Causal Discovery on Precarity and Depression

Kyuri Park[1], Leonie K. Elsenburg[2], Mary Nicolao[2], Karien Stronks[2], Vítor V. Vasconcelos[1, 3]

[1]*Computational Science Lab, Informatics Institute, University of Amsterdam, PO Box 94323, Amsterdam, 1090GH, the Netherlands*
[2]*Department of Public and Occupational Health, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherland*
[3]*Institute for Advanced Study, University of Amsterdam, Oude Turfmarkt 147, Amsterdam, 1012GC, the Netherland*

2024-12-28

Abstract

Understanding the causal mechanisms linking precarity factors and depression is critical for developing effective interventions. This study utilizes the HELIUS dataset to explore these relationships using advanced causal discovery methods. By applying algorithms such as FCI, and CCI, and combining traditional Gaussian CI tests with non-parametric approaches like RCoT, we investigate how precarity factors—including employment, social, financial, housing, and relational stress—affect depression, both as a sum score and at the individual symptom level. Our findings reveal that relational stress consistently emerges as a potential causal factor for depression, while symptoms such as sleep disturbances, guilt, and anhedonia are particularly sensitive to external stressors, acting as potential early warning signals or intervention points for prevention. Moreover, the results highlight complexities in the data, including the influence of latent confounders and the challenges of capturing cyclic relationships. Despite some limitations, such as unresolved ambiguities in causal directions and challenges with mixed data distributions, this study demonstrates the utility of causal discovery tools in disentangling the intricate interplay between social and mental health dynamics. By mapping these causal structures into computational models, future research can simulate intervention effects, providing actionable insights to mitigate the impact of precarity on mental health. This study serves as a foundational effort, offering both methodological advancements and practical implications for addressing depression at a population level.

Table of contents

1 Introduction

Mental health problems in urban areas have been reported to be on the rise. Governments have been attempting to intervene, but the complexity of mental health systems presents significant challenges in planning effective interventions, let alone understanding the underlying mechanisms driving these issues.

Recent research has aimed to identify underlying factors contributing to mental health problems, often referred to as precariousness factors. These factors encompass various aspects of life, such as employment, social connections, financial stability, housing, and cultural dimensions. This comprehensive perspective on precarity helps to highlight how different aspects of life may be interconnected. While this research has advanced our understanding of the complex interplay between precariousness factors, a key question remains unanswered: how do these factors influence mental health? Specifically, the lack of directional information — knowing what influences what — limits our ability to identify and prioritize effective intervention targets.

This study aims to investigate the causal relationships between precariousness factors and mental health outcomes, with a specific focus on depression, the most prevalent mental health issue in urban populations. Using causal discovery methods, we explore how different aspects of precarity influence depression and delve deeper into the dynamics at the symptom level. By examining individual depressive symptoms, we aim to identify which symptoms may act as initiators by being particularly sensitive to specific precariousness factors. Through this analysis, our goal is to uncover the causal mechanisms underlying mental health challenges and provide a foundation for developing more effective and targeted interventions.

2 Methods

2.1 Data

We use the HELIUS dataset, which captures the diverse population of Amsterdam across various ethnicities and provides comprehensive health and lifestyle data. To operationalize precariousness factors, we draw on the framework outlined in previous research (i.e., Leonie's paper) and select a set of relevant variables.

To ensure a robust representation of each precariousness factor, we conducted various exploratory analyses to identify consistent and meaningful factor structures. Based on these analyses, we identified five precariousness factors, including two related to recent stressors, each comprising multiple variables as outlined below. Detailed information on the exploratory analyses can be found in the Appendix.

- Employment precariousness: `emp_stat`, `work_sit`.

- Social precariousness: `soc_freq`, `soc_adq`.
- Housing precariousness: `nb_safe`, `nb_res`, `nb_rent`, `cul_rec`.
- Recent relational stressors: `frd_brk12`, `conf12`.
- Recent financial stressors: `fincri12`, `inc_diff`.

After preprocessing, the HELIUS dataset comprises 21,628 samples. In addition to the five precariousness factors, we include PHQ-9 scores — both the total sum score and individual symptom scores — to represent depression. In the subsequent causal discovery analysis, we will examine depression both as an aggregated sum score and through its individual symptom-level representations. Refer to Figure 1 for the overall distributions of the variables used in the analysis.

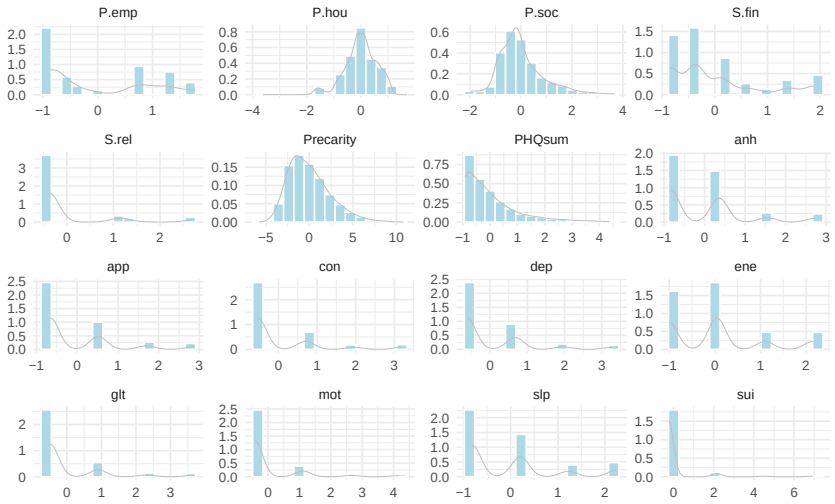


Figure 1: Distributions of variables with density overlay. *P.emp* = employment precariousness; *P.hou* = housing precariousness; *P.soc* = social precariousness; *S.fin* = recent financial stressors; *S.rel* = recent relational stressors; *PHQsum* = PHQ-9 sum score; *anh* = anhedonia; *app* = appetite; *con* = concentration; *dep* = depressed mood; *ene* = energy; *glt* = guilty; *mot* = motor; *sui* = suicidal

2.2 Causal Discovery

There are numerous causal discovery algorithms available; however, in this study, we focus on algorithms suited to the potential cyclic relationships within our system. Specifically, we use FCI (Fast Causal Inference) and CCI (Cyclic Causal Inference), both capable of accounting for such cycles under certain conditions (Mooij & Claassen, 2020; Strobl, 2019). Additionally, we include the PC algorithm as a reference, given its simplicity and prominence as one of the most widely known causal discovery methods (Spirtes et al., 2001). For a more detailed explanation of these algorithms, please refer to Park et al. (2024).

As shown in Table 1, the resulting graphs from FCI and CCI differ slightly (*PAG*:

Table 1: Assumptions of causal discovery algorithms

Algorithm	Acyclicity	Causal sufficiency	Absence of selection bias	Output
PC	✓	✓	✓	CPDAG
FCI	$-^a$	✓	$-^a$	PAG
CCI	x	x	x	MAAG

Note. [a]FCI can detect cycles in systems that lack selection bias and exhibit non-linear relationships (Mooij & Claassen, 2020).

partial ancestral graph; *MAAG*: maximal almost ancestral graph) due to their reliance on different underlying assumptions. Both encode information about causal relationships between variables, where the presence of an edge indicates causal *ancestry*. Directed edges, A *$\rightarrow$ B, specify that B is not an ancestor of A in every graph within the Markov equivalence class, $Equiv(G)$. A*—B represent cases where B is an ancestor of A in every graph in $Equiv(G)$. Circle endpoints, A*-oB, denote ambiguity in the ancestral relationship, meaning that B's ancestral status relative to A varies across graphs in $Equiv(G)$.

In contrast, the graph produced by the PC algorithm is a *CPDAG* (completed partially directed acyclic graph), where directed edges ($A \rightarrow B$) indicate that A is a direct cause (parent) of B. Unlike FCI and CCI, the CPDAG does not include circle symbols. Instead, when the PC algorithm cannot determine the direction, it represents uncertainty with bidirectional arrows. While PC serves as a useful reference, its strict assumptions of acyclicity and the absence of latent confounders limit its applicability in more complex settings. Therefore, our primary focus remains on the results from FCI and CCI. For completeness, all PC algorithm results are provided in the Appendix.

One practical challenge in applying these algorithms to the HELIUS dataset is that the data does not follow a Gaussian distribution, and the relationships among variables are unlikely to be strictly linear. To account for this, we complement the commonly used Gaussian conditional independence test (CI test), which relies on partial correlations, with a non-parametric CI test based on kernel methods. However, kernel-based non-parametric tests are computationally demanding, particularly with large datasets like ours. To mitigate this issue, we employ Randomized Conditional Independence Test (RCIT) and the Randomized conditional Correlation Test (RCoT), which uses random Fourier features to approximate the kernel methods, thereby significantly reducing the computational cost (Strobl et al., 2019). For a more detailed explanation of RCIT and RCoT, see Section 6.8.

2.3 Analysis

We analyze the causal structure using two approaches: one with the PHQ sum score representing overall depression severity, and another with individual symptom scores. The PHQ sum score simplifies the analysis by reducing dimensionality, which is computationally advantageous and provides a broad, interpretable perspective on depression's relationship with precarity factors. In contrast, individual symptom scores offer a nuanced understanding, capturing the heterogeneous ways symptoms respond to precarity factors. However, analyzing individual symptoms poses methodological challenges due to their non-standard distributions and the increased complexity introduced by the higher dimensionality. By employing both approaches, we balance simplicity and granularity, ensuring robustness in our findings: the PHQ sum score captures overarching trends, while symptom-level analysis reveals detailed dynamics, essential for tailoring specific interventions.

To evaluate the sensitivity of the results to the choice of alpha levels and ensure consistency, we test two significance levels — $\alpha = 0.01, 0.05$. To further ensure robustness, we employ bootstrapping to generate 100 bootstrap samples For each sample, we estimate causal graphs and retain only the edges and directions that appear above predefined thresholds.

The analyses are conducted under the following conditions:

- Significance levels (α): 0.01 and 0.05
- Thresholds: 0.5, 0.6, 0.7, and 0.8
- CI test: Gaussian CI test, RCoT
- Algorithms: FCI, CCI, and PC

This setup yields 16 combinations (2 significance levels × 4 thresholds × 2 CI tests), applied across three algorithms, with each combination repeated for 100 bootstrap samples, resulting in a total of 1,600 resulting graphs. For analyses involving individual symptom variables, we streamline the setup by using thresholds of 0.6 and 0.7, reducing the number of bootstrap samples to 30.

To summarize the results, we identify the most frequently occurring edge endpoints across different experimental setups.[1] This ensures that only stable and consistent edges are retained, providing a clearer and more reliable understanding of the relationships between precarity factors and depression. Refer to Figure 2 for the analysis workflow.

3 Results

3.1 Depression as sum score

The sum score graphs provide a high-level summary of how precarity factors collectively influence overall depression severity, focusing on aggregated rela-

[1]The detailed proportion of each edge endpoint occurrence is shown in Section 6.9.

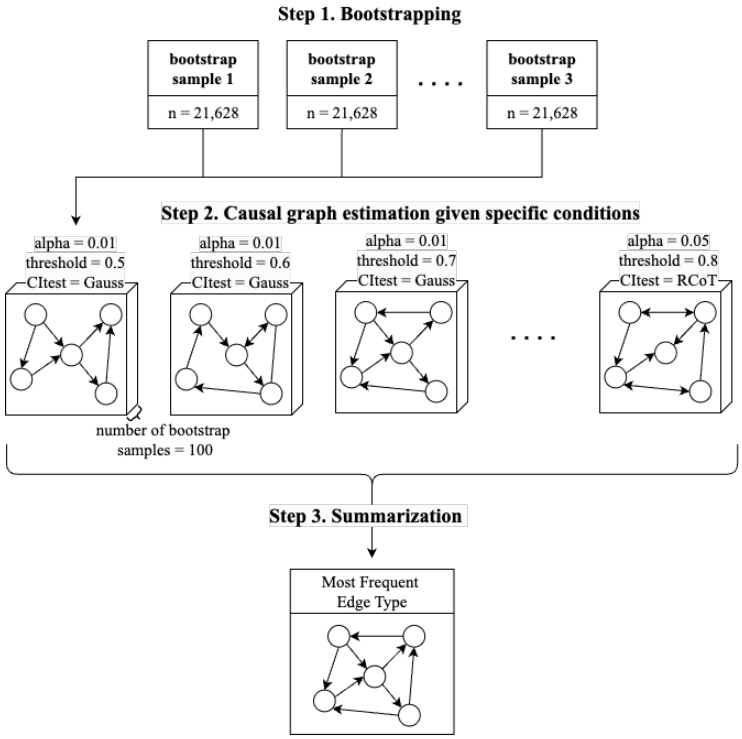


Figure 2: Analysis workflow applied across all three algorithms.

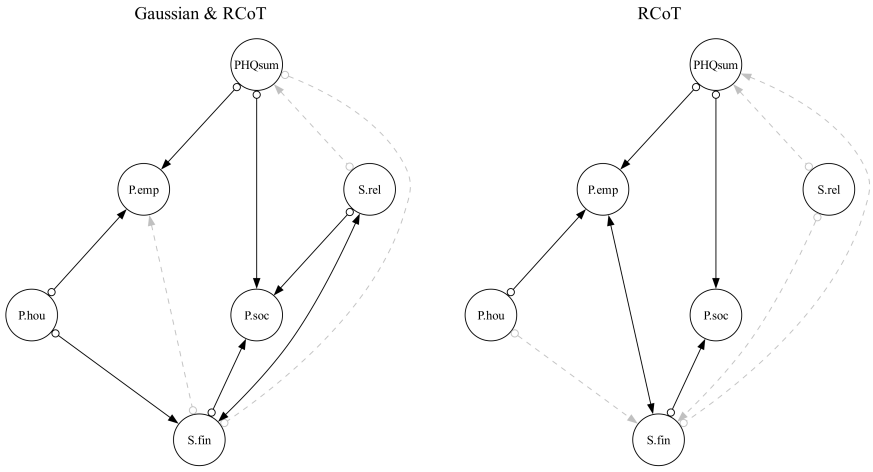
tionships. Figure 3 illustrates the causal relationships between precarity factors (*P.hou*, *P.emp*, *P.soc*, *S.rel*, *S.fin*) and the depression sum score (*PHQsum*) under two different setups: (a) using both Gaussian CI test and RCoT, and (b) using RCoT alone. Black edges represent consistent causal relationships identified by both FCI and CCI, while gray edges denote inconsistent relationships that vary between the two methods. The edge endpoints of inconsistent gray edges are marked with circles. In graph (a), the key pathways suggest that employment precarity (*P.emp*) and social precarity (*P.soc*) do not cause depression (*PHQsum*). Social precarity does not cause recent relational stress (*S.rel*) or financial stress (*S.fin*), and these stressors are likely related by a latent confounder. Additionally, *P.emp* and *S.fin* are identified as non-causes of housing precarity (*P.hou*). Both FCI and CCI detect a dependency between *P.emp* and *S.fin*, but they disagree on the direction of the relationship, leaving it unclear whether *S.fin* causes *P.emp* or if the connection is mediated by an unmeasured confounder. A similar ambiguity exists in the relationships between *S.rel* and *PHQsum* and between *S.fin* and *PHQsum*, with the methods diverging on whether these stressors influence depression or are linked through latent variables.

Graph (b), derived solely from the nonparametric RCoT test. Both graphs consistently identify that employment precarity is not a cause of housing precarity or depression, and that social precarity does not cause depression or financial stress. The relationship between stressors and depression remains unresolved between the two algorithms. However, RCoT provides greater confidence that depression is not the cause of financial stress and suggests a stronger likelihood that the financial stress may contribute to depression or that their relationship is mediated by a latent confounder. The role of relational stress has become less pronounced, as the edge between *P.soc* and *S.rel* is omitted, and the relationship between *S.rel* and *S.fin* is now inconsistent between the algorithms.

Overall, both graphs together highlight the potential roles of recent stressors as being closely linked with depression, either as causes or through latent confounders. Employment and social precarity may also be influenced by depression, either directly or through latent variables. Lastly, housing precarity does not directly impact depression but remains connected to employment precarity and financial stress, either directly or through a latent confounder.

3.2 Individual depression symptom

Moving from the sum score representation to the symptom-level graph provides a more granular perspective on the causal relationships between precarity factors and depression. This approach highlights the heterogeneity in how precarity factors influence individual depressive symptoms — *slp* (sleep), *ene* (energy), *app* (appetite), *mot* (motor), *sui* (suicidal), *anh* (anhedonia), *glt* (guilt), and *dep* (depressed mood). While the sum score graph aggregates all symptoms into a single measure—potentially obscuring nuanced relationships—the symptom graph uncovers distinct pathways for different symptoms. In the symptom-level graph (Figure 4), consistent relationships are represented by



(a) Using both GaussianCI and RCoT (b) Using only RCoT

Figure 3: Resulting graphs of precarity factors and depression sum score using FCI and CCI

black solid edges, while areas of disagreement between the FCI and CCI algorithms are denoted by gray dashed edges. Endpoints marked with circles indicate differences in directional conclusions between the algorithms. Additionally, the navy dashed edges represent relationships unique to the graphs generated using Gaussian CI testing.

The symptom-level graph reveals a complex and interconnected structure, far more intricate than the sum score graph. Its denser network highlights the strong interdependence among symptoms and suggests the presence of latent confounding influences, as indicated by numerous bidirectional arrows.

Certain nodes in the network emerge as more *causally* central, highlighting their potential importance as intervention points. Among the depressive symptoms, *anh*, *dep*, *slp*, and *glt* emerge as particularly influential, with multiple outgoing edges (o->) to other symptoms, suggesting their roles as potential key drivers within the symptom network. Especially, symptoms such as *glt*, *slp*, and *anh* are particularly critical, potentially acting as initiator or activator nodes within the network due to their apparent connections with precarity factors. These symptoms appear to be especially sensitive to external stressors, potentially manifesting early in response to such conditions and subsequently activating other interconnected symptom nodes.

The causal structure involving precarity factors in the symptom-level graph is largely consistent with the patterns observed in the sum score graphs. As in the sum graphs, relational stress (*S.rel*) emerges as a potential causal factor for depression, while employment (*P.emp*) and social precarity (*P.soc*) are more

likely to be influenced by depressive symptoms. The symptom-level analysis, however, provides more specificity by pinpointing the symptoms involved in these relationships. For instance, *slp* is identified as influencing *P.emp* or potentially through a latent confounder (*slp* o-> *P.emp*) , while *glt* appears to affect *P.soc* or may be linked via a confounder (*glt* o-> *P.soc*). Additional edges are observed in the graph generated using Gaussian CI tests, such as *slp* <-> *S.rel* and *slp* o-> *P.emp*. As seen in the sum graph, Gaussian CI testing tends to produce a denser graph. In this case, causal relationships involving *slp* are particularly prominent.

Some discrepancies, however, exist between the symptom-level and sum graphs. For example, financial stress (*S.fin*) does not have any edges with depressive symptoms in the symptom-level graph, although it maintains associations with other precarity factors. Additionally, housing precarity (*P.hou*) becomes entirely disconnected from the rest of the network, appearing as an isolated node.

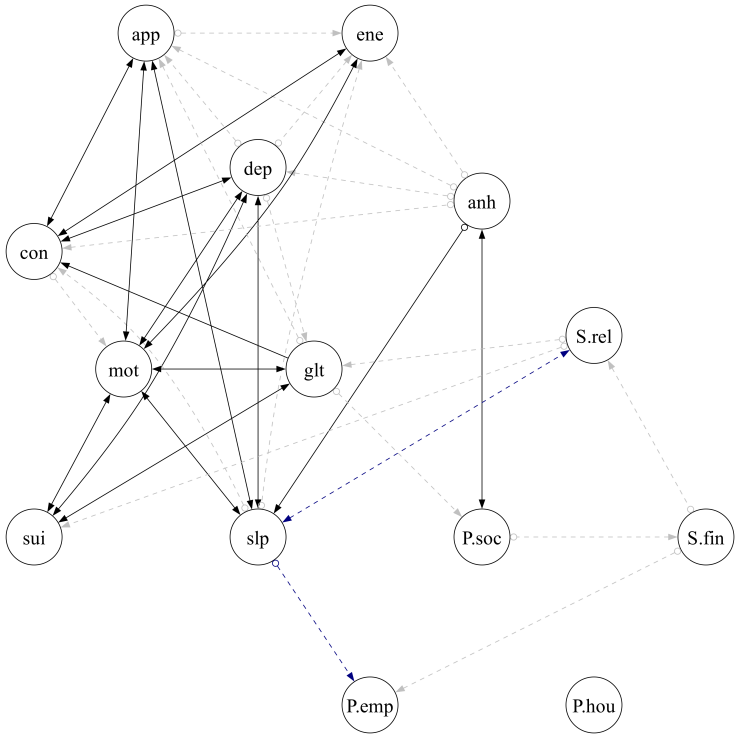


Figure 4: Resulting graphs of precarity factors and individual depression symptoms using FCI and CCI

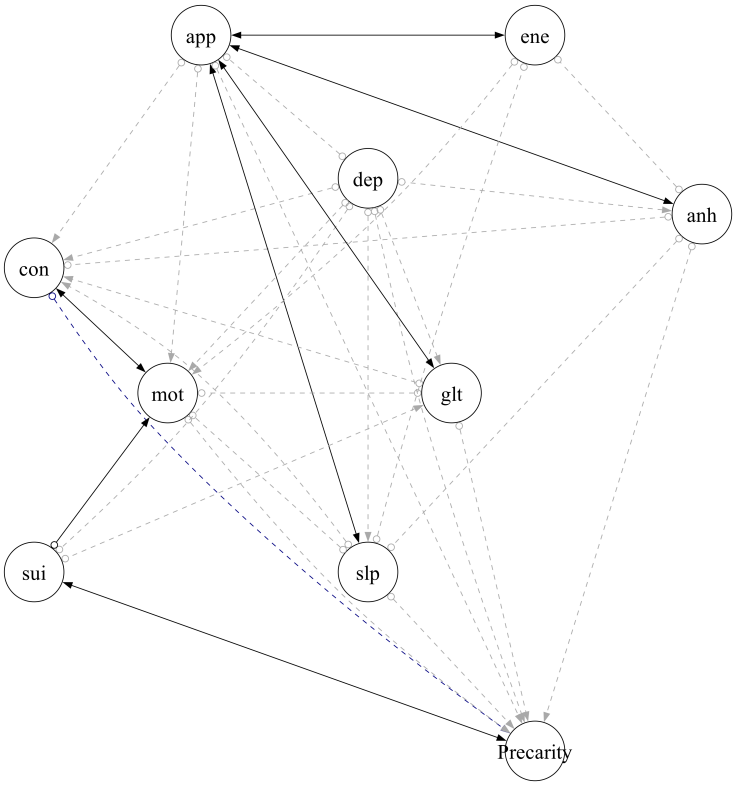


Figure 5: Resulting graphs of precarity sum score and individual depression symptoms using FCI and CCI

3.3 Precarity as sum score

4 Discussion

The present study explored the causal relationships between precarity factors and depression by employing both sum score and symptom-level analyses using causal discovery algorithms. The use of both PHQ sum scores and individual symptom scores provided a comprehensive understanding of how different aspects of precarity influence depression, revealing nuanced pathways that may be obscured when considering aggregate measures alone.

Our findings highlight the significant role of recent relational stress (*S.rel*) as a potential causal factor for depression, consistently observed across both sum score and symptom-level analyses. This consistency underscores the profound impact interpersonal relationships can have on mental health. The symptom-level analysis further identified specific symptoms, such as sleep disturbances (*slp*), guilt (*glt*), and anhedonia (*anh*), as particularly sensitive to external precarity conditions. These symptoms emerged as potential initiators or activators within the depressive symptom network, suggesting that they may serve as early warning signs or valuable targets for preventive interventions.

Employment precarity (*P.emp*) and social precarity (*P.soc*) were more likely to be influenced by depressive symptoms rather than acting as direct causes. The symptom-level graph provided additional clarity by identifying that sleep disturbances influenced employment precarity (*slp* o-> *P.emp*) and feelings of guilt affected social precarity (*glt* o-> *P.soc*). This directional insight suggests that interventions targeting specific depressive symptoms may have downstream effects on improving certain aspects of precarity.

Interestingly, financial stress (*S.fin*) did not exhibit causal relationships with individual depressive symptoms in the symptom-level analysis, contrasting with its apparent role in the sum score graphs. This divergence indicates that while financial stress may influence overall depression severity, its impact on specific symptoms may not be significant. Also, housing precarity (*P.hou*) emerged as an isolated node in the symptom-level graph, losing all connections with other precarity factors. This isolation suggests that housing precarity may operate independently of the depressive symptom network or that its effects are not captured within the scope of the measured variables.

Several limitations should be acknowledged when interpreting these findings. The prevalence of bidirectional arrows and circle-marked endpoints in the graphs reflects unresolved ambiguities in the causal relationships suggested by the data. These uncertainties underscore the need for further research, ideally incorporating datasets with a higher proportion of symptomatic individuals. The HELIUS dataset, being predominantly composed of asymptomatic samples, posed challenges in identifying clear causal directions, particularly among symptom nodes. Future studies could address these limitations by incorporating time-series data to leverage temporal information about the relationships

between precarity factors and depressive symptoms. Time-series data could provide time-specific insights and track how these relationships evolve over time. For instance, methods such as *PCMCI* (Runge et al., 2019) and *tsFCI* (Entner & Hoyer, 2010), along with other time-series adaptations of causal discovery algorithms, could help better account for temporal dependencies and refine the analysis.

Another limitation is the lack of clear evidence for cycles within the symptom network, despite using algorithms designed to account for cyclic relationships. The CCI algorithm predominantly produced bidirectional arrows, while FCI primarily generated directional arrows, yet neither displayed patterns indicative of definitive cyclic structures. Addressing cyclic relationships is particularly challenging with observational datasets alone. Future research could benefit from refined datasets that include intervention data. Methods such as *LLC* (Hyttinen et al., 2012), *NODGAS-Flow* (Sethuraman et al., 2023), and the recently developed *Bicycle* algorithm (Rohbeck et al., 2024) are specifically designed to utilize both observational and intervention data to uncover potential cycles. Additionally, if time-series data becomes available, corresponding methods, as described above, could be applied to capture repetitive patterns in variable interactions that might suggest cyclic structures.

Lastly, regarding conditional independence (CI) testing, the differences between the graphs generated by Gaussian CI and RCoT underscore the methodological sensitivities in detecting causal relationships. Gaussian CI produced denser graphs, which is somewhat counterintuitive, as the Gaussian CI's strict linearity assumption would typically result in fewer detected relationships, not more. A possible explanation for this discrepancy is that Gaussian CI's reliance on partial correlations may overestimate relationships when specific non-linear dependencies exist in the data. In contrast, RCoT, free from such assumptions, may better capture these patterns under such conditions. However, while RCoT is technically non-parametric, its performance can still be influenced by the distributional characteristics of the data. Specifically, the RBF kernel, optimized for continuous data with smooth transitions, may struggle to capture relationships in datasets with discrete or mixed distributions. In such cases, the distances between discrete points may fail to convey meaningful similarity information. As a result, RCoT might miss certain dependencies, particularly when variables in the dataset lack smooth continuity. Future research could address these issues by exploring a broader range of CI testing approaches. One traditional approach to handle this is discretizing variables and use G^2 test, which may better capture dependencies in non-continuous datasets (Dojer, 2016; Neapolitan et al., 2004). A more promising direction, however, lies in the development of hybrid kernels tailored for mixed datasets, effectively integrating both continuous and discrete variables into RCoT-like methods. By systematically employing and comparing a wider variety of CI testing techniques, researchers could gain more robust insights and mitigate the limitations inherent in specific approaches. This broader exploration holds the potential to enhance the reliability of findings, particularly in datasets with complex and heterogeneous structures.

Despite its limitations, this study marks a meaningful step toward understanding the mechanisms linking precarity factors and depression. By applying causal discovery methods, it moves beyond traditional association-based analyses, providing insights that can inform more precise and targeted interventions. While the resulting graphs are preliminary and contain unresolved ambiguities, they offer a valuable starting point for leveraging causal discovery tools to investigate the causal interplay between depression and precarity factors. A promising next step would involve integrating these causal structures into computational models, such as the symptom dynamic model proposed by **our comp-model paper**. By simulating intervention effects, such models could provide more realistic insights into how targeted actions might influence symptom networks and precarity factors over time. For example, interventions focused on improving sleep hygiene or alleviating guilt could be evaluated for their cascading effects on employment and social relationships, offering actionable guidance for designing population-level mental health strategies. As one of the early applications of causal discovery tools to the complex dynamics of depression and precarity factors, this study lays a foundation for future research. We hope it inspires further refinement of these methods and ultimately contribute to more effective solutions for alleviating depression and improving societal well-being.

5 References

Dojer, N. (2016). Learning bayesian networks from datasets joining continuous and discrete variables. *International Journal of Approximate Reasoning*, *78*, 116–124.

Entner, D., & Hoyer, P. O. (2010). On causal discovery from time series data using FCI. *Probabilistic Graphical Models*, *16*.

Hyttinen, A., Eberhardt, F., & Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, *13*(1), 3387–3439.

Lindsay, B. G., Pilla, R. S., & Basak, P. (2000). Moment-based approximations of distributions using mixtures: Theory and applications. *Annals of the Institute of Statistical Mathematics*, *52*, 215–230.

Mooij, J. M., & Claassen, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. *Conference on Uncertainty in Artificial Intelligence*, 1159–1168.

Neapolitan, R. E. et al. (2004). *Learning bayesian networks* (Vol. 38). Pearson Prentice Hall Upper Saddle River.

Park, K., Waldorp, L. J., & Ryan, O. (2024). Discovering cyclic causal models in psychological research. *Advances. In/Psychology*, *2*, e72425.

Rohbeck, M., Clarke, B., Mikulik, K., Pettet, A., Stegle, O., & Ueltzhöffer, K. (2024). Bicycle: Intervention-based causal discovery with cycles. In F. Locatello & V. Didelez (Eds.), *Proceedings of the third conference on causal learning and reasoning* (Vol. 236, pp. 209–242). PMLR. https://proceedings.mlr.press/

v236/rohbeck24a.html
Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996.
Sethuraman, M. G., Lopez, R., Mohan, R., Fekri, F., Biancalani, T., & Hütter, J.-C. (2023). NODAGS-flow: Nonlinear cyclic causal structure learning. *International Conference on Artificial Intelligence and Statistics*, 6371–6387.
Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search*. MIT press.
Strobl, E. V. (2019). A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, *8*(1), 33–56. https://doi.org/10.1007/s41060-018-0158-2
Strobl, E. V., Zhang, K., & Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, *7*(1), 20180017.
Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv Preprint arXiv:1202.3775*.

6 Appendix

6.1 Precariousness factors by Leonie

1. EMPLOYMENT PRECARIOUSNESS

- `H1_Arbeidsparticipatie`: Working status
- `H1_WerkSit`: Which work situation most applies to you?
- `H1_RecentErv8`: Experiences past 12 months: h. You were sacked from your job or became unemployed (*reverse*)

2. FINANCIAL PRECARIOUSNESS

- `H1_InkHhMoeite`: During the past year, did you have problems managing your household income?
- `H1_RecentErv9`: Experiences past 12 months: i. You had a major financial crisis (*reverse*)

3. HOUSING PRECARIOUSNESS

- `veilig_2012`: Score safety (veiligheid) in 2012 (*reverse*)
- `vrz_2012`: Score level of resources (niveau voorzieningen) in 2012 (*reverse*)
- `P_HUURWON`: Percentage Huurwoningen

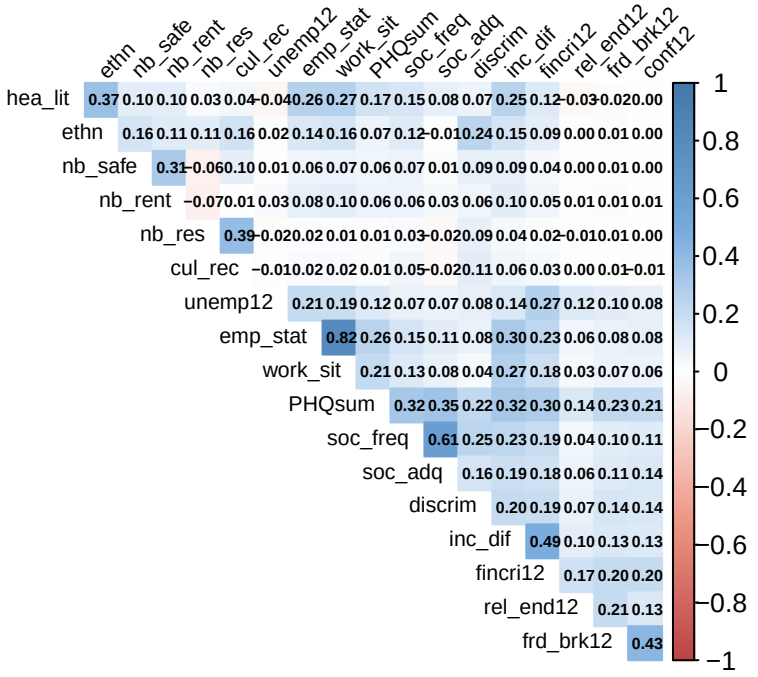
4. CULTURAL PRECARIOUSNESS

- `H1_Discr_sumscore`: Perceived discrimination: sum score of 9 items (range 9-45)

- `H1_SBSQ_meanscore`: Health literacy: SBSQ meanscore (range 1-5) (*reverse*)
- `A_BED_RU`: Aantal bedrijfsvestigingen; cultuur, recreatie, overige diensten (*reverse*)

5. SOCIAL PRECARIOUSNESS

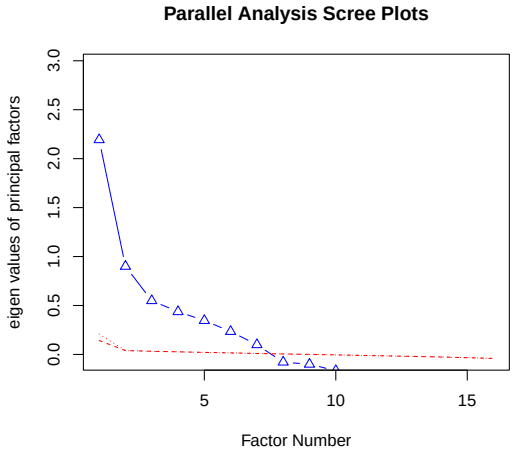
- `H1_RecentErv5`: Experiences past 12 months: e. Your steady relationship ended (*reverse*)
- `H1_RecentErv6`: Experiences past 12 months: f. A long-term friendship with a good friend or family member was broken off (*reverse*)
- `H1_RecentErv7`: Experiences past 12 months: g. You had a serious problem with a good friend or family member, or neighbour (*reverse*)
- `H1_SSQT`: SSQT (frequency of social contact): sum score of 5 items (range 5-20) (*reverse*)
- `H1_SSQSa`: SSQS (adequacy of social contact): sum score of 5 items, category 3 and 4 not combined (range 5-20) (*reverse*)

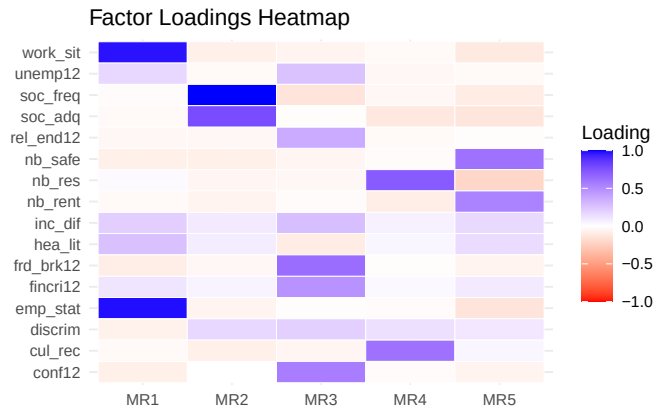


- **High** Correlations: `emp_stat` (employment status) and `work_sit` (work situation) have a strong positive correlation of 0.82. This suggests that individuals with higher employment status tend to have more secure or favorable work situations. `soc_freq` (social contact frequency) shows a strong positive correlation with `soc_adq` (social adequacy) at 0.61. This indicates that individuals with more frequent social contact also tend to have higher perceived adequacy of social interactions.

- **Moderate** Correlations: `nb_safe` (neighborhood safety) and `nb_res` (resources) have a moderate positive correlation of 0.39, suggesting that areas with higher safety also have better resources. `hea_lit` (health literacy) has moderate correlations with `emp_stat` (0.26) and `work_sit` (0.25), which could mean that higher health literacy is associated with better employment situations. `frd_brk12` (friendship breakups) and `conf12` (conflicts) have a notable correlation of 0.43, indicating a relationship between having conflicts and friendship losses.
- **Low to Moderate** Correlations in Financial Precariousness: `inc_dif` (income difficulties) has a moderate correlation with `fincri12` (financial crisis) at 0.49. This aligns with the expected relationship, where individuals who experience general income difficulties are more likely to report financial crises.
- **Low** Correlations (0.1 - 0.2): Many variables, such as `discrim` (discrimination), `unemp12` (unemployment experience), and `rel_end12` (relationship end), have low correlations with other variables, suggesting relatively independent relationships in the context of this dataset.

6.2 Exploratory Factor Analysis (EFA)





6.2.1 Factor Loadings (Pattern Matrix)

- **MR1**: High loadings on `emp_stat` and `work_sit` suggest this factor captures *employment* precariousness.
- **MR2**: Strong loadings on `soc_freq` and `soc_adq` indicate *social* precariousness.
- **MR3**: Key items like `frd_brk12`, `conf12`, and `fincri12`, suggest recent *stressful events*.
- **MR4**: High loadings on `nb_res` and `cul_rec` may reflect *community resources* precariousness.
- **MR5**: Variables `nb_safe` and `nb_rent` with high loadings indicate *housing* precariousness.

6.2.2 Variance Explained

The factors cumulatively explain **38%** of the variance, with MR1 being the most influential factor. Each factor contributes a smaller proportion to the total variance (MR1 at 12%, MR2 at 9%, etc.).

6.2.3 Factor Intercorrelations

Factors are moderately correlated, especially between *MR1 and MR5*, and *MR2 and MR3*. This indicates that while distinct, these factors are related—reasonable in a complex socio-economic context.

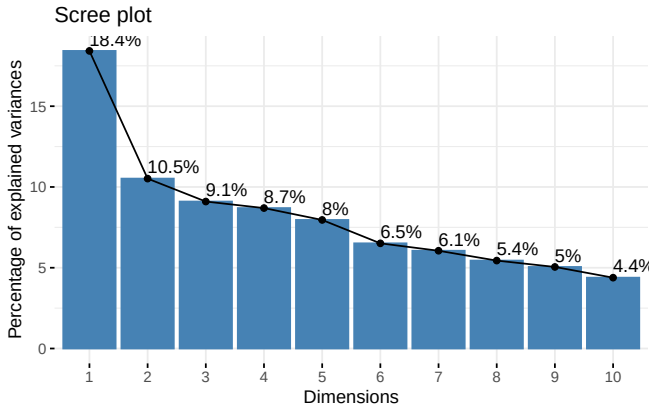
6.2.4 Model Fit Statistics

RMSEA (0.071) suggest an acceptable fit. Tucker Lewis Index (0.802) suggests moderate reliability for the model.

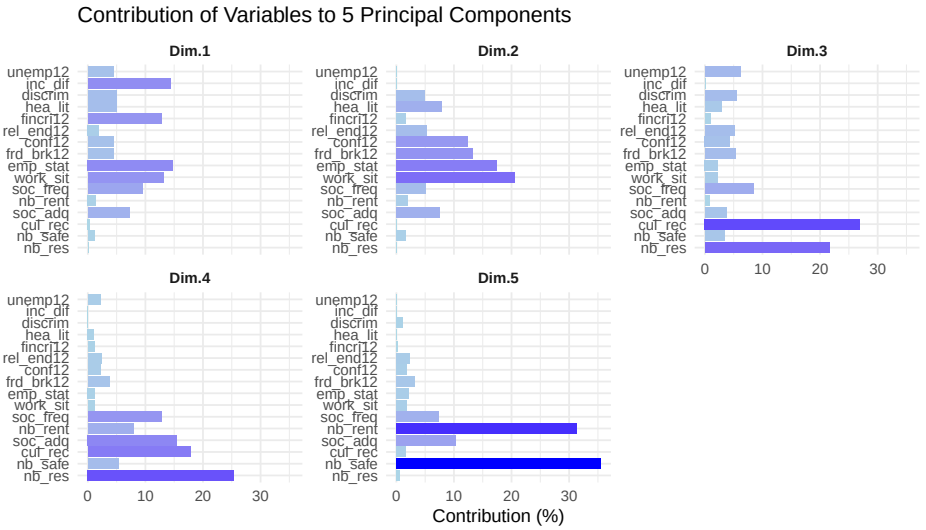
6.2.5 Summary

The 5-factor model appears interpretable and captures distinct dimensions of precariousness: *employment, social, stressors, community resources, and housing precariousness*. Although the overall fit and explained variance could be stronger, these factors offer insights into the underlying structure of the data, highlighting key areas of precariousness.

6.3 PCA



- Component Retention: The scree plot shows a clear "elbow" after the first component. This steep drop suggests that most variance is explained by the first component. After Dimension 5, the percentage of explained variance decreases slightly more gradually, indicating diminishing returns for adding more components. If we need to choose multiple components, retaining the first 5 components seems reasonable, as they capture most of the variance (cumulatively explaining about 54.7% of the total variance).



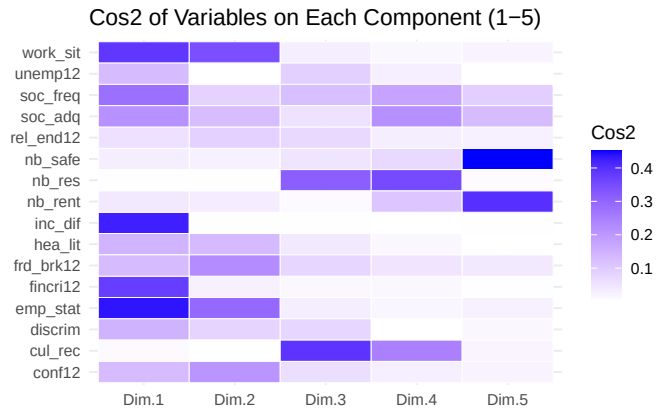
6.3.1 Explained variance (contributions) of variables

It shows the importance of variables within each component.

- **Dim1**: High contributions are observed from `emp_stat`, `work_sit`, `inc_dif`, and `fincri12`, suggesting that this dimension captures aspects of *employment and financial* security.
- **Dim2**: While `emp_stat` and `work_sit` overlap with Dim1, the strong contributions from `frd_brk12` and `rel_end12` indicate that this dimension captures a focus on *recent relationship stressors*.
- **Dim3**: `cul_rec`, `nb_res` have the highest contributions, indicating this dimension likely represents *community and cultural* factors.
- **Dim4**: `soc_freq` and `soc_adq` stand out in this dimension, suggesting an emphasis on *social* precariousness.
- **Dim5**: `nb_safe` and `nb_rent` are the top contributors, pointing to *housing* security as key themes in this component.

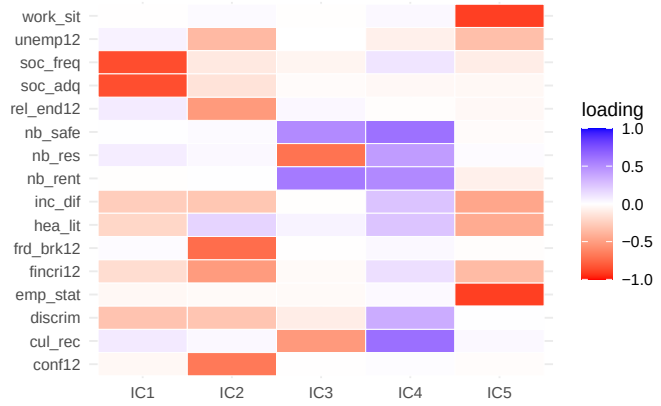
6.3.2 Cos² Values

Cos² (squared cosine) values, or the quality of representation, show how well each variable is represented by each dimension. where higher cos² values (closer to 1) indicate better representation of a variable by a component.



- **Dim.1**: Variables `emp_stat`, `work_sit`, `inc_dif`, and `fincri12` show high cos² values, meaning that PC1 primarily captures variations in employment and financial difficulties. This component could represent *employment & finance* precariousness.
- **Dim.2**: Variables `work_sit`, `emp_stat`, `frd_brk12`, and `conf12` are well-represented in this component, suggesting PC2 captures aspects of *recent relationship stressors*.
- **Dim.3**: Variables `nb_res` and `cul_rec` load strongly on PC3. This may represent community or cultural resources, indicating that this component is associated with *neighborhood resources*.
- **Dim.4**: This component has high cos² values for `nb_res`, `cul_rec`, `soc_freq`, and `soc_adq`. While `nb_res` and `cul_rec` are also prominent in PC3, PC4 uniquely captures nuanced differentiation in *social* precariousness.
- **Dim.5**: `nb_safe` and `nb_rent` are well-represented by PC5. This component might capture *housing* precariousness.

6.4 ICA



6.4.1 Dominant Variables per Component:

For each Independent Component (IC), we can identify variables with *high absolute* values in each column. These values indicate that the IC captures a strong, independent signal associated with these variables.

- **IC1**: `soc_freq` and `soc_adq` have strong negative loadings on this component, indicating that this component might represent *social* precariousness.
- **IC2**: `frd_brk12`, `conf12`, `rel_end12`, `fincri12` and `unemp12` have the most substantial loadings on this component, all with negative signs. This might point to a *recent relational or social stressor* component.
- **IC3**: `nb_res` and `cul_rec` show notable negative loadings, pointing to a focus on *community resource* precariousness.
- **IC4**: High loadings for `nb_safe`, `nb_rent`, `nb_res`, `cul_rec`, and `discrim` suggest a theme of *housing and community-based* precariousness, reflecting both safety and social challenges within the neighborhood context.
- **IC5**: `emp_stat` and `work_sit` both have strong negative loadings on this component, suggesting it captures *employment* precariousness.

6.5 Hierarchical clustering

6.5.1 Using Euclidean distance

- Ward.D's method: Minimizes the variance within clusters, producing more compact and spherical clusters.
- Single linkage: Groups clusters based on the minimum distance between points.

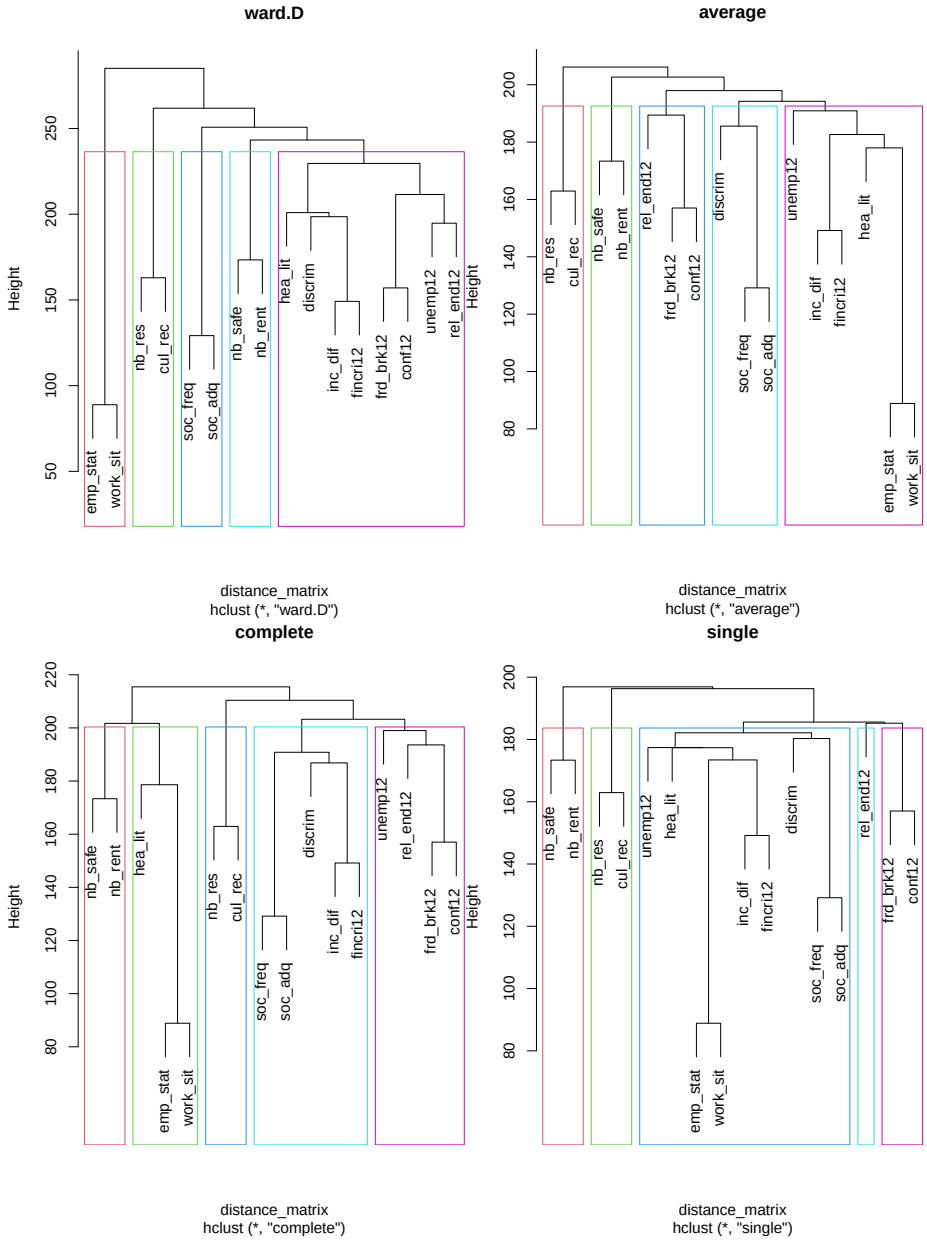
- Complete linkage: Groups clusters based on the maximum distance between points.
- Average linkage: Uses the average distance between all pairs of points in the two clusters.

6.5.1.1 Consistent Groupings (Across All or Most Methods)

- `emp_stat` and `work_sit`: This pair consistently clusters together across all linkage methods, suggesting that they are closely related variables, likely capturing a similar aspect of the data (possibly employment status or employment-related information).
- `nb_safe`, `nb_res`, and `nb_rent`: These variables are often grouped closely in several methods (especially Ward.D, average, and complete linkage). This suggests a similarity or common theme among them, potentially related to neighborhood or housing precariousness.
- `soc_freq` and `soc_adq`: These two variables frequently cluster together, indicating they likely measure aspects of social frequency and adequacy in similar ways. They appear together in Ward.D, average, and complete linkage.
- `frd_brk12` and `conf12`: These variables are often clustered closely (though they sometimes join with other variables like rel_end12), suggesting they may capture aspects of relationship or social conflict. This pair appears in close proximity, especially in average and Ward.D.

6.5.1.2 Inconsistent Groupings (Variability Across Methods)

- `hea_lit`: This variable shows inconsistent clustering across methods. In Ward.D, it joins with `fincri12`, while in other methods, it's often more isolated or grouped with variables that do not appear similar. This may suggest that `hea_lit` does not strongly correlate with other variables, or it has multidimensional aspects affecting its grouping across methods.
- `discrim`: This variable also shows variable groupings. In Ward.D, it is grouped with `hea_lit`, while in other methods (e.g., complete and single linkage), it clusters differently, sometimes on its own. This variability may indicate that `discrim` has weaker associations with the main clusters in the data or overlaps partially with multiple clusters.
- Social and Financial Variables (`inc_dif`, `fincri12`, `unemp12`): These variables appear together in some methods (e.g., Ward.D clusters `fincri12` and `inc_dif`), but in others, they are spread out. This inconsistency suggests that social and financial variables may not have strong or consistent ties across different methods, perhaps due to capturing different aspects of precariousness.

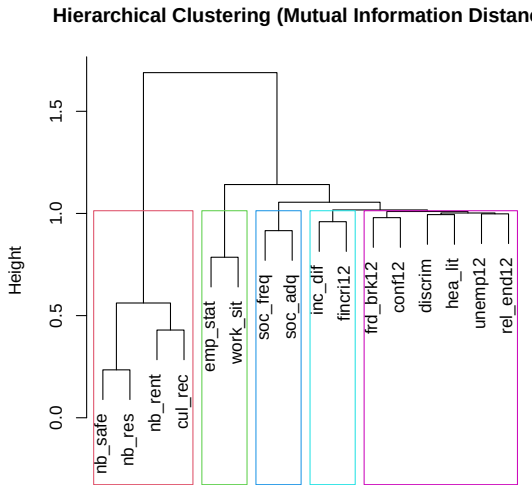

ward.D
Height
50
100
150
200
250
emp_stat
work_sit
nb_res
cul_rec
soc_freq
soc_adq
nb_safe
nb_rent
hea_lit
discrim
inc_dif
fincri12
frd_brk12
conf12
unemp12
rel_end12
Height
distance_matrix
hclust (*, "ward.D")
average
80
100
120
140
160
180
200
nb_res
cul_rec
nb_safe
nb_rent
rel_end12
frd_brk12
conf12
discrim
soc_freq
soc_adq
unemp12
inc_dif
fincri12
hea_lit
emp_stat
work_sit
distance_matrix
hclust (*, "average")
complete
Height
80
100
120
140
160
180
200
220
nb_safe
nb_rent
hea_lit
emp_stat
work_sit
nb_res
cul_rec
soc_freq
soc_adq
discrim
inc_dif
fincri12
unemp12
rel_end12
frd_brk12
conf12
Height
distance_matrix
hclust (*, "complete")
single
80
100
120
140
160
180
200
nb_safe
nb_rent
nb_res
cul_rec
unemp12
hea_lit
emp_stat
work_sit
inc_dif
fincri12
discrim
soc_freq
soc_adq
rel_end12
frd_brk12
conf12
distance_matrix
hclust (*, "single")

6.5.1.3 Summary

The consistent clusters are likely capturing distinct thematic dimensions of the data (e.g., employment, housing, social contact), while the inconsistent variables may reflect multifaceted or weakly correlated attributes that do not fit neatly into one cluster.

6.5.2 Using Mutual Information

Using mutual information (MI) as a basis for hierarchical clustering differs from using traditional distance measures (like Euclidean distance) in a few key ways.



6.5.2.1 Comparison to Euclidean Distance Clustering

- **Housing and Community** Cluster: The variables `nb_safe`, `nb_res`, `nb_rent`, and `cul_rec` cluster together, indicating a strong association among housing-related and community-based factors. This suggests a shared theme of housing or community precariousness. This grouping is also observed in the Euclidean-based clustering, but it appears more tightly connected here, potentially due to the non-linear relationships highlighted by mutual information.
- **Employment and Social Support** Cluster: `emp_stat` and `work_sit` form a cluster, linking employment status and work situation together as they did in Euclidean-based clustering. These remain closely associated regardless of the distance metric used. `soc_freq` and `soc_adq`, related to social contact frequency and adequacy, cluster nearby, indicating they have a stronger non-linear relationship with employment variables. This is a subtle difference as Euclidean distance might not capture this association as effectively.

- **Financial Stressor** Cluster: `inc_dif` and `fincri12`, representing income difficulties and recent financial crises, consistently cluster together in both approaches, showing a strong association, likely linear. However, mutual information-based clustering links these financial stressors with social support variables, suggesting that financial challenges may have complex dependencies with social support in this dataset.
- **Relational Stressor** Cluster: `frd_brk12`, `conf12`, `discrim`, `hea_lit`, `unemp12`, and `rel_end12` form a *looser* cluster focused on social and relational stressors (e.g., friendship breakup, conflicts, and discrimination). Compared to Euclidean clustering, `discrim` and `hea_lit` (health literacy) appear closer to relational stressors here, indicating that non-linear relationships might play a larger role in linking these variables.

6.5.2.2 Summary

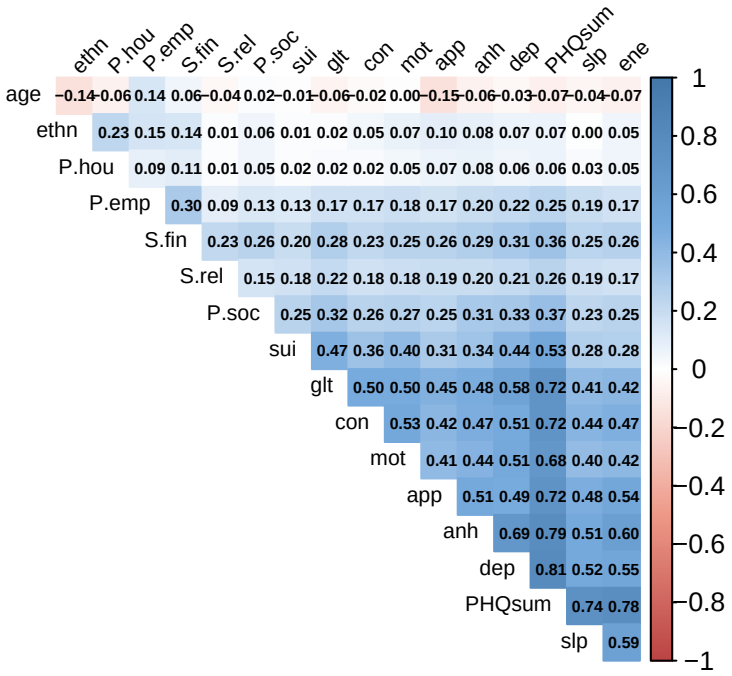
In conclusion, mutual information-based clustering provides an alternative perspective that can reveal more intricate associations between variables, especially for those with non-linear relationships. Compared to Euclidean clustering, it shows a similar high-level structure but emphasizes nuanced connections between variables, particularly around social support, employment, and financial stress.

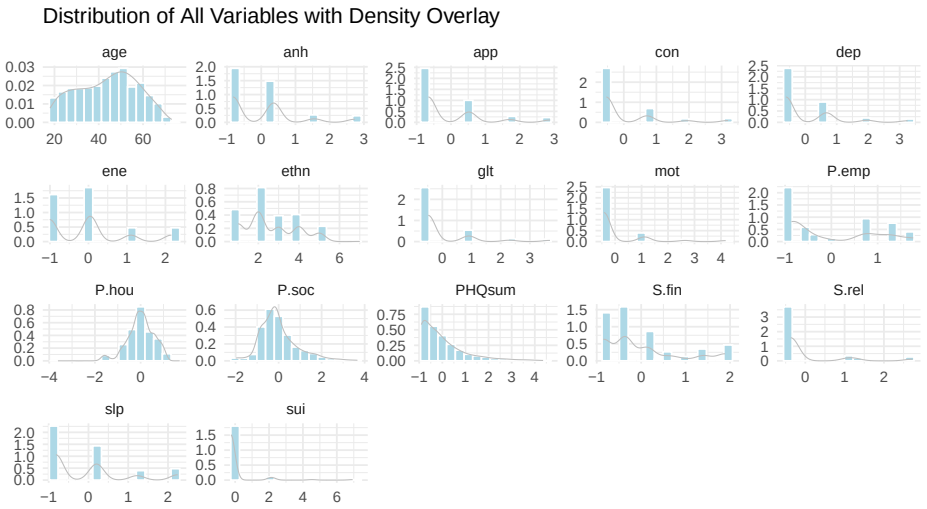
6.6 Conclusions on Precariousness factors

Based on the consistent findings across multiple analyses, we decided to exclude the variables `discrim`, `hea_lit`, `umemp12`, and `rel_end12`, as they do not clearly belong to any specific precariousness factor nor exhibit strong associations with depression (see the correlation table above). Therefore, we propose retaining the following key precariousness factors:

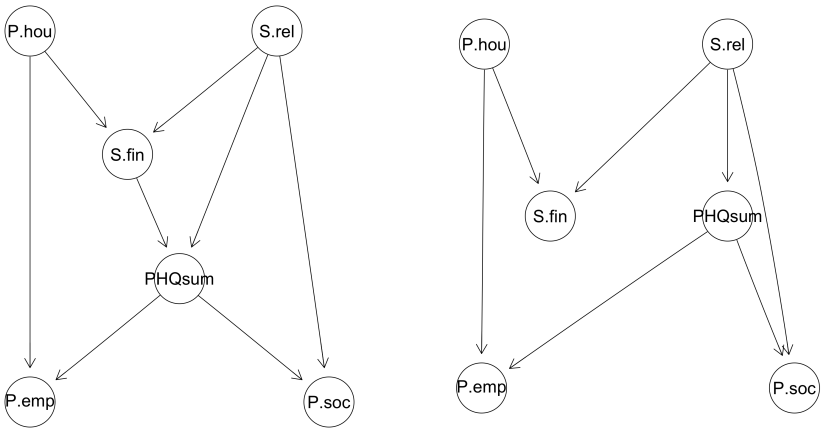
- Employment Precariousness: `emp_stat`, `work_sit`
- Social Precariousness: `soc_freq`, `soc_adq`
- Housing Precariousness: `nb_safe`, `nb_res`, `nb_rent`, `cul_rec`
- Recent Relational Stressors: `frd_brk12`, `conf12`
- Recent Financial Stressors: `fincri12`, `inc_diff`

We construct each precariousness factor by calculating the mean value of the combined variables. Below, we present the updated correlation table for the newly composed factors, along with the corresponding distributions of all variables to be used in the causal discovery analysis.



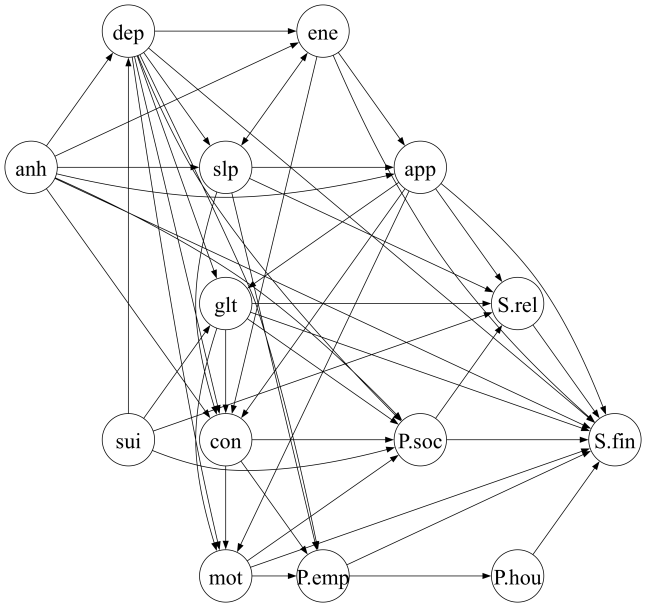


6.7 Results from PC algorithm

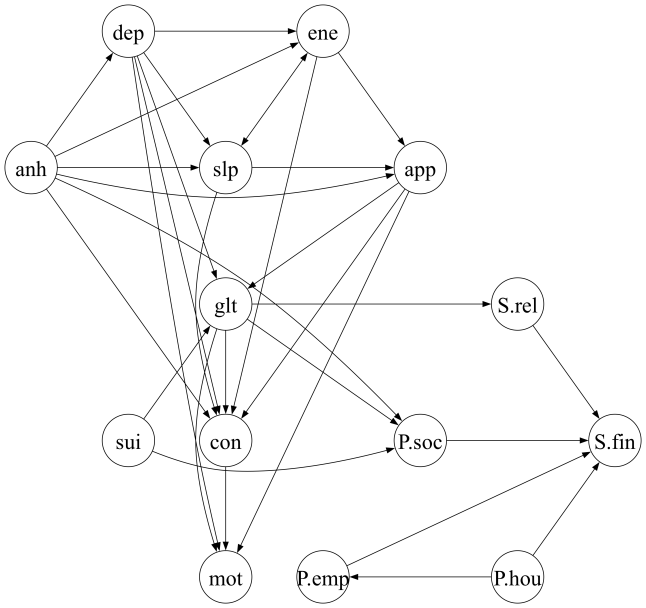


(a) Using both GaussianCI and RCoT (b) Using only RCoT

Figure 6: Resulting graphs of precarity factors and depression sum score using PC

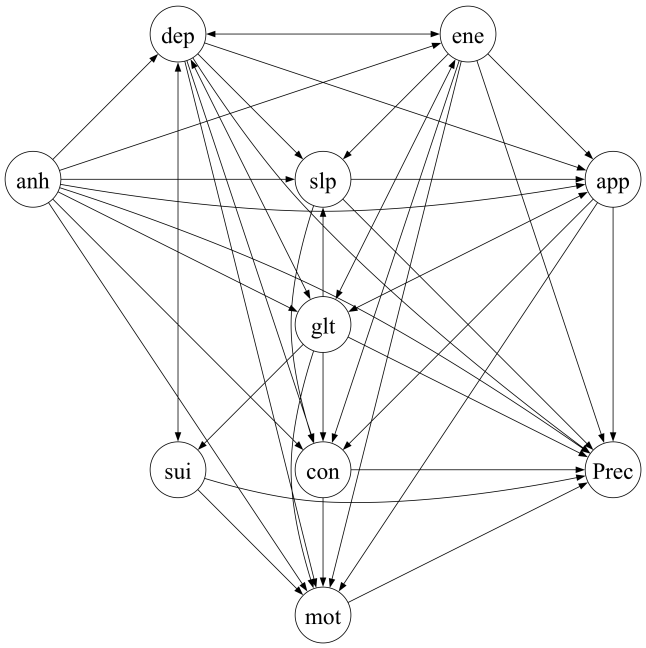


(a) Using both GaussianCI and RCoT

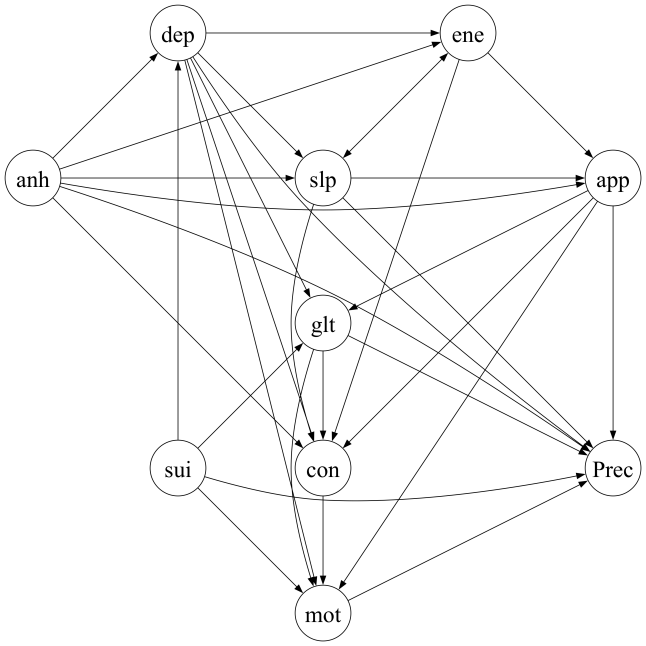


(b) Using only RCoT

Figure 7: Resulting graphs of precarity factors and individual depression symptoms using PC



(a) Using both GaussianCI and RCoT



(b) Using only RCoT

Figure 8: Resulting graphs of precarity sum score and individual depression symptoms using PC

6.8 Randomized Conditional Independence / Correlation Test (RCIT & RCoT)

RCIT (Randomized Conditional Independence Test) and RCoT (Randomized conditional Correlation Test) are advanced methods for scalable conditional independence (CI) testing, offering computational efficiency while maintaining the accuracy of kernel-based approaches. These methods evaluate conditional independence between two variables X and Y given a third variable Z while addressing computational challenges inherent in kernel-based CI tests. In this section, we provide a high-level overview of RCIT and RCoT based on (Strobl et al., 2019).

6.8.1 Kernel-Based Conditional Independence Testing

Traditional kernel-based CI tests, such as the Kernel Conditional Independence Test (KCIT), compute dependencies using the Hilbert-Schmidt Independence Criterion (HSIC) in reproducing kernel Hilbert spaces (RKHS) (Zhang et al., 2012). KCIT uses the following hypothesis framework:

$$H_0 : X \perp\!\!\!\perp Y \mid Z, \quad H_1 : X \not\perp\!\!\!\perp Y \mid Z.$$

The core quantity in KCIT is the partial cross-covariance operator:

$$\Sigma_{XY \cdot Z} = \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY},$$

where Σ_{XY} represents the cross-covariance operator between X and Y, and $\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$ removes the dependence mediated by Z.

The squared Hilbert-Schmidt (HS) norm of $\Sigma_{XY \cdot Z}$ serves as the test statistic:

$$\|\Sigma_{XY \cdot Z}\|_{HS}^2 = 0 \quad \text{if and only if} \quad X \perp\!\!\!\perp Y \mid Z.$$

KCIT estimates residual dependencies using kernel ridge regression:

$$f^*(z) = K_Z(K_Z + \lambda I)^{-1} f(x),$$

where K_Z is the kernel matrix for Z, $f(x)$ is the kernel feature map for X, and λ is the ridge regularization parameter. The residual function for X is:

$$f_{\text{res}}(x) = f(x) - f^*(z) = R_Z f(x),$$

with:

$$R_Z = I - K_Z(K_Z + \lambda I)^{-1}.$$

The kernel matrix for residualized X is:

$$K_{X \cdot Z} = R_Z K_X R_Z,$$

and similarly for Y, $K_{Y \cdot Z} = R_Z K_Y R_Z$.

The test statistic is computed as:

$$T_{XY \cdot Z} = \frac{1}{n^2} \text{tr}(K_{X \cdot Z} K_{Y \cdot Z}),$$

which estimates the Hilbert-Schmidt (HS) norm of the partial cross-covariance operator. To ensure convergence, KCIT scales the statistic by n:

$$S_K = n T_{XY \cdot Z}.$$

The null hypothesis H_0 is rejected if S_K exceeds a threshold determined by permutation or moment-matching-based null distribution (Lindsay et al., 2000).

6.8.2 Random Fourier Features (RFFs)

Kernel-based methods like KCIT face scalability issues, as they involve operations on $n \times n$ kernel matrices, which scale quadratically with the sample size n. RCIT and RCoT overcome this bottleneck using *Random Fourier Features (RFFs)* to approximate kernel operations efficiently.

6.8.2.1 Bochner's Theorem

Bochner's theorem provides the foundation for RFFs, stating that any continuous shift-invariant kernel $k(x, y)$ can be expressed as:

$$k(x, y) = \int_{\mathbb{R}^p} e^{i\omega^\top (x-y)} \, dP_\omega,$$

where P_ω is the spectral distribution of the kernel. For the widely used RBF kernel:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right),$$

P_ω follows a Gaussian distribution: $\omega \sim \mathcal{N}(0, \sigma^2 I)$.

6.8.2.2 RFF Approximation

Using Monte Carlo sampling, the kernel function is approximated as:

$$k(x, y) \approx \phi(x)^\top \phi(y),$$

where $\phi(x)$ is the random Fourier feature mapping:

$$\phi(x) = \sqrt{\frac{2}{D}} \cos(W^\top x + b),$$

with $W \sim \mathcal{N}(0, \sigma^2 I)$ and $b \sim \text{Uniform}(0, 2\pi)$. Here, D is the number of Fourier features, which balances computational efficiency and approximation accuracy.

6.8.3 Differences Between RCIT and RCoT

RCIT and RCoT differ in their test statistics, computational efficiency, and practical performance, which makes them suited for different scenarios in causal discovery. RCIT evaluates the Hilbert-Schmidt norm of the full partial cross-covariance operator, providing a general test for conditional independence but at a higher computational cost. RCoT simplifies the process by using the Frobenius norm of a finite-dimensional residualized cross-covariance matrix, significantly reducing complexity and improving scalability.

These distinctions are particularly important for large-scale datasets, where RCoT's computational efficiency makes it a practical choice for high-dimensional causal discovery tasks.

6.8.3.1 RCIT: Randomized Conditional Independence Test

RCIT tests full conditional independence by examining the squared Hilbert-Schmidt (HS) norm of the partial cross-covariance operator $\Sigma_{XY \cdot Z}$:

$$S_K = nT_{XY \cdot Z} = \frac{1}{n}\text{tr}(K_{X \cdot Z} K_{Y \cdot Z}),$$

where $T_{XY \cdot Z}$ is an empirical estimate of $\|\Sigma_{XY \cdot Z}\|^2_{HS}$. The null and alternative hypotheses are:

$$H_0 : \|\Sigma_{XY \cdot Z}\|^2_{HS} = 0, \quad H_1 : \|\Sigma_{XY \cdot Z}\|^2_{HS} > 0.$$

RCIT is a general test for conditional independence but becomes computationally demanding as the size of Z increases, due to the high-dimensional kernel operations required.

6.8.3.2 RCoT: Randomized Conditional Correlation Test

RCoT simplifies the testing process by using a finite-dimensional partial cross-covariance matrix, avoiding full HS norm calculations. Instead, it uses the Frobenius norm of the residualized cross-covariance matrix:

$$S' = n\|C_{AB \cdot C}\|^2_F,$$

where $C_{AB \cdot C}$ represents the residualized cross-covariance matrix. The hypotheses are:

$$H_0 : \|C_{AB \cdot C}\|^2_F = 0, \quad H_1 : \|C_{AB \cdot C}\|^2_F > 0.$$

RCoT is computationally efficient and well-suited for large conditioning sets ($|Z| \geq 4$). Its simplicity enables robust calibration of the null distribution and improved scalability for high-dimensional data.

6.9 Summarized Stable Edges Proportion

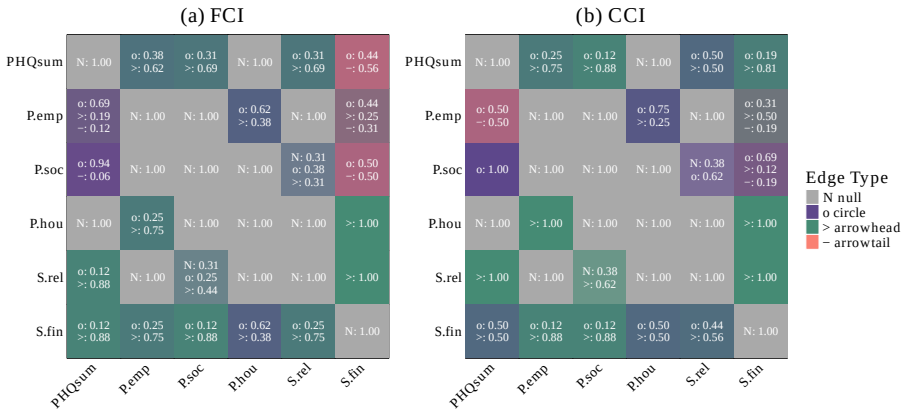
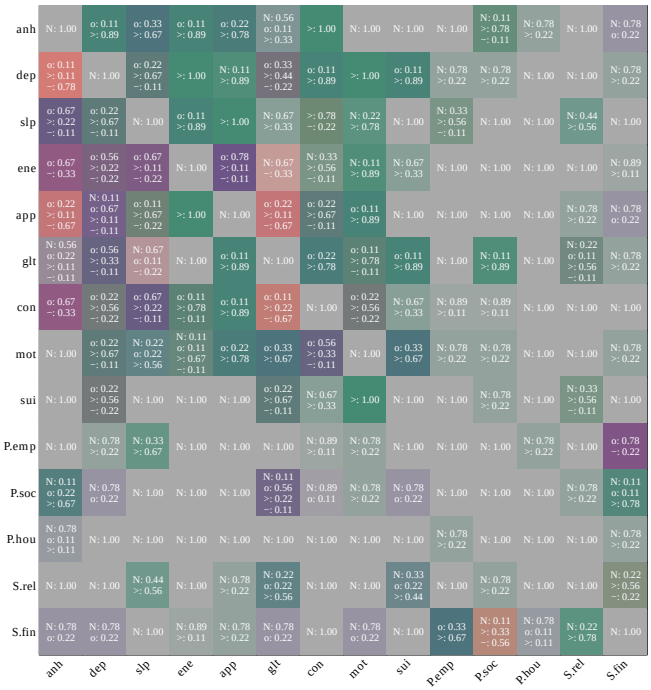
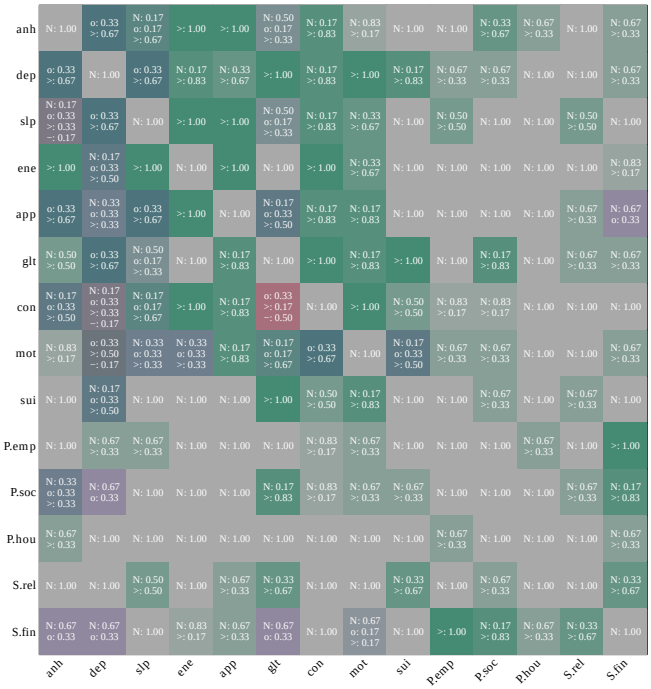


Figure 9: Proportions of edge endpoint types for graphs based on depression sum score

(a) FCI



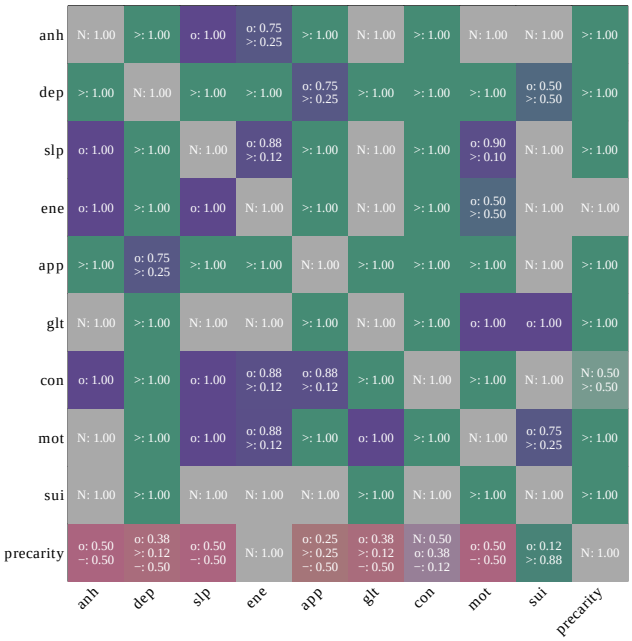
(b) CCI



Edge Type N null o circle > arrowhead − arrowtail

Figure 10: Proportions of edge endpoint types for graphs based on individual depression symptoms

(a) FCI



(b) CCI

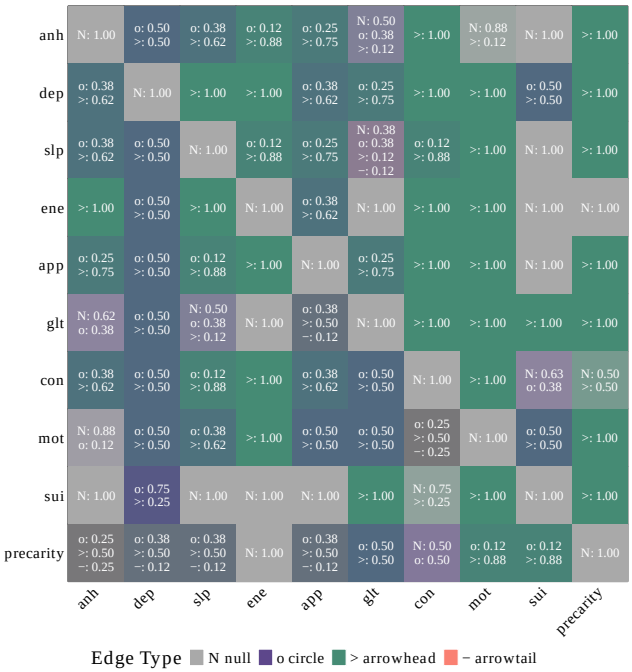


Figure 11: Proportions of edge endpoint types for graphs based on individual depression symptoms with precarity sum score