Technische Universität München
TUM School of Computation, Information and Technology

Structure Uncertainty in Causal Inference

David Maximilian Strieder

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr. Seyed Jalal Etesami

Prüfende der Dissertation:
1. Prof. Mathias Drton, Ph.D.
2. Prof. Dr. Vanessa Didelez
3. Prof. Y. Samuel Wang, Ph.D.

Die Dissertation wurde am 04.12.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 30.04.2025 angenommen.

Abstract

Inferring the effect of interventions within complex interacting systems is a fundamental challenge in statistics. In order to draw such causal conclusions from available data, it is crucial to reason about the underlying causal structure that governs the data-generating process. However, precise knowledge of this causal structure is missing in many practical applications. In this publication-based thesis, we tackle the challenge of rigorously accounting for this structure uncertainty in causal inference. Specifically, we introduce a test inversion approach that recasts the problem as a statistical testing problem and propose concrete solutions based on modified likelihood ratio theory within the context of linear structural causal models. Our framework enables the construction of calibrated confidence regions for total causal effects that rigorously capture all sources of remaining uncertainty: statistical uncertainty about the numerical size of involved effects and causal structure uncertainty stemming from the data-driven model choice as well as the potential equivalence of multiple plausible models. Moreover, motivated by the arising intricate statistical testing problem, we study the limit behavior of the employed split likelihood ratio test and suggest a new routine for selecting an optimal data splitting ratio that maximizes asymptotic power.

Zusammenfassung

Eine grundlegende Herausforderung der Statistik ist es, die Auswirkungen von Interventionen in komplexen, interagierenden Systemen vorherzusagen. Um solche kausalen Schlussfolgerungen aus verfügbaren Daten zu ziehen, ist es entscheidend, die zugrunde liegende kausale Struktur zu berücksichtigen. In vielen praktischen Anwendungen fehlt jedoch die genaue Kenntnis über eben diese kausale Struktur. In dieser publikationsbasierten Dissertation widmen wir uns der Herausforderung, diese Unsicherheit in der kausalen Struktur bei kausaler Inferenz rigoros zu berücksichtigen. Insbesondere stellen wir einen Testinversionsansatz vor, der das Problem als statistisches Testproblem neu formuliert, und schlagen konkrete Lösungen mithilfe von modifizierter Likelihood-Ratio-Theorie im Kontext linearer Kausalmodelle vor. Dieser Ansatz ermöglicht die Konstruktion kalibrierter Konfidenzbereiche für totale kausale Effekte, die alle Quellen verbleibender Unsicherheit rigoros erfassen: die statistische Unsicherheit über die numerische Größe der beteiligten Effekte und die Unsicherheit über die kausale Struktur, die aus der datengetriebenen Modellwahl sowie der potenziellen Äquivalenz mehrerer plausibler Modelle resultiert. Darüber hinaus untersuchen wir, motiviert durch das entstehende komplexe statistische Testproblem, das Grenzverhalten des verwendeten Split-Likelihood-Ratio-Tests und schlagen eine neue Routine zur Auswahl eines optimalen Daten-Splits vor, der die asymptotische Teststärke maximiert.

Acknowledgments

I want to express my sincere gratitude to everyone who has supported me throughout this journey. First and foremost, I am incredibly thankful to my supervisor, Mathias Drton, for his mentorship and guidance during my research. His expertise and constructive feedback were invaluable in shaping my ideas and helping me complete this project. I particularly appreciate his encouragement to attend conferences and workshops, which provided me with opportunities to engage with the academic community and broaden my perspective.

I would also like to thank my colleagues in the Mathematical Statistics group for creating a collaborative and enjoyable work environment. Their support and friendship have meant a great deal to me throughout this process. A special thanks goes to Andrea Grant for her assistance with administrative matters and for fostering a positive atmosphere in our group.

I am grateful for the financial support from the Munich Center for Machine Learning (MCML), which allowed me to fully concentrate on my research. The connections and learning opportunities offered by MCML have been highly beneficial.

Lastly, I am truly grateful to my parents and my partner, Nicole, for their patience, love, and encouragement. Their unconditional support has been essential throughout my studies, allowing me to wholeheartedly pursue my goals. I am forever thankful for their belief in me and for being there every step of this journey.

Contents

1 Index of Publications **1**

2 Introduction **2**

3 Causal Inference with Graphical Models **5**

3.1 Linear Structural Equation Models 5

3.2 Causal Perspective 7

3.3 Identifiability 9

4 Structure Uncertainty in Causal Inference **14**

4.1 Causal Discovery 14

4.2 Existing Approaches 17

4.3 Test Inversion Ansatz 18

5 Likelihood Ratio Tests **21**

5.1 Split Likelihood Ratio Test 21

5.2 Dual Likelihood Ratio Test 23

6 Conclusion **24**

7 Summary Core Publications **25**

7.1 Confidence in Causal Inference under Structure Uncertainty in Linear Causal Models with Equal Variances 25

7.2 On the Choice of the Splitting Ratio for the Split Likelihood Ratio Test 27

7.3 Dual Likelihood for Causal Inference under Structure Uncertainty 28

7.4 Identifying Total Causal Effects in Linear Models under Partial Homoscedasticity 29

Bibliography **30**

A Core Publications **34**

A.1 Confidence in Causal Inference under Structure Uncertainty in Linear Causal Models with Equal Variances 34

A.2 On the Choice of the Splitting Ratio for the Split Likelihood Ratio Test 62

A.3 Dual Likelihood for Causal Inference under Structure Uncertainty 86

A.4 Identifying Total Causal Effects in Linear Models under Partial Homoscedasticity 106

1 Index of Publications

Core publications as (sole) first author in journals:

1. Strieder, D. and Drton, M. (2023). Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *J. Causal Inference*, 11(1).

2. Strieder, D. and Drton, M. (2022). On the choice of the splitting ratio for the split likelihood ratio test. *Electron. J. Stat.*, 16(2):6631–6650.

Core publications as (sole) first author in conference proceedings:

3. Strieder, D. and Drton, M. (2024a). Dual likelihood for causal inference under structure uncertainty. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *PMLR*, pages 1–17.

4. Strieder, D. and Drton, M. (2024b). Identifying total causal effects in linear models under partial homoscedasticity. In *Proceedings of The 12th International Conference on Probabilistic Graphical Models*, volume 246 of *PMLR*, pages 213–230.

2 Introduction

What is the effect of COVID-19 vaccinations on the infection rate? Will raising the minimum wage decrease the employment rate? Which proteins regulate the expression of a specific gene of interest? Causal questions are the very essence of most scientific research. Across all disciplines, from medicine and biology to economics and social sciences, researchers and policymakers are interested in understanding the effects of external interventions in complex systems and aim to disentangle intricate cause-effect relations, see Sachs et al. (2005), Imbens and Rubin (2015), Runge et al. (2019) and Feuerriegel et al. (2024) for selected examples.

Every day, human brains can easily distinguish causes from effects and anticipate the consequences of actions to decide how we should act or react to our surroundings. Nevertheless, it is a fundamental statistical challenge to devise methods that go beyond describing mere correlation and reliably answer causal questions from available data. Classical statistical theory provides various powerful tools to model the associations between variables and detect significant correlations. For example, a standard linear regression model can leverage correlation into predictions, e.g., predicting a country's number of Nobel Prize winners based on its observed chocolate consumption (Messerli, 2012). However, along the lines of the well-known saying *correlation does not imply causation*, it is misleading to conclude from this statistically significant correlation that externally increasing chocolate consumption via policy changes would positively impact the number of Nobel Prize winners. In order to answer such causal questions about interventions in interacting systems, it is crucial to reason about the underlying causal structure that is behind the observed correlation. In this particular example, it seems more plausible to attribute the positive correlation between chocolate consumption and winning a Nobel Prize to a confounder, e.g., the wealth of a country, rather than a direct causal effect.

However, in many applied settings, concrete knowledge of the underlying causal structure that governs the data-generating process is missing. Therefore, researchers typically study causal questions via controlled scientific experiments guided by statistical theory in experimental design. In order to investigate the impact of specific actions we simply force those actions to be true and study the results compared to a neutral control group. In those settings, the causal perspective is modeled by the experimental design itself. In particular, randomized controlled trials are widely considered the gold standard, where randomness in the treatment assignment eliminates confounding and allows researchers to draw causal conclusions. While such controlled studies build the foundation of causal inference, in many research fields, conducting the necessary experiments and externally intervening in the system to collect experimental data may simply not be feasible due to ethical constraints, cost considerations, or technological limitations. Thus, the field of causal discovery addresses the challenge of answering causal questions from purely observational data. This means that the only available data is collected without the external influence of the researcher,

and thus, the causal perspective has to come from the statistical model itself (see, e.g., Spirtes et al., 2000; Pearl, 2009).

A widely studied approach to (statistically) model causal relations employs structural causal models that postulate noisy functional relations among a set of interacting variables. In this framework, each variable is modeled as a function of a subset of other variables (its causes) and a stochastic error term. The causal perspective emerges from interpreting these relations as assignments rather than mathematical equations, and thus, external variations on the right-hand side induce changes on the left-hand side, but not vice versa. This reflects the inherent asymmetry in cause-effect relations, which is key to distinguishing spurious correlation from actual causation. The causal structure underlying a structural causal model is naturally represented by a directed graph whose edges indicate direct causal dependencies.

An essential aspect of research in causal discovery is clarifying whether it is even theoretically possible to answer the causal inquiry of interest. In full generality, multiple differing underlying structural causal models may generate the same observational distribution and, thus, are indistinguishable even at the population level without conducting experiments. Therefore, additional assumptions are necessary to draw meaningful causal conclusions from observational data alone. Throughout this thesis, we primarily focus on structural assumptions on the functional relation between the interacting variables. Specifically, we follow a line of work initiated by Peters and Bühlmann (2014), which assumes linear relations disturbed by an additive Gaussian stochastic noise. In their influential work, the authors showed that the additional assumption of homoscedasticity among all noise variances is sufficient to render the complete underlying causal data-generating mechanism identifiable from the observational distribution alone.

The next challenge of causal discovery research is to develop practical algorithms that infer the underlying causal structure from a data sample of the observational distribution. Over the last decades, many structure learning algorithms have been proposed to tackle this challenge; see Drton and Maathuis (2017) for an overview or Peters et al. (2017) for recent advances in connection with the machine learning community. On the other hand, when the underlying causal structure is known, causal inference results allow researchers to estimate the involved causal quantities and assess the remaining uncertainty about the numerical size of the effects (see, e.g., Hernan and Robins, 2024). However, in the typical application, it is necessary to do both, that is, learn the underlying causal structure first from observational data and then apply causal inference methods within the inferred model. Thus, in order to draw reliable conclusions and make calibrated confidence statements about causal effects, it is crucial to incorporate uncertainty about the underlying causal structure into causal inference. The overarching theme of this thesis is how to mathematically rigorously account for this structure uncertainty in causal inference arising from the data-driven structure learning approach.

Contributions of the thesis. In total, this thesis comprises four core articles. In the first core article (Strieder and Drton, 2023), we introduce a framework based on test inversions to construct confidence regions for total causal effects when the underlying causal structure is unknown. We further explore this methodology in the third core article (Strieder and Drton, 2024a), where we employ a different strategy to solve the arising statistical testing problem in order to provide a closed-form solution to compute the proposed confidence regions. In the fourth core article (Strieder and Drton, 2024b), we investigate to

what extent the equal error variance assumption can be relaxed and extend the framework to the partially homoscedastic case.

Furthermore, our ideas for constructing confidence regions for causal effects under structure uncertainty lead to intricate statistical testing problems. While classical likelihood ratio tests provide a powerful tool for solving hypothesis tests, their validity depends on asymptotic approximations under regularity conditions (see, e.g., van der Vaart, 1998). When these conditions are not satisfied, the required distribution-theoretic insights may be difficult to obtain. The recently introduced framework of universal inference presents an intriguing alternative by employing data splitting to utilize a carefully chosen universal critical value instead of relying on asymptotic distribution theory (Wasserman et al., 2020). In the second core article of the thesis (Strieder and Drton, 2022), we expand our understanding of the limit behavior of the universal inference framework by deriving precise limit theory and propose an optimal choice of the data splitting ratio for tests of composite hypotheses of different dimensions.

Organization of the thesis. In Chapter 3, we review the relevant causal preliminaries and introduce the considered causal framework of structural causal models. Furthermore, we discuss the concept of identifiability and highlight the contributions of the fourth core article (Strieder and Drton, 2024b). Chapter 4 provides an overview of the existing approaches for causal inference without knowledge of the underlying causal structure and presents our test inversion ansatz. Moreover, we discuss our strategies to solve the arising statistical testing problem and illustrate the contributions of the first and third core article (Strieder and Drton, 2023, 2024a). Chapter 5 recapitulates the likelihood ratio test and introduces the various modifications used in the articles to solve statistical testing problems. Specifically, we introduce the split likelihood ratio test and outline the key contributions of the second core article (Strieder and Drton, 2022). In the concluding Chapter 6, we discuss potential extensions of our framework. Chapter 7 additionally provides (short) summaries of the four core articles of this publication-based thesis.

3 Causal Inference with Graphical Models

Graphical models constitute a valuable tool to study statistical dependencies in complex interacting systems and admit an intuitive interpretation as causal models. Specifically, we consider statistical models specified by a recursive system of (linear) structural equations, where the underlying (causal) dependency structure is naturally represented by a directed acyclic graph. In this chapter, we review structural equation models and their causal interpretation. Further, we recapitulate the key concepts of graphical models and causal inference relevant to this thesis.

3.1 Linear Structural Equation Models

A natural starting point to model noisy functional relations among a set of interacting random variables $\{X_i : i = 1, \dots, d\}$ is to describe each variable X_i as a function of a subset of the other variables and a stochastic noise term ε_i. In this work, we primarily focus on a specific functional form of the relationships among the variables. In particular, we assume linear dependencies which are disturbed by an additive Gaussian stochastic noise. To be precise, we statistically model the random vector $X = (X_1, \dots, X_d)$ as a solution to a system of linear equations

$$X_j = \sum_{i \neq j}^{d} \beta_{j,i} X_i + \varepsilon_j, \quad j = 1, \dots d, \tag{3.1}$$

where $B := [\beta_{j,i}]_{j,i=1}^{d}$ are unknown parameters representing the direct dependencies between the variables and the error terms are mutually independent, normally distributed random variables with unknown variance, i.e., $\varepsilon := (\varepsilon_1, \dots, \varepsilon_d) \sim \mathcal{N}(0, \mathrm{diag}(\omega))$, where $\omega = (\omega_1, \dots, \omega_d)$. We may assume, without loss of generality, that the mean vector is zero, as it does not convey any information about the underlying dependency structure. Furthermore, the assumption of mutually independent error terms implies that all relevant variables are included in the model. In the causal context, this (strong) assumption is referred to as *causal sufficiency*.

In the remainder of this thesis, we study the properties of statistical models defined by an underlying system of linear equations (3.1) and their associated dependency structure. In particular, in each specific *structural equation model*, a subset of the direct dependencies $\beta_{j,i}$ is constrained to be zero. Therefore, the underlying dependency structure of a structural equation model is naturally represented by a *directed graph* $G = (V, E)$, where each node i in the graph corresponds to a variable X_i (i.e., $V = \{1, \dots, d\}$), and the edge set E is equal to the support of B (i.e., a missing edge $i \to j$ indicates that $\beta_{j,i} = 0$).

Throughout this thesis, we focus on recursive systems, which entail that the matrix B is permutation similar to a lower triangular matrix or, similarly, that the naturally associated graph is *acyclic*. Furthermore,

the system (3.1) then admits the unique solution $X = (I_d - B)^{-1}\varepsilon$, where I_d denotes the $d \times d$ identity matrix. Consequently, the distribution of this solution follows a (centered) multivariate normal distribution, i.e., $X \sim \mathcal{N}(0, \Sigma)$, where the covariance matrix is given by

$$\Sigma = (I_d - B)^{-1}\mathsf{diag}(\omega)(I_d - B)^{-T}.$$

In light of the normality assumption, it is natural to parameterize such a Gaussian structural equation model in terms of the covariance matrix Σ. To be precise, every possible distribution $\mathcal{N}(0, \Sigma)$ of a solution to the considered structural equation model with underlying dependency structure G belongs to a covariance matrix from the set

$$\mathcal{M}(G) = \left\{\Sigma \in \mathsf{PD}(d) : \exists B \in \mathbb{R}^G, \omega \in \mathbb{R}^d_+ \text{ with } \Sigma = (I_d - B)^{-1}\mathsf{diag}(\omega)(I_d - B)^{-T}\right\}. \tag{3.2}$$

Here, $\mathsf{PD}(d)$ denotes the cone of positive definite matrices and $\mathbb{R}^G$ represents the set of all linear relations conform with G, that is, $\mathbb{R}^G = \{B \in \mathbb{R}^{d\times d} : \beta_{j,i} = 0 \text{ if } i \to j \notin E\}$.

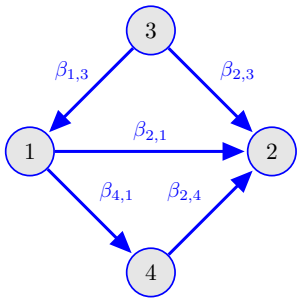
In order to study the underlying dependence structure G, it can be helpful to adopt a graphical perspective (see, e.g., Lauritzen, 1996; Maathuis et al., 2019). From a graphical standpoint, the missing edges in a directed acyclic graph (DAG) G encode a collection of conditional independence statements, which can be read off the graph using d-separation criteria. Consequently, we can define a Gaussian *graphical model* on a DAG G as the family of (centered) multivariate normal distributions that are *Markov* to the DAG, that is, distributions satisfying the (graphically) implied conditional independence relations. Furthermore, we highlight that a Gaussian graphical model can be equivalently defined as the family of all (centered) multivariate normal distributions that factorize according to the DAG.

We emphasize that both perspectives, structural equation models and graphical models, are intertwined in the sense that the joint distribution of the solution to a structural equation model is Markov to its naturally implied DAG and, vice versa, every distribution satisfying the implied conditional independence relations of a DAG G occurs as the distribution of the solution to a structural equation model with the underlying dependence structure G. In fact, the Gaussian graphical model on a DAG G and the linear Gaussian structural equation model (3.2) with dependency structure G define the same set of distributions. Therefore, we will use both approaches and terms interchangeably throughout this thesis. However, note that generally structural equation models encapsulate more information than their corresponding graph and distribution, which, in the causal context, enables counterfactual reasoning.

Furthermore, under the assumption of an underlying linear structural equation model (3.1), the covariance matrix has a specific structure that is defined by a polynomial map (3.2). Similarly, the collection of conditional independence statements implied by a Gaussian graphical model corresponds to polynomial constraints on the set of admissible covariance matrices. This correspondence translates the problem into an algebraic setting, enabling the use of algebraic geometry tools to study properties such as identifiability and equivalence of structural equation models and, thus, provides another perspective for analyzing the dependency structure in complex interacting systems through the lens of algebraic statistics (see, e.g., Sullivant, 2018; Drton, 2018).

$$
\begin{aligned}
X_1 &= \beta_{1,3} X_3 + \varepsilon_1 \\
X_2 &= \beta_{2,1} X_1 + \beta_{2,3} X_3 + \beta_{2,4} X_4 + \varepsilon_2 \\
X_3 &= \varepsilon_3 \\
X_4 &= \beta_{4,1} X_1 + \varepsilon_4
\end{aligned}
$$

(a) A linear structural equation model.



(b) The corresponding directed acyclic graph.

Figure 3.1 Example of a linear structural equation model to study a system of 4 interacting variables.

Example 3.1. We consider the linear structural equation model in Figure 3.1a with independent (centered) Gaussian errors, i.e., $\varepsilon_i \sim \mathcal{N}(0, \omega_i)$, and the corresponding naturally associated DAG representing the underlying dependency structure in Figure 3.1b as a running example for the remainder of the thesis. To illustrate the structure and parametrization of the covariance matrix of the model (3.2) we display

$$\Sigma_{1,2} = (\beta_{2,1} + \beta_{2,4}\beta_{4,1})\omega_1 + \beta_{1,3}\left(\beta_{2,3} + \beta_{1,3}(\beta_{2,1} + \beta_{2,4}\beta_{4,1})\right)\omega_3.$$

3.2 Causal Perspective

While structural equation models can be used to study complex statistical associations among variables, their true potential lies in their ability to capture causal relationships. Thus, adopting a causal perspective when specifying and interpreting structural equation models is crucial to fully leverage their capabilities.

The causal perspective of structural equation models arises when viewing the equations in (3.1) as making assignments rather than mathematical equalities. In particular, in the causal interpretation, the structural equations do not describe mere statistical correlations but specify the underlying causal mechanism that generates the system. Specifically, the left-hand side of each equation, the effect, is assigned the value determined by the right-hand side, the causes. This captures the inherent asymmetry in cause-effect relations: externally varying the causes may change the effect, but interventions on the effect do not alter its causes. The distinction between cause and effect makes *structural causal models* a powerful tool for understanding how interventions propagate through interacting systems and allows researchers to differentiate causal relationships from mere associations. We use the term structural causal model to emphasize the causal interpretation of structural equation models.

Interventions, in this context, refer to deliberate actions that externally change the value of a variable in the system. Pearl (2009) formalized this concept and introduced the do-notation to denote a (hard) intervention that externally sets the variable X_i to the value x_i as do$(X_i = x_i)$. Formally, an intervention constitutes a change in the true data-generating process and, thus, a modification in our probabilistic model. Therefore, in contrast to conditional distributions that can be seen as focusing on specific parts of the joint distribution of the observed system, interventional distributions, in theory, can be arbitrarily distinct from the observed distribution. However, the key aspect implicit in the causal interpretation of

structural causal models as causal mechanisms is the assumption that these mechanisms are invariant and independent in the sense that under an external intervention the causal mechanisms not intervened upon stay the same. Thus, the interventional distribution under the intervention do$(X_i = x_i)$ can be modeled as the distribution of the solution to a modified system of structural equations by replacing the i-th structural equation with $X_i = x_i$ (see Example 3.2). Using the graphical modeling language of structural causal models, this assumption is called the *causal Markov property*, and the underlying DAG is referred to as a causal DAG.

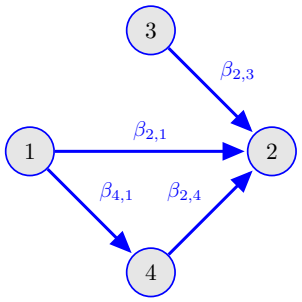
Specifically, the causal Markov property implies that the interventional distribution under the intervention do$(X_i = x_i)$ factorizes according to a mutilated DAG, where all incoming edges into X_i have been deleted. Thus, such a mutilated DAG represents the situation where X_i was drawn independently from the other variables in the system while all other relations remain unchanged. We emphasize the link to randomized controlled trials, which are widely regarded as the gold standard in scientific research. Randomization in these trials breaks the natural associations, enabling researchers to draw causal conclusions. Specifically, through random assignment, the treated variables do not depend on other variables in the system. As a result, randomized controlled trials mimic sampling from an interventional distribution that factorizes according to the mutilated DAG.

Finally, we emphasize that structural causal models are not only capable of answering interventional questions, i.e., what would happen if X_i is externally set to x_i, but also counterfactual questions, i.e., what would have happened had X_i been x_i. Counterfactual reasoning is essential for decision-making and policy analysis, where we are interested in evaluating hypothetical scenarios in retrospect.

Example 3.2. We reconsider the setup of the running Example 3.1. The intervention do$(X_1 = x_1)$ that externally sets the value of X_1 to x_1 corresponds to replacing the first structural equation with $X_1 = x_1$ while the remaining structural equations are left unchanged. Thus, the interventional distribution is given as the distribution of the solution to the modified system of structural equations in Figure 3.2a and is Markov to the mutilated DAG in Figure 3.2b.

$$
\begin{aligned}
X_1 &= x_1 \\
X_2 &= \beta_{2,1} X_1 + \beta_{2,3} X_3 + \beta_{2,4} X_4 + \varepsilon_2 \\
X_3 &= \varepsilon_3 \\
X_4 &= \beta_{4,1} X_1 + \varepsilon_4
\end{aligned}
$$

(a) Structural equations under do$(X_1 = x_1)$.



(b) The corresponding mutilated DAG.

Figure 3.2 The intervention do$(X_1 = x_1)$ in the running Example 3.1.

3.3 Identifiability

In the remainder of the thesis, the primary target of interest is the total causal effect of a variable X_i onto another variable X_j. Informally put, we are interested in estimating how much the value of X_j changes if we externally intervene in the system and set the value of X_i to x_i. Using the do-notation to indicate the external intervention, we can formally define the total causal effect $\mathcal{C}(i \to j)$ as the unit change in the expectation of X_j under the interventional distribution, that is,

$$\mathcal{C}(i \to j) := \frac{\mathsf{d}}{\mathsf{d}x_i} \mathbb{E}\left[X_j | \ \mathsf{do}(X_i = x_i)\right].$$

We emphasize that this causal quantity is formally a characteristic of the interventional distribution. In many research fields, however, access to interventional data is limited, i.e., conducting the necessary experiments for a randomized controlled trial is not feasible due to ethical concerns, financial constraints, or technological limitations. Consequently, a central question in causal inference is determining whether it is theoretically possible to identify a causal target of interest using observational data alone. However, as the causal quantities are not inherently defined by the observational distribution, their estimation from observational data must rely on (untestable) assumptions about the causal structure.

Throughout this work, we assume the causal Markov property, introduced in Section 3, that provides a link between observational and interventional distribution. Given this assumption of an underlying causal DAG, graphical criteria exist that facilitate the translation between these distributions and allow us to identify the causal effects of interest through appropriate adjustments. One such criterion is the *back-door criterion* (Pearl, 1993), which gives graphical criteria to specify a set of variables that blocks all back-door paths, that is, paths that could potentially lead to confounding. Thus, the back-door criterion enables the identification of causal effects from observational data alone by conditioning on such a valid adjustment set, provided the underlying causal DAG is known and correctly specified. We highlight that the parent set $pa(i)$ of the node i, that is, the set of all nodes k such that the underlying DAG contains a directed edge from node k to node i, is always a valid adjustment set for the total causal effect $\mathcal{C}(i \to j)$. Thus, using the back-door criterion, we can translate between interventional and observational distribution and write the causal effect of interest as the (ordinary) conditional expectation

$$\mathcal{C}(i \to j) = \frac{\mathsf{d}}{\mathsf{d}x_i} \mathbb{E}\left[X_j | \ \mathsf{do}(X_i = x_i)\right] = \frac{\mathsf{d}}{\mathsf{d}x_i} \mathbb{E}\left[\mathbb{E}\left[X_j | X_i = x_i, X_{pa(i)}\right]\right]. \tag{3.3}$$

Furthermore, we emphasize that while the back-door criterion is a sufficient condition for identification, it is not necessary. In particular, other (graphical) criteria, such as the *front-door criterion* (Pearl, 1995), also provide sufficient conditions to identify causal effects from observational data, which highlights the power and flexibility of graphical approaches in causal inference.

In this thesis, we primarily focus on an underlying linear structural causal model (3.1). From an algebraic perspective, the question of identifiability, in this case, corresponds to investigating the injectivity of the parametrization map given in (3.2). Moreover, the total causal effect can be intuitively read off the underlying DAG and the corresponding edge weights, i.e., the effect $\mathcal{C}(i \to j)$ is given by the sum of the products of edge weights along all directed paths from node i to node j. If no such directed path exists, the causal effect is zero. This result follows directly from the path properties of the matrix $(I_d - B)^{-1}$ that

computes the solution X of the structural equations from the independent errors. Moreover, considering Equation (3.3) with the parent set $pa(i)$ as an valid adjustment set, the total causal effect $\mathcal{C}(i \to j)$ is given as the regression coefficient of X_i when regressing X_j on $(X_i, X_{pa(i)})$, that is,

$$\mathcal{C}(i \to j) = \Sigma_{j,i|pa(i)} \left(\Sigma_{i,i|pa(i)}\right)^{-1}. \tag{3.4}$$

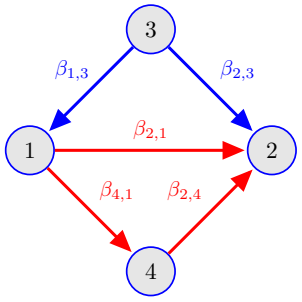
Here, and in the remainder of this thesis, $\Sigma_{j,i|pa(i)}$ denotes the conditional covariance matrix, namely,

$$\Sigma_{j,i|pa(i)} := \Sigma_{j,i} - \Sigma_{j,pa(i)} \left(\Sigma_{pa(i),pa(i)}\right)^{-1} \Sigma_{pa(i),i}.$$

In theory, any valid adjustment set can be used to identify the total causal effects of interest from the observational distribution and derive the Equations (3.3) and (3.4) as described. However, for simplicity, we focus on the parent set for the remainder of the thesis. Nevertheless, it is important to note that while the parent set is always a valid adjustment set for identifying the total causal effects, it may not always be the most efficient choice for estimation (see, e.g., Henckel et al., 2022).

Example 3.3. We reconsider the setup of the running Example 3.1 and compute the total causal effect $\mathcal{C}(1 \to 2)$ in Figure 3.3a, which intuitively corresponds to the sum of all products of edge weights along all directed paths from node 1 to node 2, highlighted in red in the corresponding DAG in Figure 3.3b. Adjusting for the parent set $pa(1) = \{3\}$ blocks the blue back-door path $1 \leftarrow 3 \to 2$, and thus, eliminates the influence of the confounder X_3. We note that this is the only valid adjustment set in this example.

$$\begin{aligned}\mathcal{C}(1 \to 2) &= \beta_{2,1} + \beta_{4,1}\beta_{2,4} \\ &= \Sigma_{2,1|3} \left(\Sigma_{1,1|3}\right)^{-1}.\end{aligned}$$



(a) The total causal effect $\mathcal{C}(1 \to 2)$. **(b)** The corresponding edges in the DAG.

Figure 3.3 The total causal effect $\mathcal{C}(1 \to 2)$ in the running Example 3.1.

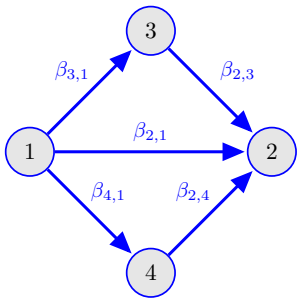
In Equation (3.4) we demonstrated that if the underlying causal DAG G is known, the causal parameter of interest can be expressed as a simple function of the covariance matrix $\varphi(G, \Sigma) = \Sigma_{j,i|pa(i)} \left(\Sigma_{i,i|pa(i)}\right)^{-1}$. However, difficulties arise in practice when the underlying causal structure is unknown, or more precisely, a valid adjustment set is unknown. A well-known result states that, in general, this causal structure, represented by the underlying DAG, can, at best, be identified up to a Markov equivalence class of indistinguishable models unless one makes additional structural assumptions (see, e.g., Drton, 2018). In particular, from the graphical perspective, each DAG naturally encodes a collection of conditional independence statements. Two DAGs are called *Markov equivalent* if they imply the same set of conditional independencies and, as a result, yield the same set of possible distributions. Furthermore, graphical criteria exist to decide whether two DAGs are Markov equivalent, and additional *faithfulness* assumptions

commonly guarantee at least being able to identify the Markov equivalence class by assuming that the given distribution does not satisfy additional conditional independence constraints, which are not implied by d-separation relations in the (true) underlying DAG. However, Markov equivalent DAGs do not necessarily entail the same causal effect of interest. Thus, without imposing further structural assumptions, we can, at best, identify the causal effect up to the Markov equivalence class, meaning we can identify a set of possible values for the causal parameter, each corresponding to a different DAG within the Markov equivalence class.

This raises a crucial question: Given an observed distribution P_Σ from a statistical model, that is, a set of possible distributions $\{P_\Sigma : \Sigma \in \mathcal{M}\}$, can we uniquely recover the causal parameter of interest $\varphi(G, \Sigma)$? More specifically, the question is under which assumptions on the statistical model $\mathcal{M}$ is it theoretically possible to identify the underlying causal structure G from observational data alone. The field of causal discovery provides various frameworks that address this question by restricting the functional class of the relations between the interacting variables. Notable examples are the assumption of linear relations disturbed by an additive non-Gaussian noise (Shimizu et al., 2006), or the assumption of nonlinear relations disturbed by an additive Gaussian noise (Hoyer et al., 2008) with the post-nonlinear model as a generalization (Zhang and Hyvärinen, 2009).

Example 3.4. We consider the linear structural causal model in Figure 3.4a with independent (centered) Gaussian errors, i.e., $\varepsilon_i \sim \mathcal{N}(0, \omega_i)$, and the corresponding DAG in Figure 3.4b. Using graphical criteria, we can easily check that this structural causal model yields the same set of possible distributions as the running Example 3.1. To be precise, the DAG in Figure 3.4b has the same skeleton and the same unshielded colliders as the DAG in Figure 3.1b corresponding to the running Example, and thus, both DAGs are Markov equivalent (see, e.g., Maathuis et al., 2019, for details on this classical result). Therefore, without additional assumptions, we cannot distinguish whether a given observed distribution P_Σ was generated by the structural causal model in Figure 3.4a or by the structural causal model in Figure 3.1a.

$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \beta_{2,1} X_1 + \beta_{2,3} X_3 + \beta_{2,4} X_4 + \varepsilon_2 \\
X_3 &= \beta_{3,1} X_1 + \varepsilon_3 \\
X_4 &= \beta_{4,1} X_1 + \varepsilon_4
\end{aligned}
$$



(a) A linear structural causal model. **(b)** The corresponding directed acyclic graph.

Figure 3.4 A linear structural causal model that is Markov equivalent to the running Example 3.1.

We emphasize that the assumption of an underlying (general) linear Gaussian structural causal model, as introduced in (3.1), is not sufficient to identify the (true) causal DAG from observational data. To be precise, under the assumption of an underlying linear Gaussian structural causal model, every possible distribution is given by the statistical model $\mathcal{M} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}(G)$, where $\mathcal{G}(d)$ denotes the set of all DAGs

on d nodes and $\mathcal{M}(G)$ as given in Equation (3.2). In this case, an underlying complete DAG does not impose any (additional) restrictions on the set of possible distributions, meaning that every complete DAG could have generated any given observed distribution. Consequently, any given observational distribution is compatible with multiple DAGs (at least all complete DAGs), rendering the causal structure unidentifiable without further assumptions.

For the remainder of the thesis, we adopt the following frequently used structural assumption that resolves this issue of identifiability. Peters and Bühlmann (2014) introduced a line of research that assumes homoscedasticity among the error variances, that is, assuming $\omega_i = \omega_j$ for all $i, j = 1, \ldots, d$, to render the underlying causal DAG identifiable. In particular, assuming homoscedasticity provides a simple criterion to identify the underlying causal order of the DAG. Since the external error variances are equal, the true source nodes in the DAG have minimal variance. Consequently, the underlying causal order is implied by recursively ordering conditional variances (see, e.g., Chen et al., 2019; Gao et al., 2022). Algebraically, the homoscedasticity assumption imposes the constraint that the variance of each node conditioned on its parents has to be equal. Thus, in the remainder of thesis, we work with the statistical model $\mathcal{M}_{EV} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{EV}(G)$, where $\mathcal{M}_{EV}(G)$ represents the set of possible distributions for a given DAG G under homoscedasticity, that is,

$$\mathcal{M}_{EV}(G) = \Big\{ \Sigma \in \mathcal{M}(G) : \exists \omega > 0 \text{ with } \omega = \Sigma_{k,k|pa(k)} \quad \forall\, k = 1, \ldots, d \Big\},$$

where ω corresponds to the common (unknown) error variance.

In the fourth core article (Strieder and Drton, 2024b), we explored the question of identifiability in linear Gaussian structural causal models in greater detail, specifically examining how the equal error variance restriction can be relaxed when practitioners are primarily interested in the effects of specific interventions, rather than identifying the entire causal structure. We concluded that the assumption of partial homoscedasticity, that is, equal error variances (only) between the potential cause X_i and response X_j of interest, is sufficient to *generically identify* the total causal effect. This means that we can uniquely recover the total causal effect of interest $\mathcal{C}(i \to j)$ for almost all observational distribution from the statistical model $\mathcal{M}_{PEV} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{PEV}(G)$, where $\mathcal{M}_{PEV}(G)$ represents the set of possible distributions for a given DAG G under partial homoscedasticity, namely,

$$\mathcal{M}_{PEV}(G) = \Big\{ \Sigma \in \mathcal{M}(G) : \Sigma_{i,i|pa(i)} = \Sigma_{j,j|pa(j)} \Big\}.$$

Our result extends the applicability of the equal variance framework to cases where the potential cause and response of interest are from similar domains or originate from measurements taken by the same machine, but not the entire interacting system.

Example 3.5. We reconsider the setup of the running Example 3.1. In this setting, the total causal effect $\mathcal{C}(1 \to 2)$ is not identifiable without additional assumptions. For example, the Markov equivalent DAG given in Example 3.4 implies the total causal effect $\mathcal{C}(1 \to 2) = \Sigma_{2,1} \left(\Sigma_{1,1}\right)^{-1}$ from X_1 onto X_2, whereas the original (indistinguishable) model implies $\mathcal{C}(1 \to 2) = \Sigma_{2,1|3} \left(\Sigma_{1,1|3}\right)^{-1}$, see Figure 3.3a. In the fourth core article (Strieder and Drton, 2024b), we showed that assuming equal error variances between X_1 and X_2, i.e., $\omega_1 = \omega_2$, suffices to generically identify the total causal effect $\mathcal{C}(1 \to 2)$. Nevertheless,

we emphasize that in the article, we were also able to explicitly construct “unfaithful” distributions in the intersection between two causal models that would imply not only different magnitudes of the causal effect of interest but also different directions of the causal relation. Thus, not even the causal ordering of the potential cause and response is globally identifiable. However, if all error variances are assumed to be equal, then the causal effect (and the full causal model) is globally identifiable.

4 Structure Uncertainty in Causal Inference

In the previous chapter, we explored the theoretical foundations of structural causal models and outlined when the underlying causal structure is theoretically identifiable at the population level. However, in practice, we only have access to finite samples from the observational distribution, raising a key question: How can we reliably infer the causal quantity of interest when the underlying causal structure must be learned from available data? In this chapter, we highlight established causal discovery algorithms, explain why "naïve" approaches fail, and introduce our test inversion framework to construct rigorous confidence regions for total causal effects without requiring prior knowledge of the underlying causal structure.

4.1 Causal Discovery

We consider a typical application where we are interested in the total causal effect $\mathcal{C}(i \to j)$ that an external intervention on variable X_i has onto variable X_j. More precisely, to draw reliable conclusions, we aim to construct a confidence interval for this causal quantity of interest, that is, in the technical sense, a region that guarantees a desired frequentist coverage probability of the (true) total causal effect. The challenge is that we typically only have access to observational data consisting of n independent copies $X^1, \dots, X^n$ of the random vector $X = (X_1, \dots, X_d)$. Importantly, we cannot perform additional experiments to study the causal relationships directly, and there is not enough expert domain knowledge available to justify assuming a specific underlying DAG priori. However, since knowledge of the causal structure is essential for causal inference, the underlying causal DAG, or more specifically, at least a valid adjustment set, has to be learned from available data.

Over the last decade, numerous structure learning algorithms have been developed to tackle this challenge of inferring the causal DAG from observational data alone. In the following, we highlight some well-established examples. However, again, we emphasize that, in general, the underlying causal DAG can only be identified up to a Markov equivalence class of indistinguishable models. Thus, under the assumption of faithfulness, constraint-based approaches, i.e., the Peter Spirtes and Clark Glymour (PC) algorithm (Spirtes and Glymour, 1991), use conditional independence tests to estimate this Markov equivalence class from observational data. Specifically, the faithfulness assumption guarantees that all conditional independencies in the observational distribution arise solely due to d-separation relations in the true underlying causal DAG. The PC algorithm leverages these conditional independencies to determine the skeleton of the underlying causal structure and, subsequently, applies a set of orientation rules to estimate the underlying Markov equivalence class, represented by a completed partially directed acyclic graph (CPDAG). Moreover, score-based approaches, i.e., the Greedy Equivalence Search (GES) (Chickering, 2002), learn the underlying causal structure by optimizing a specified score. The GES algorithm, for

example, greedily searches over the space of equivalence classes in two phases, a forward edge insertion phase, and a backward edge deletion phase, to estimate the score-optimal Markov equivalence class.

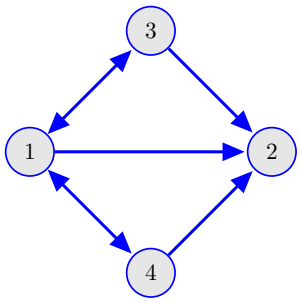
In Section 3.3, we showed that, in the context of structural causal models, restricting the functional form of the relationships among the variables allows us to identify the underlying causal structure beyond Markov equivalence classes. In particular, under the assumption of a linear Gaussian structural causal model with homoscedastic error variances, it is theoretically possible to uniquely identify the individual causal DAG from the observational distribution. Using this insight, specific assumption-tailored causal discovery algorithms, i.e., the greedy directed acyclic graph search with equal error variances (GDS) (Peters and Bühlmann, 2014), leverage the functional restrictions of the underlying structural causal model to infer the causal DAG from available data. For example, the GDS algorithm uses edge insertions, deletions, and reversals to greedily search over the space of DAGs and maximize the BIC score under the assumption of homoscedasticity.

Furthermore, we highlight that in a separate side project, not part of this thesis, we proposed a novel causal discovery algorithm for the post-nonlinear causal model (Keropyan et al., 2023). The core idea of this approach is to employ rank-based methods to learn the functional relationships between the variables and then test the independence of residuals and predictors to infer the causal structure.

Example 4.1. In this example, we perform a small simulation study to demonstrate how challenging the task of causal discovery from observational data is. In particular, we generate 1000 synthetic data sets based on the running Example 3.1 with standard normal distributed errors and randomly selected edge weights according to a normal distribution $\mathcal{N}(\beta, 0.1)$ for a range of different average direct effects β and sample sizes n. Then, we infer the underlying Markov equivalence class (see Figure 4.1b) using various causal discovery algorithms (PC, GES, and GDS) implemented via the R package pcalg (Kalisch et al., 2012). It is important to note that the GDS algorithm essentially solves a more complex task by searching over the DAG space, however, it explicitly uses the (satisfied) equal variance assumption. Figure 4.1a shows the proportion of times the algorithm perfectly infers the Markov equivalence class of the underlying causal structure. Even in this small system of four interacting variables, the causal discovery algorithms often fail to learn the exact underlying causal structure, especially for weak causal relations. This highlights the importance of rigorously addressing structure uncertainty in causal inference.

n	500		1000	
method$\backslash\beta$	0.1	0.5	0.1	0.5
PC	0.00	0.38	0.01	0.58
GES	0.00	0.26	0.02	0.49
GDS	0.01	0.76	0.03	0.87

(a) Proportion of times true CPDAG is inferred.



(b) The true underlying CPDAG.

Figure 4.1 The performance of causal discovery algorithms using synthetic data based on the running Example 3.1.

Without knowledge of the underlying causal structure, a commonly used "naïve" approach to construct confidence intervals for total causal effects from observational data alone involves splitting the task into two steps. In the first step, a suitable causal discovery algorithm is used to estimate the underlying causal structure. Then, in the second step, confidence intervals for the causal parameters are calculated with classical statistical inference methods, i.e., Wald-type confidence intervals, within the inferred model. However, the resulting confidence intervals are inherently agnostic to structure uncertainty and invalid if the causal learning algorithm infers the wrong causal structure. For instance, if the algorithm incorrectly infers a causal DAG where node j precedes node i, the two-step method would report with certainty yet wrongly a confidence interval containing only zero, regardless of the chosen significance level. Thus, this "naïve" approach implicitly conditions away the uncertainty stemming from the data-driven model choice.

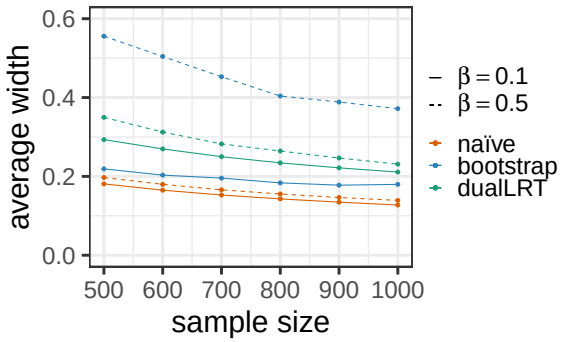
In order to properly account for this structure uncertainty in the calculations of a valid confidence interval, it is necessary to take all possible underlying structures into account. However, as a result, the causal target of interest has a complicated composite structure, and thus, obtaining sample distributions of estimators is difficult. In such cases, bootstrapping and other resampling methods are often employed to circumvent the technical difficulties posed by intractable distribution theory of estimators. However, for our considered problem, these methods fail due to singularities in the model space where multiple causal models intersect. Specifically, bootstrapping methods may fail drastically when the test statistic, here the composition of a consistent model selection method that learns the graph and a consistent estimator of the causal effect, lacks smoothness (Andrews and Guggenberger, 2010; Drton and Williams, 2011). Consequently, bootstrapping cannot be used to capture structure uncertainty correctly.

In the remainder of this chapter, we provide an overview of methods that circumvent the issues caused by the non-smooth nature of the total causal effect and may offer a solution to causal inference under structure uncertainty. In particular, in Section 4.3, we introduce our approach to construct mathematically rigorous confidence regions for total causal effects when the causal structure is unknown.

Example 4.2. We reconsider the setup of the simulation study in Example 4.1 and compare the empirical coverage probabilities of different approaches to construct confidence regions for causal effects under structure uncertainty. Specifically, for each of the 1000 synthetic data sets, we compute a 95%-confidence region for the total causal effect $\mathcal{C}(1 \to 2)$ using a "naïve" two-step approach based on Wald-type confidence intervals of (linear) regression parameters within the inferred model by the GDS algorithm, using a bootstrapping approach based on repeated runs of the GDS algorithm on resampled data, and using our proposed framework based on test inversions of (dual) likelihood ratio tests (see Section 4.3 and the core articles (Strieder and Drton, 2023, 2024a)). Figure 4.2a shows the proportion of times the constructed confidence regions cover the true total causal effect. The "naïve" two-step approach and the (classical) bootstrapping approach fail to achieve the desired coverage frequency under high structure uncertainty, i.e., if the causal relations are weak ($\beta = 0.1$). In contrast, our proposed framework (dualLRT) always yields valid confidence regions that hold level even under high structure uncertainty. Moreover, in Figure 4.2b, we compare the performance in terms of information contained in the confidence regions by displaying the average width of the non-zero part of the confidence sets against the sample size. Even in situations with low structure uncertainty ($\beta = 0.5$), where bootstrapping achieves the desired coverage,

n	500		1000	
method$\backslash\beta$	0.1	0.5	0.1	0.5
naïve	0.40	0.82	0.45	0.88
bootstrap	0.82	0.96	0.84	0.96
dualLRT	0.99	0.99	0.99	0.99

(a) Proportion of times true causal effect is covered.



(b) Mean width of non-zero part of confidence intervals.

Figure 4.2 The performance of different approaches to construct 95%-confidence intervals for the total causal effect $\mathcal{C}(1 \to 2)$ under structure uncertainty using synthetic data based on the running Example 3.1.

our method yields more conclusive results. Thus, our proposed method constitutes a valuable tool for reasoning about causal effects while correctly accounting for remaining uncertainty.

4.2 Existing Approaches

In the previous chapter, we highlighted the growing number of causal discovery methods that estimate causal structures from observational data. Moreover, when the underlying causal structure is known, causal inference results enable researchers to estimate causal quantities in complex systems and assess uncertainty about the numerical size of effects. However, when the underlying causal structure is learned from data, the remaining structure uncertainty must be incorporated into causal inference in order to draw reliable conclusions from causal estimates, especially considering the difficulty of the task, see Example 4.1. Despite the popularity and growing interest in causality, the question of how to bridge this gap between causal discovery and causal inference has only recently gained attention. Before introducing our solution, in the following, we highlight and compare newly arising approaches that tackle this issue.

We emphasize that, in general, structure uncertainty is not only aleatoric due to the data-driven model choice but also epistemic due to multiple plausible models being indistinguishable (Markov equivalent). The goal is to construct reliable confidence intervals that consider all sources of uncertainty and cover all indistinguishable effects with the desired frequency. When the causal effect of interest is identifiable only up to a Markov equivalence class, a natural approach is to target the set of effects corresponding to each DAG in the equivalence class, thereby accounting for structure uncertainty stemming from model equivalence, e.g., as in the IDA-framework by Maathuis et al. (2009).

In order to respect the uncertainty that arises due to learning the underlying causal structure from data, in recent work, Gradu et al. (2024) treat the task as a classical post-selection inference problem. Specifically, to avoid the double-dipping problem that occurs when both the (causal) model and the target of interest are estimated from one data set, they propose to use a noisy version of the GES algorithm. Introducing randomness into the process of causal discovery reduces the bias caused by reusing data and, thus, enables valid inference after model selection. However, to treat the problem as a post-selection task,

Gradu et al. (2024) view the causal target of interest as a functional of the selected DAG. Thus, their target may be viewed as a projection of the true underlying causal effect on the selected model, but it potentially differs from the true effect. In contrast, we aim to construct a valid confidence interval for the true underlying effect, irrespective of the selected model.

Chang et al. (2024) share our perspective and propose the following two-step approach to construct a confidence interval for the total causal effect under structure uncertainty. Their approach also introduces randomness into the structure learning process (here, the PC algorithm). However, they propose to leverage the randomness in the algorithm to perform causal discovery multiple times to compute a set of plausible causal models and then construct confidence regions as the union of individual confidence intervals for each selected DAG. With practical application in mind, this approach can build upon existing causal discovery algorithms and provides a flexible method to account for structure uncertainty in causal inference by selecting multiple plausible models. Nevertheless, it relies on a chosen tuning parameter that implicitly controls the number of considered causal models. In contrast, our goal is to rigorously construct confidence regions that mathematically capture the full extent of the remaining uncertainty without relying on a tuning parameter.

Wang et al. (2023) propose a mathematically rigorous two-step approach that disentangles the two tasks of accounting for structure uncertainty and statistical uncertainty about the numerical size of the effects. Specifically, they employ a test inversion approach of goodness-of-fit tests to construct a valid confidence set of plausible causal orderings of the involved variables by leveraging the independence of the noise term in an additive structural causal model and (true) causal ancestors. Subsequently, their idea is to construct a confidence region as the union of individual confidence intervals within each selected order. We aim to address both tasks jointly to fully reflect the remaining uncertainty in causal effect estimation, where precise knowledge of the underlying DAG or causal order is not necessary.

4.3 Test Inversion Ansatz

In the remainder of this chapter, we introduce our ideas to circumvent the problems posed by the non-smooth nature of the total causal effect and illustrate how to construct mathematically rigorous confidence intervals for causal effects under structure uncertainty. The core idea of our framework is to leverage the classical duality between statistical hypothesis tests and confidence regions. Specifically, if we perform valid hypothesis tests for all attainable causal effects, then all values that we cannot reject form a confidence region for the total causal effect (see, e.g., Casella and Berger, 1990).

The first ideas for using test inversions to construct confidence regions for causal effects emerged in a conference paper (Strieder et al., 2021), which is not directly part of this thesis, where we considered various approaches for simple bivariate models. In the first core article (Strieder and Drton, 2023), we refined and generalized this test inversion approach to multivariate systems. Under the assumption of an underlying linear Gaussian structural causal model with equal error variances, this test inversion ansatz

of inverting joint tests for causal structure and effect size leads to the following statistical testing problem, which we must solve for all $\psi \in \mathbb{R}$, namely,

$$H_0^{(\psi)} : \Sigma \in \mathcal{M}_{EV}^{\psi} \quad \text{against} \quad H_1 : \Sigma \in \mathcal{M}_{EV} \backslash \mathcal{M}_{EV}^{\psi}, \tag{4.1}$$

where $\mathcal{M}_{EV}^{\psi} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{EV}^{\psi}(G)$ with $\mathcal{M}_{EV}^{\psi}(G) := \big\{\Sigma \in \mathcal{M}_{EV}(G) : \mathcal{C}(i \to j) = \psi\big\}$.

The test inversion ansatz transforms the task of accounting for structure uncertainty in causal inference into a concrete statistical testing problem. Nevertheless, solving this testing problem remains challenging. We have to test a composite union hypothesis $\mathcal{M}_{EV}^{\psi} = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{EV}^{\psi}(G)$ within a model space that itself is an intricate union over graphs $\mathcal{M}_{EV} = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{EV}(G)$ and, notably, is not equal to the entire cone of positive definite matrices.

In the first core article (Strieder and Drton, 2022), we propose two concrete solutions based on likelihood ratio theory. The first solution addresses the challenge of both the null hypothesis and alternative being intricate unions that do not satisfy typical regularity conditions by employing intersection-union theory and stochastic upper bounds. Specifically, we compare the likelihood ratio test statistic for each single hypothesis $\mathcal{M}_{EV}^{\psi}(G)$ to quantiles of the limit distribution of a stochastic upper bound on the statistic. By inverting these tests, we obtain an asymptotically valid confidence interval for the total causal effect that mathematically rigorously captures both types of uncertainty, the uncertainty about the causal structure and the uncertainty about the numerical size of the effect. Furthermore, the second solution offers an alternative that is valid in finite samples by using the framework of universal inference (Wasserman et al., 2020), see also Section 5.1, which avoids the complexities of distribution theory and the need for stochastic upper bounds.

From a theoretical standpoint, the proposed framework provides a rigorous method for constructing confidence regions for causal effects that account for structure uncertainty. However, in practice, computing this solution using likelihood ratio tests requires maximizing the Gaussian likelihood for every single hypothesis, i.e., for every possible (complete) graph under every fixed-size effect constraint. Considering the superexponential growth of the number of DAGs with the number of nodes, combinatorial shortcuts are crucial to compute the confidence intervals efficiently. To address this, in the first core article (Strieder and Drton, 2022), we introduce a branch and bound type algorithm that quickly reduces the space of plausible causal orders. Nonetheless, maximizing the Gaussian likelihood under constraints on total causal effects leads to complex optimization problems involving polynomial constraints, which must be solved with numerical optimization routines.

In contrast, in the third core article (Strieder and Drton, 2024a), we propose an alternative solution to the test inversion problem (Equation (4.1)) based on dual likelihood ratio theory (see Section 5.2). In linear Gaussian structural causal models, maximizing the dual likelihood under total effect constraints offers the key advantage over the classical likelihood of a closed-form solution. Thus, computing confidence regions for total causal effects by inverting dual likelihood ratio tests does not rely on time-consuming optimization routines while providing (asymptotically) valid confidence regions for causal effects that capture structure uncertainty.

Finally, we emphasize that the general approach of inverting joint tests of causal structure and effect size to construct confidence regions for causal effects under structure uncertainty is generalizable to other

modeling assumptions. Specifically, for parametric models where the causal structure and causal quantity of interest induce constraints that can be tested, (dual) likelihood ratio tests may be employed with the same stochastic approximation strategies. This is demonstrated in the fourth core article (Strieder and Drton, 2024b), where we extend the framework to the partially homoscedastic case and the more general case without variance restrictions.

5 Likelihood Ratio Tests

Likelihood ratio tests offer powerful solutions for a wide range of statistical hypothesis testing problems. Specifically, consider a given (parametric) statistical model $\{P_\theta : \theta \in \Theta\}$ with parameter space $\Theta \subset \mathbb{R}^d$. The distributions P_θ are assumed to be dominated by a measure μ, and we denote the corresponding μ-density with p_θ. Given an i.i.d. sample $X^1, \ldots, X^n$ from an unknown distribution P_θ in the model, we are interested in testing the hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta \setminus \Theta_0, \tag{5.1}$$

where $\Theta_0 \subset \Theta$. This setup also includes the testing problem that arises via the test inversion approach to account for structure uncertainty in causal inference see Equation (4.1).

With the log-likelihood function $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta\left(X^i\right)$, the likelihood ratio statistic for the testing problem (5.1) is given by

$$\lambda_n := 2\left(\sup_{\theta\in\Theta} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n\left(\theta\right)\right).$$

Under regularity conditions, which include probabilistic conditions on the model and geometric smoothness assumptions on Θ_0, the null distribution of the likelihood ratio statistic for the testing problem can (asymptotically) be approximated by a chi-square distribution. Notably, the arising limit distribution depends on the local geometry of Θ_0 at the true parameter θ_0, characterized by the tangent cone.

However, when these regularity conditions are not met, the needed distribution-theoretic insights may be difficult to obtain. In particular, at boundary points or model singularities, the tangent cone may not be a linear space, leading to limiting distributions other than the standard chi-square distribution (Drton, 2009). In such cases, the framework of universal inference offers a flexible alternative.

5.1 Split Likelihood Ratio Test

The recently introduced framework of *universal inference* provides a new approach to constructing hypothesis tests and confidence regions that are valid in finite samples and, importantly, do not rely on specific regularity assumptions (Wasserman et al., 2020). At the core of the methodology is the *split likelihood ratio*, a modified likelihood ratio statistic that is formed under data splitting. In particular, the available n data points are partitioned into two disjoint subsets, $D_0 = \{X^1, \ldots, X^{m_0}\}$ and $D_1 = \{X^{m_0+1}, \ldots, X^n\}$, based on a chosen splitting ratio m_0. One part of the available data points is used to form a maximum likelihood estimate of the distribution under the full model, and the other is used to compare the likelihood

under the estimate versus the null hypothesis. Thus, writing $\ell_0(\theta)$ and $\ell_1(\theta)$ for the log-likelihood functions based on D_0 and D_1, respectively, this split likelihood ratio statistic is defined as

$$\widetilde{\lambda}_n := 2\left(\ell_0(\widehat{\theta}_1) - \sup_{\theta\in\Theta_0} \ell_0(\theta)\right),$$

where $\widehat{\theta}_1 =: \text{argmax}_{\theta\in\Theta}\ell_1(\theta)$, is the (unconstrained) maximum likelihood estimate based on D_1. Crucially, the independence of the split data sets allows the use of an universal critical value instead of relying on asymptotic distribution theory. Type-I error control is ensured by a simple application of Markov's inequality, that is, we have

$$P_\theta\left(\widetilde{\lambda}_n > -2\log(\alpha)\right) \leq \alpha\mathbb{E}_\theta\left[\frac{\prod_{i=1}^{m_0} p_{\widehat{\theta}_1}(X^i)}{\sup_{\theta\in\Theta_0}\prod_{i=1}^{m_0} p_\theta(X^i)}\right] \leq \alpha.$$

Therefore, the decision rule

$$\text{reject } H_0 \text{ if } \widetilde{\lambda}_n > -2\log(\alpha)$$

constitutes a valid level α test. Notably, this split likelihood ratio test controls Type-I error at level α without any regularity conditions and, thus, makes it possible to conduct rather simple analyses of complex hypotheses. However, this critical value can be very conservative. Thus, in the second core article (Strieder and Drton, 2022), we investigate how to choose the splitting ratio optimally in the sense that the potential loss of power is minimal. In particular, we study the split likelihood ratio test under local alternatives and the classical regularity conditions that lead to chi-square limits for the ordinary likelihood ratio tests and introduce the resulting class of noncentral split chi-square distributions that arises as the limit distribution of the split likelihood ratio statistic. While the classical (noncentral) chi-square distribution corresponds to the distribution of the squared distance from a standard normal vector to some fixed point in the space, the noncentral split chi-square distribution is the distribution of the difference between two squared distances. The first part is analogously the squared distance from a standard normal vector to a fixed point in the space. However, a new second part arises arises due to splitting the data that behaves like a scaled chi-square distributed random variable where the scaling factor depends only on the chosen splitting ratio. Using properties of this new class of distributions, we propose an optimal choice of this data splitting ratio to maximize the (asymptotic) power for tests of composite hypotheses of different dimensions.

We note that Tse and Davison (2022) study universal inference from a similar perspective and investigate the loss of power of the split likelihood ratio test in the basic normal location model. Their objective is to leverage insights from the regular setting to justify using a smaller critical value, which increases the power while still controlling the size in regular cases. However, with the modified critical value, one no longer has a universal guarantee to control the size of general irregular problems. In contrast, our approach of modifying the splitting ratio remains universally valid in finite samples, regardless of whether the problem is regular or irregular. We discuss and explain these differences further in the (short) discussion by Drton et al. (2023), which is not directly part of this thesis.

Finally, we highlight that the split likelihood ratio statistic can be used as an E-process, offering a natural extension to anytime-valid inference in the context of sequential testing (Ramdas et al., 2023).

5.2 Dual Likelihood Ratio Test

In situations where directly maximizing the classical (Gaussian) likelihood has no closed-form solution, a dual formulation of the problem may offer an alternative. For example, in the context of constructing confidence regions under structural uncertainty via test inversions (see the testing problem (4.1)), direct maximization of the Gaussian likelihood with constraints on the total causal effect has no closed-form solution. Consequently, computing the solution requires numerical optimization routines and grid searches. In contrast to that, in the third core article (Strieder and Drton, 2024a), we employed dual likelihood theory for Gaussian models (e.g., Brown, 1986; Kauermann, 1996) to obtain a simple closed-form solution.

The dual formulation emerges through the connection of maximum likelihood estimation and Kullback-Leibler divergence. Specifically, we consider the statistical testing problem that arises in the context of structure uncertainty in linear Gaussian structural causal models (4.1). The maximum likelihood estimate is the closest point in the statistical model $\mathcal{M}$ to the observed sample covariance in the Kullback-Leibler sense, that is, maximizing the Gaussian likelihood corresponds to the minimization problem

$$\operatorname*{arg\,min}_{\Sigma \in \mathcal{M}} \mathrm{KL}(P_{\widehat{\Sigma}}, P_{\Sigma}) = \operatorname*{arg\,min}_{\Sigma \in \mathcal{M}} \left(\operatorname{tr}\left(\Sigma^{-1} \widehat{\Sigma} \right) + \log \det(\Sigma) \right),$$

where KL is the Kullback–Leibler divergence with sample covariance $\widehat{\Sigma}$. As the Kullback–Leibler divergence is not symmetrical, swapping its arguments yields a different minimization problem, namely,

$$\operatorname*{arg\,min}_{\Sigma \in \mathcal{M}} \mathrm{KL}(P_{\Sigma}, P_{\widehat{\Sigma}}) = \operatorname*{arg\,min}_{\Sigma \in \mathcal{M}} \left(\operatorname{tr}\left(\widehat{\Sigma}^{-1} \Sigma \right) + \log \det(\Sigma^{-1}) \right).$$

The solution to this minimization problem is called the *dual maximum likelihood estimator*. We highlight that this solution coincides with the classical maximum likelihood estimate if the statistical model is a full exponential family (Brown, 1986), i.e., similar to the classical likelihood, $\widehat{\Sigma}$ is the (unconstrained) dual maximum likelihood estimate among all centered multivariate normal distributions. However, this equality is not necessarily true for submodels, and thus, this dual formulation generally yields a different estimator.

Moreover, by substituting the classical maximum likelihood estimation with this dual minimization problem, we can define the dual likelihood ratio statistic analogously to the classical likelihood ratio statistic. We emphasize that the corresponding *dual likelihood ratio test* can be seen as solving a modified testing problem with the classical likelihood ratio test and observed sample covariance $\widehat{\Sigma}^{-1}$. Thus, many properties of maximum likelihood estimation apply to the corresponding dual maximum likelihood estimate. In the third core article (Strieder and Drton, 2024a), we leverage this correspondence to derive the limit distribution of the dual likelihood ratio test statistic to construct (asymptotically) valid confidence regions for total causal effects under structure uncertainty.

6 Conclusion

Concrete knowledge of the causal mechanism that governs the data-generation process is crucial to draw causal conclusions from observed data in interacting systems. However, in typical applications, this underlying causal structure is unknown and must be inferred from available data. In this publication-based thesis, we address the challenge of rigorously accounting for this structure uncertainty in causal inference. In particular, we demonstrate the complexity of this task and introduce a test inversion approach that recasts the problem as a concrete statistical testing problem. Furthermore, we offer concrete solutions to the arising testing problem within the context of linear structural causal models to construct (asymptotically) valid confidence regions that rigorously reflect all types of remaining uncertainty: classical statistical uncertainty about the numerical size of involved effects and causal structure uncertainty, stemming from a data-driven structure learning approach and the equivalence of multiple plausible models.

Without prior knowledge about the underlying causal structure, we have to consider all possible causal structures to make calibrated confidence statements that include structure uncertainty. This highlights the fundamental practical bottleneck of the task itself: the superexponential growth of the number of possible causal structures with the number of nodes. Therefore, leveraging combinatorial shortcuts is essential to compute rigorous confidence regions under structure uncertainty. In particular, we highlight that due to the test inversion ansatz, we only need to consider complete DAGs, immensely decreasing the worst-case number of causal structures we need to take into account. Thus, we combine these shortcuts in a branch and bound type algorithm that allows us to compute the proposed confidence region under (full) structure uncertainty in reasonable time for moderate dimensions. In higher dimensional systems, our method still offers a valuable tool to practitioners by at least partially quantifying the remaining uncertainty in the causal structure. That is, rather than solely relying on a pre-specified DAG by experts or a single output of causal learning algorithms, practitioners can employ our framework to a set of (most plausible) causal structures and, thus, incorporate some degree of structure uncertainty in causal effect estimates.

Throughout this thesis, we focused on the concrete setting of linear relations disturbed by an additive Gaussian noise. Even in this simplified setting, we already observe high levels of uncertainty about causal directions and size of effects, highlighting the importance of methods that properly account for structure uncertainty in causal inference. We emphasize that our general approach of leveraging test inversions of joint tests for causal structure and effect size to reformulate the task as a concrete statistical testing problem is generalizable to other modeling settings. The arising testing problem may then be solved with similar stochastic approximation strategies based on (modified) likelihood ratio tests, particularly if one deals with parametric (causal) models for which the causal structure is identifiable. From this perspective, we consider our work as building the foundation of a potential new line of research on rigorously incorporating structure uncertainty in causal inference.

7 Summary Core Publications

7.1 Confidence in Causal Inference under Structure Uncertainty in Linear Causal Models with Equal Variances

Summary. In this article (full published version in Appendix A.1), we raise the problem of rigorously accounting for uncertainty in causal inference when the causal structure is unknown. We demonstrate the intricacies of this task and introduce a test-inversion framework that circumvents the problems posed by the non-smooth nature of the causal quantity of interest. Specifically, we present our main conceptual idea to leverage the classical duality between statistical hypothesis tests and confidence regions. Using this ansatz, we highlight how to recast the problem as a concrete statistical testing problem of joint tests for causal structure and effect size. In the concrete context of linear structural causal models with homoscedastic Gaussian errors, we propose two precise solutions for constructing rigorous confidence regions for total causal effects under structure uncertainty. In particular, by employing intersection-union theory and stochastic upper bounds, our solution addresses the challenge of both the null hypothesis and the alternative being intricate unions that do not satisfy typical regularity conditions. Furthermore, we present a top-down procedure to compute our solution, which is feasible for problems of small to moderate scale—e.g., in the simulation experiments, we considered systems with up to 12 variables.

The paper is organized as follows. We first introduce the topic in Section 1. In Section 2 we review the considered restricted class of structural causal models that ensures identifiability from observational data alone, namely, linear models with homoscedastic Gaussian errors, and define the causal target of interest, the total causal effect. In Section 3, we introduce our main ansatz to construct confidence intervals for total causal effects in multivariate data, the test inversion approach, and derive the concrete hypothesis that subsequently needs to be tested. In Section 4, we present two concrete tests based on likelihood ratio theory to solve the arising statistical testing problem: a classical constrained likelihood ratio test and an approach based on the split likelihood ratio test. Based on these solutions, we state our main result, which is the construction of valid confidence regions for total causal effects when the causal structure is unknown. Furthermore, in Section 5 we present computational details and combinatorial shortcuts to compute the proposed confidence regions and analyze the performance of the introduced framework in numerical experiments. In the concluding Section 6 we discuss extensions of the framework to partially quantify the uncertainty over causal structures.

Individual contributions. I am the sole first author of this article. The idea of investigating this topic was developed jointly with my supervisor Mathias Drton. Specifically, the concept of using test inversion to construct confidence regions first emerged in our earlier conference paper (Strieder et al., 2021) for simple

bivariate models. In this article, I generalized this idea to multivariate data, formalized the methodology, and proposed an improved solution. I was responsible for writing the manuscript, developing the proofs for all statements, implementing the algorithm, and conducting the simulation study. Mathias Drton provided valuable feedback on both the content and presentation through regular discussions.

7.2 On the Choice of the Splitting Ratio for the Split Likelihood Ratio Test

Summary. The framework of universal inference provides a new approach to constructing hypothesis tests and confidence regions that are valid in finite samples and do not rely on any specific regularity assumptions on the underlying statistical model. At the core of the methodology is a split likelihood ratio statistic, which is formed under data splitting and compared to a cleverly selected universal critical value. As this critical value can be very conservative, it is interesting to mitigate the potential loss of power by carefully choosing the ratio according to which data are split. In this article (full published version in Appendix A.2), we expand our insights about the (limit) behavior of this split likelihood ratio test. In particular, we shed light on the impact of the dimensionality of the tested null and alternative hypotheses by studying the case of smooth hypotheses in regular parametric models that are differentiable in quadratic mean. We derive precise limit theory and calculate the large-sample asymptotic distribution, allowing for local alternatives. This arising limit distribution belongs to a "split-version" of noncentral chi-square distributions, for which we derive explicit properties, i.e., the moments. We then use properties of this new class of noncentral split chi-square distributions to propose a routine for calculating the optimal splitting ratio for the split likelihood ratio test based on the dimensionality of the tested null and alternative hypotheses. Here, optimal refers to the splitting ratio that achieves the highest (asymptotic) power in the regular setting. Importantly, this modification employs the obtained insights from the regular setting to tune the data splitting ratio, but the test remains universally valid in finite samples, regardless of whether the considered problem is regular or not.

The paper is organized as follows. In Sections 1 and 2, we introduce the topic of universal inference and provide the needed background on the split likelihood ratio test. In Section 3, we derive the asymptotic distribution of the split likelihood ratio test statistic assuming the usual smoothness properties that lead to chi-square limits for the ordinary likelihood ratio test statistic and introduce the arising new class of noncentral split chi-square distributions. Based on these insights, we propose a new routine for calculating the optimal splitting ratio. In Section 4, we present the results of a simulation study that investigates the asymptotic behavior of the split likelihood ratio test and compare its performance in different regular and irregular model settings. We conclude with Sections 5 and 6, where we discuss extensions to the cross-fit variant of the split likelihood ratio test.

Individual contributions. I am the sole first author of this article. The initial idea for investigating this topic was developed during collaborative discussions with my supervisor Mathias Drton. I was responsible for writing the manuscript, developing the theoretical concepts and proofs for all statements, and conducting the simulation study. Mathias Drton made helpful suggestions regarding both the theory and the writing of the article during regular discussions.

7.3 Dual Likelihood for Causal Inference under Structure Uncertainty

Summary. In this article (full published version in Appendix A.3), we address the important problem of accounting for structure uncertainty in causal inference. To be precise, knowledge of the underlying causal relations is essential for inferring the effect of interventions in complex systems. However, in the typical application, this underlying causal structure is unknown and, thus, has to be learned from data. In order to draw reliable conclusions from causal estimates, the remaining structure uncertainty needs to be incorporated into causal inference. In our previous work, we introduced a test inversion approach that provides an ansatz to account for this data-driven model choice and, therefore, bridges the gap between causal discovery and causal inference. This test inversion ansatz transforms the problem into a concrete statistical testing problem. In this article, we take up the idea of using a test inversion ansatz to construct mathematically rigorous confidence regions for total causal effects under structure uncertainty. However, we propose using dual likelihood theory to significantly simplify the treatment of the involved testing problem. In particular, in the concrete context of linear structural causal models with homoscedastic Gaussian errors, we present a closed-form solution that avoids the need for numerical optimization routines and, thus, significantly saves on computation. Furthermore, we provide theoretical coverage guarantees for the proposed confidence regions for total causal effects using intersection-union theory and the limit behavior of stochastic upper bounds of the dual likelihood ratio test statistic. Finally, we introduce a bottom-up, branch and bound type procedure starting from sink nodes to compute the proposed confidence regions and show in a simulation study, that the differences in performance and explanatory power that stem from employing dual likelihood ratio tests instead of classical likelihood ratio tests are negligible and heavily outweighed by the significantly reduced computation times due to the available closed-form solution.

The paper is organized as follows. In Sections 1 and 2, we formulate the research question and introduce the topic of structure uncertainty in causal inference. We recapitulate the necessary (causal) preliminaries and review the test inversion ansatz to recast the problem as a concrete statistical testing problem. Section 3 introduces the dual likelihood ratio test and presents a closed-form solution to construct calibrated confidence regions for total causal effects under structure uncertainty. In Section 4, we present computational details and propose a bottom-up procedure to compute the confidence regions. Furthermore, we compare the performance and computation times of the different approaches in a simulation study. In the concluding Section 5, we discuss extensions and possible future work.

Individual contributions. I am the sole first author of this article. The idea of employing dual likelihood theory to tackle the statistical testing problem arising from the test inversion approach was proposed by my supervisor, Mathias Drton, during our regular discussions. I was responsible for writing the manuscript, developing the theory and proofs of all statements, implementing the solution, and performing the numerical experiments. Mathias Drton provided valuable feedback and comments on both the content and presentation of the paper.

7.4 Identifying Total Causal Effects in Linear Models under Partial Homoscedasticity

Summary. An essential aspect of causal discovery is clarifying under which conditions it is theoretically possible to identify causal quantities from observational data alone. In this article (full published version in Appendix A.4), we investigate how different assumptions on the error variances lead to the (non-)identifiability of total causal effects in the context of linear structural causal models with Gaussian errors. In particular, classical identifiability results suggest that, without additional structural assumptions, the underlying causal structure can only be identified up to a Markov equivalence class of indistinguishable models. Thus, the causal effect of interest can only be recovered up to a (finite) set of indistinguishable plausible effects. The frequently studied structural assumption of homoscedasticity among the error variances resolves this issue and entails global identifiability of the underlying causal structure. Consequently, the causal effect of interest can be uniquely recovered. However, practitioners are often primarily interested in the effects of specific interventions, rendering the complete identification of the underlying causal structure unnecessary. We demonstrate that the less stringent modeling assumption of partially equal error variances (only) between the potential cause and response of interest is sufficient to generically identify the total causal effect. Furthermore, we follow the test inversion approach and extend the framework of constructing confidence regions for total causal effect under structure uncertainty to the partially homoscedastic case and the general unrestricted error variances cases.

The paper is organized as follows. In the first Section, we introduce the topic and the (causal) question of interest. In Section 2, we review linear structural causal models and the definition of the total causal effect, which is the causal target of interest in this study. Further, we recapitulate the notions of identifiability and Markov equivalence of models. In Section 3, we investigate how introducing homoscedasticity constraints among the error variances leads to identifiability of the (causal) parameters of interest. In particular, we compare the unrestricted, the partially homoscedastic, and the fully homoscedastic case and show that the assumption of partial homoscedasticity is sufficient to generically identify the total causal effect. In Section 4, we investigate how causal inference differs under these three error variance assumptions by highlighting how to estimate causal effects and how to construct confidence regions for causal effects under structure uncertainty. In the simulation study in Section 5, we compare the width of the constructed confidence regions and, thus, exemplify the performance gained by relying on stricter structural assumptions.

Individual contributions. I am the sole first author of this article. The idea of investigating this topic was developed jointly with my supervisor Mathias Drton. I was responsible for writing the manuscript, developing the proofs for all statements, implementing the algorithm, and conducting the simulation study. Mathias Drton provided valuable feedback and suggestions on both the theory and writing in our regular discussions.

Bibliography

Andrews, D. W. K. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2):426–468.

Brown, L. D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA.

Casella, G. and Berger, R. L. (1990). *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Chang, T.-H., Guo, Z., and Malinsky, D. (2024). Post-selection inference for causal effects after causal discovery. *arXiv preprint arXiv:2405.06763*.

Chen, W., Drton, M., and Wang, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(3):507–554. Computational learning theory.

Drton, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012.

Drton, M. (2018). Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77 of *Adv. Stud. Pure Math.*, pages 35–86. Math. Soc. Japan, Tokyo.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.*, 4:365–393.

Drton, M., Shi, H., and Strieder, D. (2023). A discussion of "a note on universal inference" by Tse and Davison. *Stat*, 12(1):e574.

Drton, M. and Williams, B. (2011). Quantifying the failure of bootstrap likelihood ratio tests. *Biometrika*, 98(4):919–934.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., and van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nat. Med.*, 30(4):958–968.

Gao, M., Ming Tai, W., and Aragam, B. (2022). Optimal estimation of gaussian dag models. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *PMLR*, pages 8738–8757.

Gradu, P., Zrnic, T., Wang, Y., and Jordan, M. I. (2024). Valid inference after causal discovery. *J. Amer. Statist. Assoc.*, 0(ja):1–21.

Henckel, L., Perković, E., and Maathuis, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84(2):579–599.

Hernan, M. and Robins, J. (2024). *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference—for statistics, social, and biomedical sciences*. Cambridge University Press, New York. An introduction.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.*, 47(11):1–26.

Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.*, 23(1):105–116.

Keropyan, G., Strieder, D., and Drton, M. (2023). Rank-based causal discovery for post-nonlinear models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pages 7849–7870.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York.

Maathuis, M. H., Drton, M., Lauritzen, S., and Wainwright, M., editors (2019). *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164.

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and nobel laureates. *N. Engl. J. Med.*, 367(16):1562–1564.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statist. Sci.*, 8(3):266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019). Inferring causation from time series in earth system sciences. *Nat. Commun.*, 10(1):2553.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.*, 9(1):62–72.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Strieder, D. and Drton, M. (2022). On the choice of the splitting ratio for the split likelihood ratio test. *Electron. J. Stat.*, 16(2):6631–6650.

Strieder, D. and Drton, M. (2023). Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *J. Causal Inference*, 11(1).

Strieder, D. and Drton, M. (2024a). Dual likelihood for causal inference under structure uncertainty. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *PMLR*, pages 1–17.

Strieder, D. and Drton, M. (2024b). Identifying total causal effects in linear models under partial homoscedasticity. In *Proceedings of The 12th International Conference on Probabilistic Graphical Models*, volume 246 of *PMLR*, pages 213–230.

Strieder, D., Freidling, T., Haffner, S., and Drton, M. (2021). Confidence in causal discovery with linear causal models. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *PMLR*, pages 1217–1226.

Sullivant, S. (2018). *Algebraic statistics*, volume 194 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

Tse, T. and Davison, A. C. (2022). A note on universal inference. *Stat*, 11:Paper No. e501, 9.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Wang, Y. S., Kolar, M., and Drton, M. (2023). Confidence sets for causal orderings. *arXiv preprint arXiv:2305.14506*.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA*, 117(29):16880–16890.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 647–655. AUAI Press.

A Core Publications

A.1 Confidence in Causal Inference under Structure Uncertainty in Linear Causal Models with Equal Variances

Research Article

David Strieder* and Mathias Drton

Confidence in causal inference under structure uncertainty in linear causal models with equal variances

https://doi.org/10.1515/jci-2023-0030
received May 16, 2023; accepted September 21, 2023

Abstract: Inferring the effect of interventions within complex systems is a fundamental problem of statistics. A widely studied approach uses structural causal models that postulate noisy functional relations among a set of interacting variables. The underlying causal structure is then naturally represented by a directed graph whose edges indicate direct causal dependencies. In a recent line of work, additional assumptions on the causal models have been shown to render this causal graph identifiable from observational data alone. One example is the assumption of linear causal relations with equal error variances that we will take up in this work. When the graph structure is known, classical methods may be used for calculating estimates and confidence intervals for causal-effects. However, in many applications, expert knowledge that provides an a priori valid causal structure is not available. Lacking alternatives, a commonly used two-step approach first learns a graph and then treats the graph as known in inference. This, however, yields confidence intervals that are overly optimistic and fail to account for the data-driven model choice. We argue that to draw reliable conclusions, it is necessary to incorporate the remaining uncertainty about the underlying causal structure in confidence statements about causal-effects. To address this issue, we present a framework based on test inversion that allows us to give confidence regions for total causal-effects that capture both sources of uncertainty: causal structure and numerical size of non-zero effects.

Keywords: confidence intervals, causal-effects, linear structural equation models, equal error variances, graphical models

MSC 2020: 62D20, 62H22

1 Introduction

Questions about causal relations are at the heart of many research problems. Researchers across various fields, ranging from economics to biology and medicine, are interested in distinguishing causes from effects to predict the outcome of interventions in complex systems. While controlled experiments provide an established method for inferring causal relations, in many applied settings, interventional studies are not feasible for ethical or cost considerations. Thus, inferring causal relationships based solely on observational data is an important task that the field of causal discovery addresses.

In the last few decades, causal discovery has gained popularity, with much fundamental research being done on the topics of identification and estimation [1,2]. Identifiability results clarify when it is theoretically

*** Corresponding author: David Strieder,** TUM School of Computation, Information and Technology, Munich Center for Machine Learning, Technical University of Munich, 80333 Munich, Germany, e-mail: david.strieder@tum.de
Mathias Drton: TUM School of Computation, Information and Technology, Munich Center for Machine Learning, Technical University of Munich, Munich, Germany, e-mail: mathias.drton@tum.de

possible to go beyond mere correlations. More specifically, they clarify under which circumstances, assumptions lead to unique identification of interventional distributions from a given joint probability distribution. Moreover, numerous structure learning algorithms have been proposed that infer the underlying causal structure based on observed data. Knowledge of the underlying causal structure then allows researchers to reason about the effect of interventions in complex systems, and given a causal structure, standard statistical methods may be applied to estimate the involved causal-effects and provide an uncertainty assessment for the numerical size of the effects. However, in the typical applied setting, the exact causal structure is (at least partially) unknown. In order to draw reliable conclusions and make calibrated confidence statements, the remaining uncertainty about the underlying causal structure needs to be incorporated in causal inference.

Without prior expert knowledge about the underlying causal structure, a commonly employed "naïve" approach would involve splitting the task. In the first step, a causal learning algorithm estimates the underlying causal structure. In the second step, within the inferred model, confidence intervals for the causal parameters can be calculated using classical statistical inference methods. However, the resulting intervals are not valid if the causal learning algorithm infers the wrong causal structure. Such a two-step method conditions away the uncertainty in the causal structure arising from the data-driven model choice, and the classical "double-dipping" problem occurs since both causal structure and the effect are estimated from one data set. Thus, the approach is overly optimistic in its conclusion about the strength and existence of causal-effects, and its confidence regions do not achieve the desired coverage probability, especially under high uncertainty with respect to the underlying structure.

Despite the popularity and growing interest in causal discovery, we are unaware of prior attempts to rigorously account for uncertainty in the causal structure when providing confidence statements about causal-effects. Here, we use the term "confidence" in the technical sense for a set with a given desired frequentist coverage probability. A completely different approach for uncertainty quantification that we do not consider here is to form Bayesian credible sets. In fact, a number of authors have used Bayesian approaches for uncertainty quantification in causal discovery, but primarily with an emphasis on the uncertainty in the graphical structure as opposed to causal-effects (see [3–5] for three selected examples).

In this article, we want to shed light on the problem by highlighting the challenges of the task of handling structure uncertainty and by proposing a solution for constructing confidence intervals for total causal-effects that capture both types of uncertainty: the uncertainty regarding the causal structure resulting from a data-driven model selection and the uncertainty about the numerical size of the effect. Our solution is feasible for problems of small to moderate scale – e.g., our simulation experiments will consider problems with up to 12 variables. However, already at this scale, rigorously accounting for structure uncertainty means to account for more than 10^{26} possible structures, reflecting the fast superexponential growth with the number of nodes in the causal graph.

In Section 2 of this article, we review the considered framework of structural causal models [6,7]. More specifically, in order to ensure identifiability from observational data alone, we consider a restricted class of causal models, namely, linear structural equation models (LSEMs) with Gaussian errors and equal variances [8], and review the definition of a total causal-effect, which is the target of interest in this study. We then generalize and improve the ideas introduced for bivariate problems by Strieder et al. [9] and present a general framework for constructing confidence intervals for total causal-effects in multivariate data (Sections 3 and 4). Furthermore, we present computational shortcuts and analyze the performance of the introduced framework in numerical experiments (Section 5). In Section 6, we discuss extensions of the framework to partially quantify the uncertainty over causal structures.

2 Background

This section reviews the LSEMs and the total causal-effect, which is the target of interest in this study.

2.1 LSEMs

Structural causal models constitute a commonly used tool to study causal relationships and represent noisy causal relations among a set of interacting variables. Each variable is modeled as a function of a subset of other variables, its causes, and a stochastic error term. The causal perspective emerges from viewing those relations as making assignments rather than representing mathematical equations. The left-hand side, the effect, is assigned the value specified on the right-hand side, the causes. Externally varying the values on the right-hand side results in a change on the left-hand side, but not vice versa. This reflects the asymmetry inherent in cause–effect relations.

We assume access to observational data in the form of a sample of independent copies of a random vector $X = (X_1, \ldots, X_d)$; without loss of generality, we assume the random vector to have zero mean. Without any further assumptions on the structural causal model, e.g., the involved functions, identification of the underlying causal structure is only possible up to a Markov equivalence class. Therefore, to ensure unique identifiability, following a line of research initiated by Peters and Bühlmann [8], we focus on linear relations and normally distributed errors with equal variances by assuming that X solves the equation system:

$$X_j = \sum_{i \neq j} \beta_{j,i} X_i + \varepsilon_j, \quad j = 1, \ldots, d, \tag{1}$$

where $B \coloneqq [\beta_{j,i}]_{j,i=1}^{d}$ are unknown parameters that represent the direct causal-effects between the variables, and ε_j are independent, normally distributed error terms:

$$\varepsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \ldots, d,$$

with common, unknown variance $\sigma^2 > 0$. Each specific LSEM restricts a subset of the direct effects $\beta_{j,i}$ to be zero. Thus, LSEMs are naturally represented by a (minimal) directed graph $G(B)$ with vertex set $V = \{1, \ldots, d\}$. In the associated directed graph, the edge set equals the support of B; thus, a missing edge $i \to j$ indicates that $\beta_{j,i} = 0$. As in related work, we assume the underlying directed graph to be acyclic (DAG), which entails that B is permutation similar to a triangular matrix and System (1) admits the unique solution $X = (I_d - B)^{-1}\varepsilon$, with I_d denoting the $d \times d$ identity matrix. Incorporating the equal variance assumption, the covariance matrix of X is seen to be

$$\mathbb{E}[XX^T] = \sigma^2 (I_d - B)^{-1} (I_d - B)^{-T}.$$

In the sequel, we will use the following graphical concepts. Each DAG admits a (not necessarily unique) topological ordering of its vertices, i.e., a permutation π of $\{1, \ldots, d\}$ such that the existence of a directed path from node $\pi(i)$ to node $\pi(j)$ implies $i < j$. We write $i <_G j$ if node i precedes node j in such a causal ordering of the graph. If the DAG contains an edge from node i to node j, then node i is a parent of node j, and we denote the set of all parents of node j with $p(j)$. It has been shown that under the aforementioned modeling assumptions with equal error variances, i.e., homoscedasticity across all interacting variables, the causal ordering is identifiable and corresponds to an ordering of conditional variances [10,11]. Hence, causal-effects can also be uniquely identified.

2.2 Causal-effects

Informally put, we are interested in estimating how much the value of X_j changes if we externally intervene in the system and set the value of X_i to x_i. Such an intervention represents a change in the true data-generating process and, thus, a modification in our probabilistic model. In the structural equation framework, the effect of such an external intervention can be expressed by replacing the ith equation with $X_i = x_i$ and is denoted as $\mathrm{do}(X_i = x_i)$.

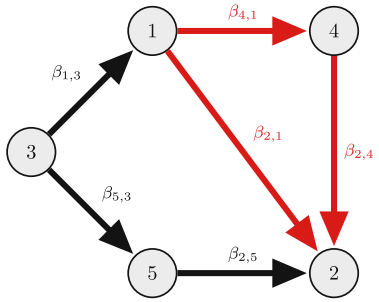
In this work, we seek to construct confidence intervals for the total causal-effect $C(i \to j)$ in LSEMs. This effect is formally defined as the unit change in the expectation of X_j with respect to an intervention in X_i, i.e.,

$$C(i \to j) \coloneqq \frac{\mathrm{d}}{\mathrm{d}x_i}\mathbb{E}[X_j | \operatorname{do}(X_i = x_i)].$$

Remark 2.1. Note that the total causal-effect in an LSEM (B, σ^2) can be intuitively expressed based on the respective underlying DAG $G(B)$. The effect $C(i \to j)$ equals the sum of all products of edge weights along all directed paths from i to j. Furthermore, the effect is zero if there does not exist a directed path from i to j. This follows from the path properties of the matrix $(I_d - B)^{-1}$ that computes the solution X of the structural equations from the independent errors.

Example 2.2. Consider the following LSEM and the corresponding (minimal) DAG:

$$\begin{aligned}
X_1 &= \beta_{1,3}X_3 + \varepsilon_1 \\
X_2 &= \beta_{2,1}X_1 + \beta_{2,4}X_4 + \beta_{2,5}X_5 + \varepsilon_2 \\
X_3 &= \varepsilon_3 \\
X_4 &= \beta_{4,1}X_1 + \varepsilon_4 \\
X_5 &= \beta_{5,3}X_3 + \varepsilon_5
\end{aligned}$$



The total causal-effect is given by the sum of all products of edge weights along all directed paths. For example, we obtain $C(1 \to 2) = \beta_{2,1} + \beta_{4,1}\beta_{2,4}$, whereas $C(2 \to 1) = 0$.

Our aim is to construct a confidence interval for the total causal-effect $C(i \to j)$ when the underlying causal structure is not known and has to be learned. We stress that, under the identifiability assumption of an underlying Gaussian LSEM with equal error variances, this is a well-defined problem. In fact, every admissible distribution can be represented by a single (minimal) DAG. This DAG then uniquely defines the total causal-effect.

A "naïve" two-step approach that first learns a causal structure and then calculates confidence intervals in the inferred model does not incorporate the uncertainty regarding the underlying causal structure. To respect this uncertainty in the calculations of a valid confidence interval, it is necessary to take all possible structures, i.e., graphs, into account. As a result, the target quantity has a composite structure, and thus, obtaining finite sample distributions of estimators is difficult.

In many situations where the sampling distribution of estimators is unobtainable, bootstrapping and other resampling methods can offer an appealing framework to construct confidence intervals. At first sight, resampling may appear to provide a simple solution to the problem under discussion. This solution relies on computing for each resampled data set a causal-effect estimate. To obtain this estimate, one may compose a consistent model selection method that learns the causal graph with a consistent causal-effect estimator in the learned model. Note that this composition is well defined due to the identifiability of LSEMs with equal error variances. However, as a statistic, this composition lacks smoothness, and thus, bootstrapping procedures may fail severely [12,13]. We emphasize that Drton and Williams [13] demonstrate that even in low dimensions, the asymptotic behavior of confidence intervals and bootstrap tests is difficult to predict for complex composite settings as encountered here. The subtleties stem from the fact that the set of covariance matrices associated with at least one possible DAG form a union of smooth manifolds. This union contains singularities at the intersections of the manifolds, which is highlighted in the bivariate case by Strieder et al. [9]. They demonstrated that singular points arise at the intersections of the two models. Consequently, bootstrapping cannot be used to capture model uncertainty correctly. This failure of the bootstrapping methods can also be observed in our simulations in Section 5.3.

In conclusion, there is a need to develop new frameworks for constructing confidence intervals that circumvent the problems that originate from the non-smooth nature of the target of interest.

3 Confidence intervals via test inversion

In the remainder of this article, we develop a framework that circumvents the problems posed by the non-smooth nature of the total causal-effect and provides valid confidence intervals for causal inference under structure uncertainty. In this section, we introduce our main ansatz, the test inversion approach, and derive the concrete hypothesis that subsequently needs to be tested.

3.1 Inversion of tests

Our main idea is to leverage the classical duality between statistical hypothesis tests and confidence regions in the following way. If we perform valid hypothesis tests for all attainable causal-effects, then all values that we cannot reject form a confidence region for the total causal-effect. More specifically, let $\alpha \in (0,1)$ be a fixed significance level. We perform (asymptotically) valid-level α tests for the hypothesis of a fixed total causal-effect ψ for all attainable values of the causal-effect, i.e., for all $\psi \in \mathbb{R}$, we test

$$\mathrm{H}_0 : C(1 \to 2) = \psi,$$

with a test whose acceptance region we denoted by $A(\psi)$. Then. a (asymptotically) valid $(1-\alpha)$-confidence region for the total causal-effect $C(1 \to 2)$ for the data X is given by:

$$C(X) := \{\psi \in \mathbb{R} : X \in A(\psi)\}.$$

A proof of this classical result can be found, e.g., in [14, Theorem 9.2.2].

With the aforementioned perspective, our problem is shifted to developing suitable hypothesis tests for all attainable causal-effects. Due to the lack of concrete knowledge about the model, all possible causal graphs must be respected. Thus, we have to construct hypothesis tests in a composite setting that remains challenging. In the next section, we explicitly describe the hypotheses that need to be tested to construct confidence regions for the total causal-effect.

3.2 Hypothesis of a fixed causal-effect

Recalling the assumptions of an underlying LSEM with homoscedastic errors across all interacting variables, we consider the following task. We assume that the given data consist of random vectors $X^{(1)}, \dots, X^{(n)}$ drawn independently from a centered multivariate normal distribution $N(0, \Sigma_0)$. The distribution's density $p(x|\Sigma_0)$ factorizes according to an unknown DAG while satisfying the assumption of equal error variances. To account for the uncertainty in the causal structure, we need to test at level α whether any multivariate centered normal distribution that factorizes to any DAG under the equal variance assumption allows for a total causal-effect of size ψ, and repeat that test for all values $\psi \in \mathbb{R}$. In light of the normality assumption, it is natural to parameterize our statistical model in terms of the covariance matrix Σ. The additional assumption of homoscedasticity across all interacting variables restricts the set of possible covariance matrices Furthermore, as we detail now.

Given a fixed DAG G, under the assumption of equal error variances, all possible normal distributions with a density that factorizes according to G are defined by the family:

$$\mathcal{N}(G) := \{N(0,\Sigma) : \exists B \in \mathbb{R}^G, \sigma^2 > 0, \quad \text{with } \Sigma = \sigma^2 (I_d - B)^{-1}(I_d - B)^{-T}\}, \tag{2}$$

where $\mathbb{R}^G = \{B \in \mathbb{R}^{d\times d} : \beta_{j,i} = 0 \text{ if } i \notin p(j)\}$. Considering and combining all possible DAG models, our overall model corresponds to the following union of sets of covariance matrices:

$$\mathcal{M} := \bigcup_{G \in \mathcal{G}(d)} \{\Sigma \in \mathrm{PD}(d) : N(0,\Sigma) \in \mathcal{N}(G)\}, \tag{3}$$

where $\mathcal{G}(d)$ is the set of all DAGs with d nodes.

Remark 3.1. If G^* is a supergraph of G, then the set $\mathcal{N}(G)$ is a subset of $\mathcal{N}(G^*)$. Hence, in the definition of $\mathcal{M}$, we only need to consider the union over all complete DAGs with d nodes, i.e., all possible orderings.

The following proposition shows that we can identify the model $\mathcal{M}$, which is a subset of all positive definite matrices satisfying the assumption of equal error variances, solely based on polynomial constraints in the entries of the covariance matrix. Here, and in the remainder of this article, we write $\Sigma_{i,j|p(i)}$ for the conditional covariance matrix, i.e.,

$$\Sigma_{j,i|p(i)} := \Sigma_{j,i} - \Sigma_{j,p(i)}(\Sigma_{p(i),p(i)})^{-1}\Sigma_{p(i),i}.$$

Proposition 3.2. *For a covariance matrix* $\Sigma \in \mathrm{PD}(d)$, *the following statements are equivalent*:
(*i*) $\Sigma \in \mathcal{M}$,
(*ii*) *There exists a complete DAG* G *and* $\sigma^2 > 0$ *such that for all* $k = 1, \dots, d$

$$\sigma^2 = \Sigma_{k,k|p(k)} = \Sigma_{k,k} - \Sigma_{k,p(k)}(\Sigma_{p(k),p(k)})^{-1}\Sigma_{p(k),k}. \tag{4}$$

Proof. " $\Rightarrow$ " Let $\Sigma \in \mathcal{M}$. Then, there exists a DAG G such that $N(0, \Sigma) \in \mathcal{N}(G)$. Considering Remark 3.1, we can assume without loss of generality that G is complete. Straightforward calculations with the multivariate normal density then imply that the variances of each node conditioned on its parents are equal to σ^2, or see a combinatorial argument as in [15, Equation (7.4)]. Thus, for all $k = 1, \dots, d$,

$$\sigma^2 = \mathrm{Var}(X_k|X_{p(k)}) = \Sigma_{k,k} - \Sigma_{k,p(k)}(\Sigma_{p(k),p(k)})^{-1}\Sigma_{p(k),k}.$$

" $\Leftarrow$ " Since G is complete, the multivariate normal distribution $N(0, \Sigma)$ factorizes according to the graph G. This implies the existence of a matrix $B \in \mathbb{R}^G$ and a positive definite diagonal matrix $\Omega \in \mathbb{R}^{d\times d}$ with $\Sigma = (I_d - B)^{-1}\Omega(I_d - B)^{-T}$. Equation (4) implies that $\Omega_{k,k} = \Sigma_{k,k|p(k)}$ is equal for all $k = 1, \dots, d$, and the claim follows. □

Remark 3.3. Note the identifiability result under homoscedasticity across all interacting variables, e.g., Theorem 1 in [10]. Each $\Sigma \in \mathcal{M}$ corresponds to a unique (B, σ^2) and, thus, a unique minimal DAG $G(B)$. As explained in Remark 3.1, the union in the definition of $\mathcal{M}$ (3) is not necessarily disjoint, and analogously, the (complete) graph satisfying Condition 2 in Proposition 3.2 is not necessarily unique. Nevertheless, since B is unique, all graphs satisfying Condition 2 in Proposition 3.2 share the same edge set with non-zero edge coefficients and correspond to all valid causal orderings of the minimal DAG. Thus, every "additional" parent of node i in a complete DAG G from Condition 2 compared to the unique minimal DAG $G(B)$ is conditionally independent of node i given $p_{G(B)}(i)$. This follows by d-separation since every (undirected) path between node i and an additional parent either traverses the parents $p_{G(B)}(i)$ in the minimal DAG or contains a collider c with $i <_{G(B)} c$, and thus, this collider has no descendent in $p_{G(B)}(i)$.

Our testing problem is the following classical statistical problem. We are concerned with a parametric model $\mathcal{M}$, parameterized by the covariance matrix, which is an intricate union of subsets of the cone of positive definite matrices given by polynomial constraints. Within that statistical model, we need to test for all attainable values of the total causal-effect ψ. In the following proposition, we show that the parameter of interest is a function of the covariance matrix. Thus, the considered statistical testing problem can be fully parameterized in terms of the covariance matrix.

Proposition 3.4. *Let* $\Sigma \in \mathcal{M}$. *Then, the total causal-effect* $C(i \to j)$ *is given by*:

$$C(i \to j) = \Sigma_{j,i|p(i)}(\Sigma_{i,i|p(i)})^{-1},$$

with parent set $p(i)$ *based on any corresponding DAG* G *from Proposition* 3.2.

Proof. The case $j <_G i$ is trivially true, since in this case, the total causal-effect $C(i \to j)$ is zero and, considering G is complete, the parent set $p(i)$ contains the node j, which yields $\Sigma_{j,i|p(i)} = 0$.

Assume $i <_G j$. Since the set $p(i)$ satisfies the back-door criterion with respect to i and j, we have

$$C(i \to j) := \frac{\mathrm{d}}{\mathrm{d}x_i}\mathbb{E}[X_j|\,\mathrm{do}\,(X_i = x_i)] = \frac{\mathrm{d}}{\mathrm{d}x_i}\mathbb{E}[\mathbb{E}[X_j|X_i = x_i, X_{p(i)}]].$$

It follows that the total causal-effect $C(i \to j)$ is the regression coefficient of X_i when regressing X_j on $(X_i, X_{p(i)})$, which is given by $\Sigma_{j,i|p(i)}(\Sigma_{i,i|p(i)})^{-1}$.

We emphasize that this expression does not depend on the specific choice of the DAG G. All graphs satisfying Condition 2 in Proposition 3.2 share the same edge set with non-zero edge weights, and thus, the parent set $p(i)$ always satisfies the back-door criterion in the corresponding unique minimal DAG (see Remark 3.3). □

Using Proposition 3.4, we can define the hypothesis space $\mathcal{M}_\psi \subset \mathcal{M}$, given by all covariance matrices $\Sigma \in \mathcal{M}$ that allow for a total causal-effect of size ψ via

$$\mathcal{M}_\psi := \{\Sigma \in \mathrm{PD}(d) : \exists \text{ complete DAG } G \text{ and } \sigma^2 > 0 \text{ with } \psi\sigma^2 = \Sigma_{j,i|p(i)} \text{ and } \sigma^2 = \Sigma_{k,k|p(k)} \quad \forall k = 1,\dots,d\}.$$

Put together, we have to solve the statistical testing problem

$$H_0^{(\psi)} : \Sigma \in \mathcal{M}_\psi \text{ against } H_1 : \Sigma \in \mathcal{M}\backslash\mathcal{M}_\psi, \tag{5}$$

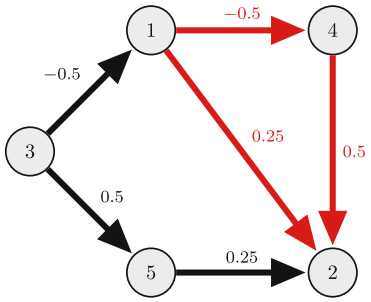
for all attainable $\psi \in \mathbb{R}$. Then, we invert the tests to construct valid confidence intervals for the total causal-effect. Similar to the union $\mathcal{M}$, we can write the hypothesis space as a union of single hypotheses over all (complete) DAGs, i.e., $H_0^{(\psi)} := \cup_{G\in\mathcal{G}(d)} H_0^{(\psi)}(G)$ with single hypotheses $H_0^{(\psi)}(G) : \Sigma \in \mathcal{M}_\psi(G)$, where

$$\mathcal{M}_\psi(G) := \{\Sigma \in \mathrm{PD}(d) : \exists\sigma^2 > 0 \text{ with } \psi\sigma^2 = \Sigma_{j,i|p(i)} \text{ and } \sigma^2 = \Sigma_{k,k|p(k)} \quad \forall k = 1,\dots,d\}. \tag{6}$$

Remark 3.5. It is clear that for $\psi \neq 0$, we only need to consider the union over all graphs with $i <_G j$. In contrast, for a zero effect, we have to consider all graphs, since the effects of multiple directed paths might "cancel" each other. Furthermore, for graphs with $j <_G i$, the constraint given by the size zero effect is not informative, since $j \in p(i)$ and consequently, $\Sigma_{j,i|p(i)} = 0$. Therefore, the corresponding hypothesis space has higher dimension.

Example 3.6. Consider the following LSEM and the corresponding (minimal) DAG:
There exist two directed paths from node 1 to node 2 in the corresponding (minimal) DAG, and thus, $1 <_G 2$ in all valid causal orderings. However, the effect of the two paths "cancel" each other, such that $C(1 \to 2) = 0.25 - 0.5^2 = 0$.

$$\begin{aligned}
X_1 &= -0.5X_3 + \varepsilon_1\\
X_2 &= 0.25X_1 + 0.5X_4 + 0.25X_5 + \varepsilon_2\\
X_3 &= \varepsilon_3\\
X_4 &= -0.5X_1 + \varepsilon_4\\
X_5 &= 0.5X_3 + \varepsilon_5
\end{aligned}$$



4 Testing for total causal-effects

For constructing confidence intervals for the total causal-effect, we invert tests of Hypothesis (5). In this section, we present two concrete tests based on likelihood ratio theory: (i) a classical constrained likelihood ratio test [16] and (ii) an approach based on the recently proposed split likelihood ratio test [17].

4.1 Constrained likelihood ratio test

The first option we consider is to form a classical likelihood ratio statistic for the testing Problem (5). Writing ℓ_n for the Gaussian log-likelihood function based on the sample $X^{(1)}, \dots, X^{(n)}$, the likelihood ratio statistic for (5) is

$$\check{\lambda}_n^{(\psi)} := 2\left(\sup_{\Sigma \in \mathcal{M}} \ell_n(\Sigma) - \sup_{\Sigma \in \mathcal{M}_\psi} \ell_n(\Sigma)\right). \tag{7}$$

In regular settings, the statistic may be compared to a chi-square quantile for an asymptotic test. Here, however, asymptotic distribution theory is difficult because both the null hypothesis and the alternative are intricate unions and do not satisfy the usual regularity assumptions. In the simplified bivariate setting, Strieder et al. [9] avoid the difficulties posed by the existence of singularities by testing modified problems in two different approaches. In the first approach, the hypothesis and alternative are relaxed to test an ordering of (conditional) variances. However, it is not feasible to optimize this modified testing problem in problems beyond the bivariate case. Furthermore, the mixture weights of the arising limit distribution, which is a mixture of chi-square distributions, are difficult to be determined in higher dimensions. The second approach relaxes the alternative to an unconstrained Gaussian model. In this modification, the statistic (7) is changed to a larger statistic for which the optimization of $\mathcal{M}$ is replaced by an optimization over the positive definite cone $\mathrm{PD}(d)$. This approach, however, may cause problems in the test inversion approach, as the union of tested null hypotheses no longer coincides with the alternative. In particular, small confidence regions may arise due to model misspecification.

In this article, we will derive our confidence region for the total causal-effect from tests that reject for too large values of the original statistic $\check{\lambda}_n^{(\psi)}$ from (7). However, we will consider relaxing the alternative $\mathcal{M}$ and use the theory of intersection union tests to obtain a simple yet effective upper bound on the distribution of $\check{\lambda}_n^{(\psi)}$.

To obtain this upper bound, first note that the hypothesis we need to test is an intricate union of single hypotheses given by $\bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_\psi(G)$. While we will later show that every single hypothesis defines a smooth submanifold, their union is not necessarily a smooth submanifold again. Nevertheless, the original test statistic (7) can then be written as $\check{\lambda}_n^{(\psi)} = \min_{G \in \mathcal{G}(d)} \check{\lambda}_n^{(\psi)}(G)$, with

$$\check{\lambda}_n^{(\psi)}(G) := 2\left(\sup_{\Sigma \in \mathcal{M}} \ell_n(\Sigma) - \sup_{\Sigma \in \mathcal{M}_\psi(G)} \ell_n(\Sigma)\right). \tag{8}$$

Thus, the original test statistic can be upper-bounded by every single test statistic $\check{\lambda}_n^{(\psi)}(G)$, and a valid-level α test of the union of single hypotheses can be constructed by testing every single hypothesis. We then reject the union null hypothesis if we reject all single hypotheses. A formal proof of this classical theory of intersection union tests can be found, e.g., in [14, Theorem 8.3.23].

In the second step, we relax the alternative to an unconstrained Gaussian model to obtain the following larger test statistic:

$$\lambda_n^{(\psi)}(G) := 2\left(\sup_{\Sigma \in \mathrm{PD}(d)} \ell_n(\Sigma) - \sup_{\Sigma \in \mathcal{M}_\psi(G)} \ell_n(\Sigma)\right).$$

Relaxing the alternative to an optimization over the entire cone of positive definite matrices $\mathrm{PD}(d)$ potentially increases the achieved likelihood and consequently upper-bounds each single test statistic $\check{\lambda}_n^{(\psi)}(G)$.

Combining both approximations yields an upper bound of the original test statistic $\check{\lambda}_n^{(\psi)}$ and, thus, a simple way to construct asymptotically conservative hypothesis tests. In the following, we state our main result by turning them into conservative confidence intervals for the total causal-effect $C(i \to j)$ via test inversion.

Theorem 4.1. *Let $\alpha \in (0,1)$. Then, an asymptotic $(1-\alpha)$-confidence set for the causal-effect $C(i \to j)$ is given by*:

$$C = \{\psi \in \mathbb{R} : \check{\lambda}_n^{(\psi)}(i <_G j) \leq \chi^2_{d,1-\alpha}\} \cup \{0 : \check{\lambda}_n^{(0)}(j <_G i) \leq \chi^2_{d-1,1-\alpha}\},$$

where $\check{\lambda}_n^{(\psi)}(i <_G j) := \min_{G \in \mathcal{G}(d) \,:\, i <_G j} \check{\lambda}_n^{(\psi)}(G)$ and $\check{\lambda}_n^{(0)}(j <_G i) := \min_{G \in \mathcal{G}(d) \,:\, j <_G i} \check{\lambda}_n^{(0)}(G)$.

Remark 4.2. Note that we only need to consider complete DAGs, which drastically reduces the number of single hypothesis tests that need to be evaluated and thus the computational burden. Furthermore, we need to respect graphs with $j <_G i$ only for zero-sized effects. Therefore, our proposed confidence regions may contain a non-zero interval and an isolated zero. In contrast to the bivariate case, however, zero effects can also correspond to orderings with $i <_G j$ due to cancellation (see also Remark 3.5).

Proof. Let $\psi \in \mathbb{R}$ and $\Sigma \in \mathcal{M}_\psi$. Since $\mathcal{M}_\psi = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_\psi(G)$, there exists a complete graph G, such that $\Sigma \in \mathcal{M}_\psi(G)$. If $i <_G j$, it follows from the introduced upper bound and Lemma 4.3

$$P_\Sigma(\check{\lambda}_n^{(\psi)} > \chi^2_{d,1-\alpha}) \leq P_\Sigma(\check{\lambda}_n^{(\psi)}(G) > \chi^2_{d,1-\alpha}) \leq P_\Sigma(\lambda_n^{(\psi)}(G) > \chi^2_{d,1-\alpha}) \to \alpha.$$

We obtain an analogous result for $j <_G i$ with Lemma 4.4. Using the test inversion approach yields the claim. □

In the following, we derive the asymptotic distribution of the upper bound $\lambda_n^{(\psi)}(G)$ of our test statistic under every single hypothesis $\mathcal{M}_\psi(G)$. As previously indicated, the cases $i <_G j$ versus $j <_G i$ define different dimensional hypothesis spaces; thus, we consider both instances separately.

First, we consider the case $i <_G j$ and assume without loss of generality that $(1, 2, \ldots, d)$ is the causal ordering. In this case, the single hypothesis can be written as $\mathcal{M}_\psi(G) = \{\Sigma \in \mathrm{PD}(d) : f_\psi(\Sigma|G) = 0\}$, with

$$f_\psi(\Sigma|G) := \begin{pmatrix} \Sigma_{j,i|p(i)} - \psi\Sigma_{1,1} \\ \Sigma_{2,2|p(2)} - \Sigma_{1,1} \\ \vdots \\ \Sigma_{d,d|p(d)} - \Sigma_{1,1} \end{pmatrix}.$$

The relaxed model space of all positive definite matrices $\mathrm{PD}(d)$ can be identified with an open subspace in $\mathbb{R}^{\frac{1}{2}(d^2+d)}$. The hypothesis space $\mathcal{M}_\psi(G)$ then defines a $\frac{1}{2}(d^2-d)$-dimensional submanifold of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$, since the Jacobian of f_ψ has full rank d. To see that, note its (non-zero) triangular structure for derivatives for $(\Sigma_{k,k})_{k=2,\ldots,d}$ in all rows except the first row. The first row, however, has a non-zero derivative for $\Sigma_{j,i}$, and zero derivatives for $(\Sigma_{k,k})_{k=j,\ldots,d}$.

Using the fact that this single hypothesis defines a smooth submanifold, we determine the limit distribution of the upper bound $\lambda_n^{(\psi)}(G)$.

Lemma 4.3. *Let $G \in \mathcal{G}(d)$ be a DAG with $i <_G j$ and $\psi \in \mathbb{R}$. Under the hypothesis $H_0^{(\psi)}(G)$, the likelihood ratio test statistic $\lambda_n^{\psi}(G)$ satisfies*

$$\lambda_n^{(\psi)}(G) \xrightarrow{\mathcal{D}} \chi^2_d.$$

Proof. Since the hypothesis space $\mathcal{M}_\psi(G)$ defines a $\frac{1}{2}(d^2-d)$-dimensional submanifold of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$, the tangent cone at every $\Sigma \in \mathcal{M}_\psi(G)$ is a $\frac{1}{2}(d^2-d)$-dimensional linear subspace of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$ and the claim follows. For details on this classical result, we refer to the study of Drton [18]. □

In the case $j <_G i$, the causal effect from i to j can only be zero. Therefore, we solely need to test the single hypothesis $H_0^{(0)}(G)$. We assume again without loss of generality that $(1, \ldots, d)$ is the underlying causal order. The polynomial constraint in Hypothesis (6) corresponding to the fixed-size effect is not informative, as the node j is included in $p(i)$, and thus, $\Sigma_{j,i|p(i)} = 0$. Therefore, the hypothesis $H_0^{(0)}(G)$ is solely defined by $d - 1$ polynomial constraints corresponding to the assumption of equal error variances, i.e., in this case, we have $\mathcal{M}_0(G) = \{\Sigma \in \mathrm{PD}(d) : f_0(\Sigma|G) = 0\}$, where

$$f_0(\Sigma|G) \coloneqq \begin{pmatrix} \Sigma_{2,2|p(2)} - \Sigma_{1,1} \\ \vdots \\ \Sigma_{d,d|p(d)} - \Sigma_{1,1} \end{pmatrix}.$$

Thus, similar to the previous case, we obtain a chi-square distribution as the limit distribution, however with fewer degrees of freedom.

Lemma 4.4. *Let $G \in \mathcal{G}(d)$ be a DAG with $j <_G i$. Under Hypothesis $H_0^{(0)}(G)$, the likelihood ratio test statistic $\lambda_n^{(0)}(G)$ satisfies*

$$\lambda_n^{(0)}(G) \xrightarrow{\mathcal{D}} \chi^2_{d-1}.$$

Proof. Analogously to Lemma 4.3. Note that the Jacobian has full rank $d - 1$ due to its triangular structure. □

With Theorem 4.1, we defined an asymptotically valid confidence interval for the total causal-effect that is able to capture both types of uncertainty: the uncertainty about the causal structure and the uncertainty about the numerical size of the effect. Besides the theoretical guarantees for the asymptotic coverage of the proposed confidence set, we demonstrate in the simulation section, that even in finite samples, the confidence set achieves the desired coverage probability. We emphasize again that our new approach uses the original test statistic (7) with an upper bound on its distribution. Thus, we avoid the issues that may arise under model misspecification in the (bivariate) method by Strieder et al. [9] due to contrasting a hypothesis restricted to equal error variances with a general alternative.

In the following, we propose an alternative framework for constructing hypothesis tests for (5) based on the theory of universal inference [17]. While this framework leads to more conservative intervals, it comes with a finite sample guarantee.

4.2 Split likelihood ratio test

Likelihood ratio tests provide a powerful tool for constructing hypothesis tests that are calibrated via asymptotic distribution theory. However, in this context, the needed distribution-theoretic insights are difficult to obtain and we adopted stochastic upper bounds. An interesting alternative is to conduct the needed tests in the framework of universal inference that was introduced by Wasserman et al. [17]. Their method uses a modification of the likelihood ratio statistic, called split likelihood ratio, based on a data-splitting approach. Due to the independence of the split data, one can apply a universal critical value, instead of relying on asymptotic distribution theory. Type I error control is then guaranteed by an application of Markov's inequality. In the following, we use their methodology for constructing conservative confidence regions for total causal-effects with finite sample guarantee.

To construct a split likelihood ratio test for the test Problem (5), we proceed as follows. First, we split the data into two subsets $D_0 = \{X^{(1)}, \ldots, X^{(k)}\}$ and $D_1 = \{X^{(k+1)}, \ldots, X^{(n)}\}$ and define the log-likelihood functions:

$$\ell_j(\Sigma) \coloneqq \sum_{X^{(i)} \in D_j} \log p(X^{(i)}|\Sigma), \quad j = 0, 1,$$

based on D_0 and D_1, respectively. Let $\tilde{\Sigma}_1 \coloneqq \operatorname{argmax}_{\Sigma \in \mathcal{M}} \ell_1(\Sigma)$ be the maximum-likelihood estimator of Σ restricted to LSEMs with equal error variances based on D_1. Then, the split likelihood ratio test statistic is defined as:

$$\tilde{\lambda}_n^{(\psi)} := 2(\ell_0(\tilde{\Sigma}_1) - \sup_{\Sigma \in \mathcal{M}_\psi} \ell_0(\Sigma)).$$

The insight of Wasserman, Ramdas, and Balakrishnan [17] is that for any fixed $\Sigma^* \in \mathrm{PD}(d)$, it holds that

$$\mathbb{E}_\Sigma\left[\prod_{X^{(i)} \in D_0} \frac{p(X^{(i)}|\Sigma^*)}{p(X^{(i)}|\Sigma)}\right] \leq \int \prod_{X^{(i)} \in D_0} p(X^{(i)}|\Sigma^*) = 1,$$

and, thus, Markov's inequality implies that for any $\alpha \in (0,1)$, under $H_0^{(\psi)} : \Sigma \in \mathcal{M}_\psi$,

$$\begin{aligned} P_\Sigma(\tilde{\lambda}_n^{(\psi)} > -2\log\alpha) &\leq \alpha \mathbb{E}_\Sigma\left[\prod_{X^{(i)} \in D_0} \frac{p(X^{(i)}|\tilde{\Sigma}_1)}{\sup_{\Sigma \in \mathcal{M}_\psi} p(X^{(i)}|\Sigma)}\right] \\ &\leq \alpha \mathbb{E}_\Sigma\left[\mathbb{E}_\Sigma\left[\prod_{X^{(i)} \in D_0} \frac{p(X^{(i)}|\tilde{\Sigma}_1)}{\sup_{\Sigma \in \mathcal{M}_\psi} p(X^{(i)}|\Sigma)} \middle| D_1\right]\right] \leq \alpha. \end{aligned} \tag{9}$$

Thus, the decision rule

$$\text{Reject } H_0^{(\psi)} \quad \text{if } \tilde{\lambda}_n^{(\psi)} > -2\log\alpha$$

constitutes a valid-level α test. This test holds level in finite samples and without any regularity conditions. Thus, we can define the following confidence region for the total causal-effect based on the split likelihood ratio test.

Theorem 4.5. *Let $\alpha \in (0,1)$. Then, a $(1-\alpha)$-confidence set for the causal-effect $C(i \to j)$ is given by*:

$$C = \{\psi \in \mathbb{R} : \tilde{\lambda}_n^{(\psi)}(i <_G j) \leq -2\log(\alpha)\} \cup \{0 : \tilde{\lambda}_n^{(0)}(j <_G i) \leq -2\log(\alpha)\},$$

where $\tilde{\lambda}_n^{(\psi)}(i <_G j) := \min_{G \in \mathcal{G}(d) \,:\, i <_G j} \tilde{\lambda}_n^{(\psi)}(G)$ and $\tilde{\lambda}_n^{(0)}(j <_G i) := \min_{G \in \mathcal{G}(d) \,:\, j <_G i} \tilde{\lambda}_n^{(0)}(G)$ and

$$\tilde{\lambda}_n^{(\psi)}(G) := 2(\ell_0(\tilde{\Sigma}_1) - \sup_{\Sigma \in \mathcal{M}_\psi(G)} \ell_0(\Sigma)).$$

Remark 4.6. Similar to the asymptotic confidence set given by Theorem 4.1, we need to consider all graphs when testing for zero effects. Thus, the confidence region may consist of a non-zero interval and a disconnected zero that reflects the remaining uncertainty about the existence of an effect.

Proof. This result follows directly from the test inversion approach and (9). Note that for non-zero effects, we only need to consider graphs with $i <_G j$. □

5 Computational details and numerical experiments

In this section, we present computational details as well as shortcuts in our proposed algorithm that significantly improve the computation time. We also present a simulation study in which we analyze and compare the coverage probabilities and performance of the different proposed testing procedures.

5.1 Algorithm and computational shortcuts

We focus on the algorithm for calculating the likelihood ratio-based confidence regions introduced in Theorem 4.1, denoted with `LRT`. By adjusting the critical values in the algorithm and incorporating the data-splitting procedure, the split likelihood ratio-based confidence regions can be computed analogously. As the number of possible underlying causal structures, i.e., the number of DAGs, grows superexponentially with the number of nodes, we need to carefully use combinatorial shortcuts in order to be able to compute our proposed confidence intervals for the total causal-effect.

Our algorithm, presented as Algorithm 1, consists of two subroutines. The first routine `possible.order` (Algorithm 2) immediately rejects all implausible causal orderings and, thus, reduces the set of possible causal orderings in which we subsequently test for all attainable effects with the second routine `testeffect` (Algorithm 3). Carefully reducing the space of plausible causal orderings drastically decreases the computation time and allows us to compute our proposed confidence intervals, which take uncertainty over all possible DAGs into account, in a reasonable time for a moderate (but far from trivial) number of involved variables. Precise computation times can be found in Section 5.3, which presents the results of a simulation study.

Algorithm 1 Returns `LRT` confidence region for $C(i \to j)$

Input: data, level α, stepsize s
(orderings, likelihood, effect) ←`possible.order`(data, α) ▷ all plausible orderings
L1← likelihood[1] ▷ maximum likelihood alternative
startvalues ←unique(effect)
while startvalues is not empty **do** ▷ start with effects from unrestricted MLEs
  (leftbound, rightbound) ←min(startvalues)
  (left, right) ← TRUE
  while left **do**
    leftbound ← leftbound−s
    left ←`test.effect`(data, leftbound, orderings, L1, α) ▷ test for size leftbound effect
  end while
  while right **do**
    rightbound ← rightbound+s
    right ←`test.effect`(data, rightbound, orderings, L1, α)
  end while
  append (leftbound+$\frac{s}{2}$,rightbound−$\frac{s}{2}$) to intervals ▷possibly multiple intervals
  remove all values smaller than rightbound from startvalues
end while
zero ←`test.effect`(data, 0, orderings, L1, α) ▷test for size zero effect
return (zero, intervals)

In order to reduce the space of plausible graphical structures as much as possible, we note again that we do not need to consider all possible DAGs but merely complete ones, represented by all possible causal orderings. This immensely decreases the worst-case number of single hypotheses we need to perform to reject an effect of size ψ to d factorial. Our subroutine `possible.order` (Algorithm 2) further reduces the number of plausible causal orderings as follows.

First, we quickly approximate the maximum likelihood under the alternative with the maximum likelihood $\widehat{\mathrm{L1}}$ obtained under the ordering obtained from the method in the study by Chen et al. [10] that sorts conditional variances. Subsequently, we calculate the maximum likelihood of all possible orderings without restrictions on the causal-effect and immediately reject implausible orderings using the (largest) critical value $\chi^2_{d,1-\alpha}$. Note that we can search through the space of orderings recursively starting from the first node and immediately reject (partial) orderings if the corresponding (partial) maximum likelihood already exceeds the threshold. Furthermore, for the subsequent testing procedure, the specific ordering of the parent set $p(i)$ is not relevant. Thus, for all orderings that only differ in the specific ordering of $p(i)$, we subsequently just test with the parent order that achieves the maximum likelihood. Additionally, we only need to test for a zero effect in the maximum-likelihood ordering out of all orderings with $j <_{\mathcal{G}} i$. Finally, we include an attainable effect in our confidence set if it cannot be rejected for at least one of the possible causal orderings using the subroutine `testeffect`, Algorithm 3. Therefore, we do not need to continue testing for a specific value of the effect if we find an ordering such that we cannot reject the value. Thus, first sorting the causal orderings by their maximum likelihood further improves the computation time.

Algorithm 2 `possible.order()` returns plausible causal orderings

Input: data, level α
crit.val $\leftarrow$ qchisq$(1-\alpha, d)$ ▷$1-\alpha$ quantile of χ^2_d-distr.
$\hat{\Sigma} \leftarrow$ cov(data)
for $1:d$ **do**
  append which.min$(\hat{\Sigma}_{\cdot\cdot|\text{var.order}})$ to var.order ▷ordering of cond. variances
end for
$\widehat{\text{L1}} \leftarrow$`get.likelihood`(var.order) ▷(unrestricted) maximum likelihood
for $1:d$ **do**
  for order in orderings **do**
    for v in setdiff$(1:d,$ order$)$ **do**
      part.order $\leftarrow$ c(order,v)
      likelihood $\leftarrow$`get.likelihood`(part.order) ▷(partial) maximum likelihood
      if $2(\widehat{\text{L1}}-$likelihood$) \leq$ crit.val **then**
        append part.order to orderings
        if v equals i **then**
          effect $\leftarrow \hat{\Sigma}_{i,j|\text{order}}/\hat{\Sigma}_{i,i|\text{order}}$ ▷causal-effect of maximum likelihood
        end if
      end if
    end for
    remove order from orderings
  end for ▷order only relevant after i
  choose order with max likelihood out of all orders in orderings that do not contain i and are defined on the same set of nodes
end for ▷keep only max likelihood for zero effect
choose order with max likelihood out of all orders in orderings with $j <_{\mathcal{G}} i$
sort orderings for decreasing likelihood
return (orderings, likelihood, effect)

Algorithm 3. `test.effect()` tests causal-effect $C(i \to j) = \psi$

Input: data, causal-effect ψ, orderings, likelihood alternative L1, level α
for order in orderings **do**
  if $j <_{\mathcal{G}} i$ **then**
    crit.val $\leftarrow$ qchisq$(1-\alpha, d-1)$
    if ψ equals zero **then**
      likelihood $\leftarrow$`get.likelihood`(order) ▷(unrestricted) maximum likelihood
    else
      likelihood $\leftarrow -\infty$ ▷reject
    end if
  else
    crit.val $\leftarrow$ qchisq$(1-\alpha, d)$
    likelihood $\leftarrow$`get.likelihood`(order,ψ) ▷(restricted) maximum likelihood
  end if
  if 2(L1-likelihood)$\leq$ crit.val **then**
    return TRUE
  end if
end for
return FALSE

We start testing within the plausible causal orderings with the minimal causal-effect that corresponds to the maximum-likelihood estimates. Then, we step-wise decrease the value until we reject in all orderings. The last value we did not reject is a lower bound for our confidence region. We repeat the procedure with increasing steps until we reject in all orderings and the last value we did not reject is an upper bound for this part of our confidence region. If there remain causal-effects from the maximum-likelihood estimates that exceed this upper bound, we repeat this procedure. Thus, using this strategy, we might obtain a confidence region that consists of multiple intervals. Finally, we test whether we have to include zero in our confidence interval, i.e., whether there remains uncertainty about the existence of an effect.

5.2 Maximum-likelihood estimation

Implementing the proposed algorithm involves the routine `get.likelihood`, which maximizes the centered Gaussian (log-)likelihood function $\ell_n(\Sigma)$ for a given causal ordering. The function is given by:

$$\frac{2}{n}\ell_n(\Sigma) = -\log\det(2\pi\Sigma) - \operatorname{tr}(\Sigma^{-1}\hat{\Sigma}),$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{l=1}^{n} X^{(l)}(X^{(l)})^T$ is the empirical covariance. Due to the underlying assumption of homoscedasticity across all interacting variables, the set of possible distributions can be uniquely identified via (2). Therefore, equivalently, we maximize

$$\frac{2}{n}\ell_n(B,\sigma^2) = -d\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\operatorname{tr}((I_d - B)^T(I_d - B)\hat{\Sigma}),$$

using the edge matrix $B \in \mathbb{R}^G$ and the common variance $\sigma^2 > 0$. The zero entries of the matrix B are implied by the given causal ordering. Furthermore, due to the acyclic structure, B is permutation similar to a lower triangular matrix. Straightforward calculations yield that the maximum of $\ell_n(B,\sigma^2)$ over the equal variance $\sigma^2 > 0$ is given by:

$$\sup_{\sigma^2>0}\ell_n(B,\sigma^2) = -\frac{nd}{2}\log\left(\frac{2\pi}{d}\operatorname{tr}((I_d - B)^T(I_d - B)\hat{\Sigma})\right) - \frac{nd}{2}. \tag{10}$$

Maximizing equation (10) over $B \in \mathbb{R}^G$ is then equivalent to minimizing

$$\operatorname{tr}((I_d - B)^T(I_d - B)\hat{\Sigma}) = \frac{1}{n}\sum_{k=1}^{d}\sum_{l=1}^{n}(X_k^{(l)} - \sum_{p\in p(k)}\beta_{k,p}X_p^{(l)})^2. \tag{11}$$

Thus, in the unrestricted case without constraints for a fixed-size causal-effect, the problem is equivalent to solving d separate linear least-squares problems with minimizer:

$$\hat{\beta}^T_{k,p(k)} = (\hat{\Sigma}_{p(k),p(k)})^{-1}\hat{\Sigma}_{p(k),k}, \quad \text{for} \quad k = 1,\dots,d, \tag{12}$$

and corresponding minimum

$$\frac{1}{n}\sum_{k=1}^{d}\sum_{l=1}^{n}\left(X_k^{(l)} - \sum_{p\in p(k)}\hat{\beta}_{k,p}X_p^{(l)}\right)^2 = \sum_{k=1}^{d}\hat{\Sigma}_{k,k|p(k)}.$$

Furthermore, in this case, we can calculate the maximum likelihood of any partial ordering that starts from the source node, since it collapses into subproblems. Thus, we can reject all orderings that start with a given partial ordering if the corresponding partial maximum likelihood already exceeds the threshold.

In addition, we need to calculate the maximum-likelihood estimates under each single hypothesis $H_0^{(\psi)}(G)$. For $j <_G i$, we only test Hypothesis $H_0^{(0)}(G)$. However, in that case, the constraint of a size zero causal-effect $C(i \to j) = 0$ is not restrictive, and thus the minimizer is similarly given by (12). Therefore, we assume in the following $i <_G j$. Analogously to the previous calculations, maximizing the Gaussian likelihood for a given

causal ordering and a fixed total causal-effect can be solved by minimizing the linear least-squares Problem (11). However, the constraint of a fixed total causal-effect $C(i \to j) = \psi$ links a subset of the regression parameters, and we cannot solve all d linear least-squares problems separately.

Recalling the path properties of the total causal-effect (see Remark 2.1), the assumption $C(i \to j) = (I_d - B)^{-1}_{j,i} = \psi$ restricts all $\beta_{k,p(k)}$ for $k \in p(j) \cup \{j\} \backslash p(i) \cup \{i\}$. This follows by observing that

$$(I_d - B)^{-1}_{j,i} = \sum_{k=1}^{d-1} (B^k)_{j,i}$$

and further inductively

$$(B^k)_{j,i} = \sum_{p \in p(j) \backslash p(i)} \beta_{j,p} (B^{k-1})_{p,i}.$$

For all other $k \notin p(j) \cup \{j\} \backslash p(i) \cup \{i\}$, the corresponding part of the optimization Problem (11) can be solved separately with the (unrestricted) minimizer given by (12). However, it remains to minimize the following nontrivial joint least-squares problem:

$$\sum_{k \in p(j) \cup \{j\} \backslash p(i) \cup \{i\}} \sum_{l=1}^{n} (X_k^{(l)} - \sum_{p \in p(k)} \beta_{k,p} X_p^{(l)})^2, \quad \text{where } \beta_{j,i} = \psi - \sum_{m=2}^{d-1} (B^m)_{j,i}.$$

Note that we can reduce the constraint to the submatrix B_s, which is formed by selecting all rows and columns corresponding to the set $p(j) \cup \{j\} \backslash p(i)$, since $(B^m)_{j,i} = (B_s^m)_{j,i}$. Then, the constraint yields $\beta_{j,i}$ as a function of the remaining $\beta_{k,p(k)}$ for $k \in p(j) \cup \{j\} \backslash p(i) \cup \{i\}$, and we solve the remaining least-squares problem numerically with quasi-Newton methods.

5.3 Simulations

In this section, we present the results of a simulation study that compares the coverage frequencies and the widths of the proposed confidence regions as well as their computing times. Our simulation experiments are designed as follows. We generate synthetic data sets based on LSEMs with Gaussian errors and equal variances associated with randomly selected directed acyclic graphs in a three-step procedure. First, we randomly select a permutation of d nodes, representing the causal ordering of $(X_1, \ldots, X_d)$. Second, we generate edge weights for all edges in the complete graph that corresponds to the chosen causal ordering according to a normal distribution $N(\beta, 0.1)$. In that way, β reflects the average strength of the direct effects in the graph, and thus, smaller values indicate higher uncertainty about the underlying causal structure. In the next step, we prune the complete graph, i.e., we include an edge with a probability of 0.9 for dense and 0.5 for sparse graphs. Finally, we generate n samples according to the LSEM, represented by the chosen graph, with standard normal distributed errors. For a range of edge weights β, sample sizes n and dimensions d, we generated multiple independent data sets and determined the confidence sets for the total causal-effect of X_1 on X_2 with confidence level $\alpha = 0.05$ using the different proposed methods (`LRT` given by Theorem 4.1 and `SLRT` given by Theorem 4.5). We repeated this procedure twice, for the case of a true non-zero effect and the case of no effect. We chose the tuning parameter of the `SLRT` interval, the splitting ratio, optimal in the sense of Strieder and Drton [19].

We report the observed empirical coverage probabilities for six-dimensional graphs in Table 1. All proposed methods achieve the desired coverage frequency of 0.95 and seem to be conservative. This is expected for the split likelihood ratio method since it is based solely on Markov's inequality. However, it is also not surprising that the `LRT` method is conservative, considering the use of intersection union testing and the fact that we set critical values based on the relaxed alternative. For comparison, we also included classical bootstrapping results in our simulation study. The bootstrapping method uses `GDS`, an established causal discovery algorithm for LSEMs with equal variance, proposed by Peters and Bühlmann [8]. For each resampled data set, the `GDS` method uses a greedy search algorithm to maximize the likelihood and estimate the total causal-effect. The results validate the theoretical findings in the studies by Strieder et al. [9] and Drton and Williams [13].

Table 1: Empirical coverage of 95% confidence intervals for the total causal-effect of X_1 on X_2 in randomly generated 6-dim. DAGs (1,000 replications)

Method	$n \backslash \beta$	True effect						No true effect					
		Sparse			Dense			Sparse			Dense		
		0.05	0.1	0.5	0.05	0.1	0.5	0.05	0.1	0.5	0.05	0.1	0.5
LRT	100	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1,000	0.98	0.98	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
SLRT	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1,000	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Bootstrap	100	**0.70**	**0.75**	0.97	**0.75**	**0.80**	0.97	1.00	1.00	1.00	1.00	1.00	1.00
	500	**0.72**	**0.77**	0.97	**0.78**	**0.85**	0.98	1.00	1.00	1.00	1.00	0.99	1.00
	1,000	**0.76**	**0.80**	0.95	**0.79**	**0.88**	0.98	0.99	0.99	1.00	1.00	1.00	1.00

Bootstrapping methods do not correctly account for uncertainty in the causal structure and do not achieve the required coverage frequency, as highlighted in bold.

In addition to the coverage frequencies, we investigated how conclusive the confidence regions are, i.e., how informative are intervals for practitioners in deciding about the existence and size of the causal-effects, while still providing a valid uncertainty assessment. To this, we show in Figure 1 the average width of the non-zero part of the confidence sets. We report the results for dimension $d = 6$, where the data were generated according to random DAGs with expected edge weight $\beta = 0.5$, against the sample size. In contrast to the bootstrapping method, our proposed LRT method constitutes a valid confidence region. Nevertheless, even in the considered setting with edge weight $\beta = 0.5$, where all variants achieved the desired coverage, our method significantly outperforms bootstrapping in the dense setting, while performing similarly in the sparse setting. In contrast, the split likelihood ratio tests yield wider confidence intervals. We also note that the confidence intervals are generally wider in the dense setting, i.e., the remaining uncertainty about the numerical size of the effects is higher for dense DAGs. Our intuition here is that in dense DAGs, more directed paths contribute to the total causal-effect; thus, more edge weights are involved with remaining uncertainty. Figure 1 shows that our proposed methods successfully help to locate the numerical size of existing causal-effects while correctly quantifying remaining uncertainty.

In addition to locating the numerical size of existing effects, an informative confidence interval should help in deciding whether there exists remaining uncertainty about the existence of an effect. Therefore, to

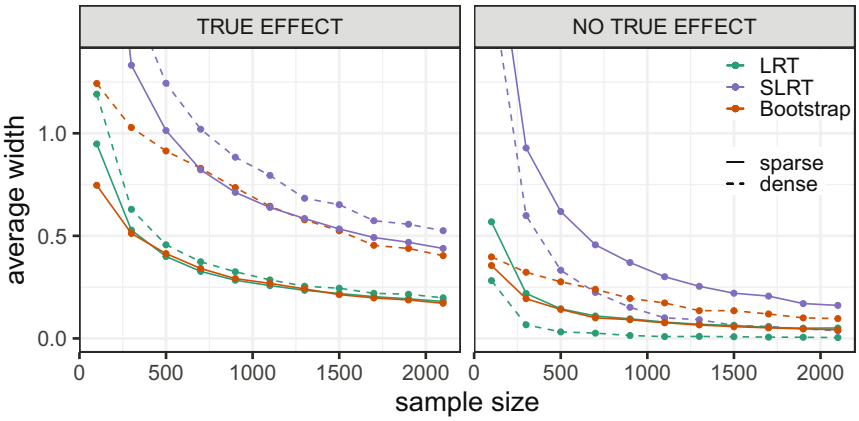


Figure 1: Mean width of 95% confidence intervals for the total causal-effect of X_1 on X_2 in randomly generated 6-dim. DAGs (1,000 replications).

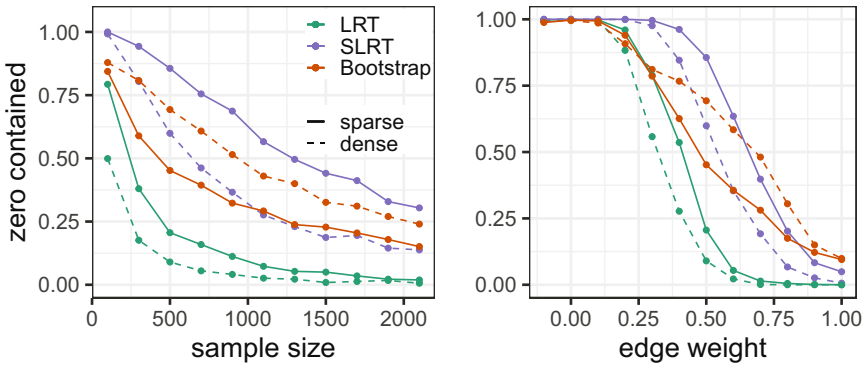


Figure 2: Percentages of times zero contained in 95% confidence intervals in randomly generated 6-dim. DAGs with non-zero effect (1,000 replications): (Left) against sample size and (Right) against expected edge weights.

investigate whether the proposed methods pick up on the causal direction of the effect between two variables, we plotted the proportions of times zero is included in the confidence sets when there is a true non-zero effect. Figure 2 shows the results for dimensions $d = 6$ with expected edge weight $\beta = 0.5$ against the sample size and for the sample size $n = 500$ against the expected edge weight. The LRT method eliminates the remaining uncertainty about the direction of the causal-effect the quickest with increasing sample size, significantly outperforming the bootstrapping method and the split likelihood ratio tests. As intuitively expected, for higher average edge weights, there is less uncertainty about the underlying causal structure, and thus, the methods more often exclude the possibility of no causal-effect.

Furthermore, one observes that in contrast to the bootstrapping method, the confidence regions of our proposed methods contain the zero less often in the dense setting compared to those in the sparse setting. This stems from the fact that there exists less uncertainty about the causal ordering for dense DAGs, and thus, the remaining uncertainty about the direction of the involved effects is lower than in the sparse setting. While there might be less uncertainty about the numerical size of the effect and thus, narrower confidence regions in the sparse setting, the uncertainty about the causal structure is higher.

We obtain similar results for larger causal structures with 12 involved variables. Figure 3 shows the percentages of times our proposed LRT-confidence regions contain zero as well as the mean width of the non-zero part. The data were generated with an expected edge weight $\beta = 0.5$ and increasing sample sizes. In the case of no true effect, the confidence intervals always correctly contain zero. The average width of the non-zero part is close to zero, even for low sample sizes. If there is a non-zero total causal-effect, the remaining uncertainty about the direction of the effect is already eliminated in samples of size around 2,000. We highlight again that the remaining uncertainty about the causal structure (reflected by the percentage of times zero is

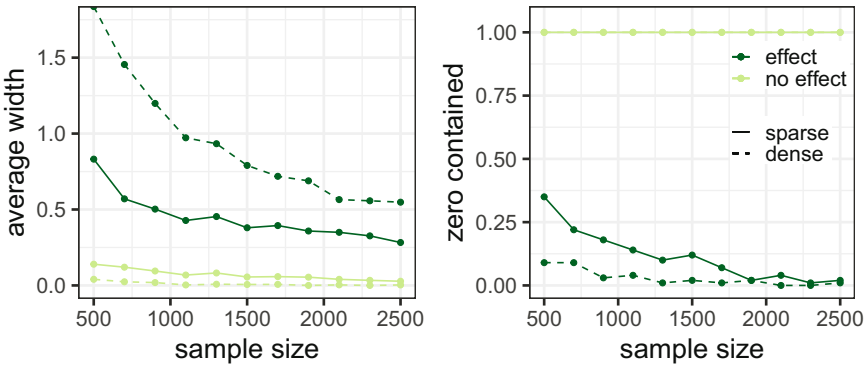


Figure 3: Mean width and percentages of times zero contained in 95% confidence intervals for the total causal-effect of X_1 on X_2 in randomly generated 12-dim. DAGs (100 replications).

contained in the confidence interval) is higher in the sparse setting. In contrast, the remaining uncertainty about the numerical size of the effect, reflected by the average width of the non-zero part of the confidence intervals, is lower.

In summary, the proposed framework for constructing confidence intervals for total causal-effects successfully detects the direction and the numerical size of causal-effects and correctly quantifies the remaining uncertainty in causal structure and effect size.

Figure 4 compares the average computation times over 100 randomly drawn DAGs. As the number of possible causal structures grows superexponentially with the dimension, the computation times for our proposed confidence intervals naturally increase. Since the intervals take uncertainty over all possible structures into account, the computation times similarly increase superexponentially. However, Figure 4 shows that it is feasible to compute this complete uncertainty over all causal structures for moderate dimensions with lower computation times than the bootstrapping method. We plotted the computation times for an expected edge weight $\beta = 0.5$ and sample size $n = 1{,}000$ against the dimension (up to $d = 12$). Contrasting the bootstrapping procedure based on the greedy search algorithm GDS, the computation times for our proposed confidence intervals are lower in the dense compared to those in the sparse setting. This is not unexpected, as dense DAGs generally coincide with less uncertainty in the causal structure and thus lead to fewer plausible causal orderings passed on to the testing procedure. Similarly, if there is no true effect, computation times are lower since we need to test within fewer plausible causal orderings. In general, the computation times are mainly determined by reducing the set of plausible orderings with Algorithm 2 and the uncertainty in the causal structure within the data.

Furthermore, Figure 4 also includes the average computation times for increasing sample sizes with expected edge weight $\beta = 0.5$ and dimension $d = 6$. In contrast to the bootstrapping procedure, increasing the sample size reduces the computation times for our proposed confidence intervals. More available data reduce the uncertainty about the underlying structure, which reduces the computation time.

Our framework is based on the assumption of an underlying Gaussian LSEM with equal error variances. We emphasize that if the true data-generating mechanism follows a Gaussian LSEM with differing error variances, the causal structure and, consequently, the true causal-effects are generally not identifiable. Therefore, our task of constructing a confidence interval for this causal-effect is not well defined. However, in this setting, our method targets the causal-effect in an approximation of the data-generating mechanism under the equal error variances assumption. With practical applications in mind, we investigated this behavior and, thus, the sensitivity of our method toward deviations from equal error variances. We generated data according to the same procedure as described earlier (for $d = 6$, $\beta = 0.1$, and $n = 500$), but with error variances sampled uniformly from $[1 - 0.5v, 1 + 0.5v]$. In that way, v quantifies the degree of deviation from homoscedasticity across the interacting variables, ranging from 0 to 1.8. The empirical coverage probabilities in Figure 5 show that our framework is rather robust to small deviations from equal error variances. For small

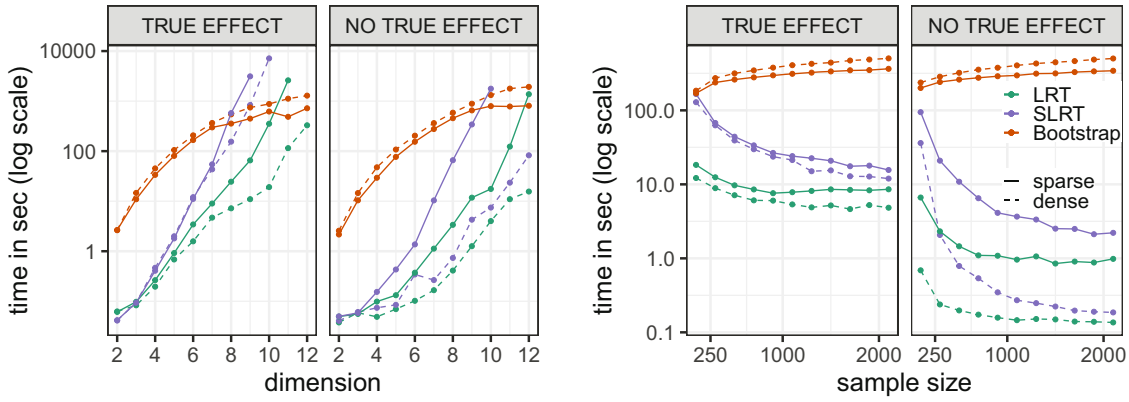


Figure 4: Mean computation times of 95% confidence intervals for causal-effect in randomly generated DAGs in seconds (100 replications): (Left) against dimension and (Right) against sample size.

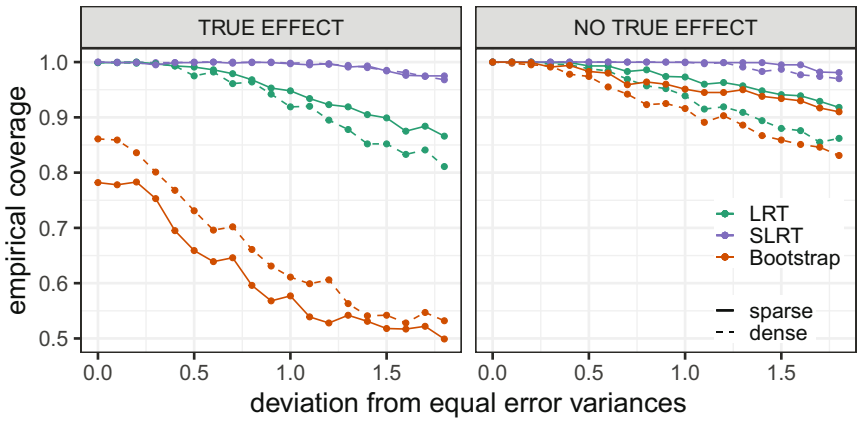


Figure 5: Empirical coverage of 95% confidence intervals for the total causal-effect of X_1 on X_2 in randomly generated 6-dim. DAGs under departure from equal error variances (1,000 replications).

deviations, the targeted causal-effect in the approximated equal error variance model is close to the true causal-effect, and thus, our confidence intervals, nevertheless, cover the (unidentifiable) true effect.

In conclusion, the likelihood ratio method performs best in our simulation studies and yields the most informative confidence intervals. Nonetheless, while the split likelihood ratio tests are more conservative, we emphasize that it is the only method with theoretical finite sample guarantees.

5.4 Real data example

To illustrate the importance of correctly accounting for uncertainty in the causal structure, we consider the frequently studied Sachs protein data [20]. This collection of 14 data sets comprises expression levels for 11 proteins and phospholipids in human T-cells obtained under different experimental conditions. In applications involving variables from similar domains and measurements obtained by similar techniques, the assumption of equal error variances offers a simple way to conduct exploratory analyzes of cause-effect relations, also in light of the robustness of our method toward small deviations from equal variance that was seen in our simulations. In our data analyzis, we compared our LRT method against naive bootstrapping with the GDS method [8] as well as bootstrapping with the LiNGAM method [21]. The LiNGAM method explicitly assumes non-Gaussian errors to ensure identifiability. We considered the first of the 14 data sets with a sample size of 853 and focused on a subset of 10 proteins (excluding PKA) that have measurements on a similar scale. The protein PKA is a highly variable source node in the ground truth causal DAG reported in the study by Sachs et al. [20]. To account for its influence, we first performed linear regressions of each protein on PKA and continued our analyzis using the residuals.

Figure 6 shows the calculated 95% confidence intervals for six exemplary causal-effects. The regression coefficient of X_i when regressing X_j on $(X_i, X_{p(i)})$, where the conventionally accepted ground truth structure defines the parent set, is seen as the true target effect $C(i \to j)$ and marked in red. In total, our proposed confidence intervals cover 84 of the 90 pairwise causal-effects among the 10 involved variables. In contrast, naive bootstrapping with GDS and with LiNGAM cover only 66 and 72, respectively. Our confidence intervals indicate that, even within our strict modeling assumptions, there is substantial structure uncertainty. Out of the 90 calculated intervals, 67 contain zero as well as non-zero effects, thus reflecting uncertainty about the direction and existence of effects. In comparison, the naive methods with GDS and with LiNGAM contain both directions in only 35 and 55 cases, respectively. Thus, while the bootstrapping methods yield narrower intervals on average, they do not correctly account for structure uncertainty and tend to miss more minor target effects.

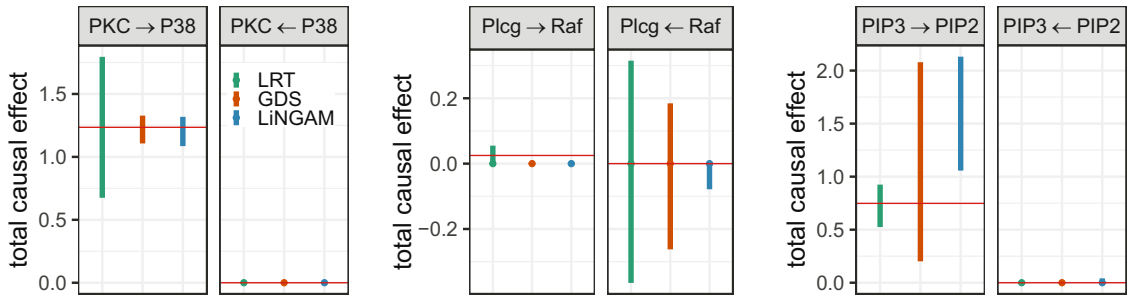


Figure 6: 95% confidence intervals for the total causal-effect between proteins from the real world expression data.

6 Discussion

In this article, we raise the problem of rigorously quantifying uncertainty in causal inference when the causal structure is unknown. We demonstrate the intricacies of this task and propose a precise solution for constructing confidence regions for total causal-effects, which capture both types of uncertainty: numerical size of effects and causal structure. Using a test inversion approach with intersection union tests and exploiting combinatorial shortcuts, our proposed framework mathematically rigorously quantifies the remaining uncertainty in the concrete setting of LSEMs with homoscedastic Gaussian errors across all interacting variables.

Our proposal of leveraging test inversions of joint tests for causal structure and size of effect to obtain confidence regions for causal-effect estimates under structure uncertainty has a general nature and with suitable statistical tests could be extended to other modeling assumptions (see Strieder et al.'s study [9] for results from a heuristic for the example of bivariate LSEMs with non-Gaussian errors). Moreover, our framework could be adjusted to target different (causal) quantities by reformulating the corresponding constraint.

Assuming LSEMs with equal error variances ensures identifiability. Thus, a confidence region corresponds to a set of covariance matrices, which all uniquely entail a corresponding total causal-effect. Without equal error variances (or rather identifiability in general), this unique correspondence does not hold. While our computational framework of focusing on causal orders and complete graphs might be adapted to consider a set of plausible covariance matrices, every covariance matrix then entails a set of possible causal-effects that always contains zero. Thus, the corresponding confidence regions are not conclusive about the direction of effects. In contrast, we demonstrated that under the assumption of equal error variances, our proposed confidence regions not only correctly quantify the remaining uncertainty in causal structure and numerical effect size, but also provide conclusive information that can help practitioners to decide about existence and numerical size of effects.

The computational shortcuts we adopt allow us to quantify structure uncertainty over all possible DAGs among a moderate number of nodes. For higher-dimensional systems, it is not feasible to consider all possible DAGs given that there already exist more DAGs on 22 nodes (10^{87}) than estimated atoms in the universe. However, our method can still be used by practitioners as a valuable tool for analyzing higher-dimensional causal systems by partially quantifying the uncertainty in the causal structure.

In the *status quo* scenario, practitioners would either confer with experts that provide the underlying causal structure or use structure learning algorithms to estimate the DAG from available data. Starting from this inferred model, classical statistical methods can then be used to calculate confidence intervals for the involved causal parameters. These confidence intervals quantify uncertainty in the numerical size of the causal-effect conditional on the inferred model. However, they do not incorporate uncertainty about the underlying causal structure and the (data-driven) model choice. While in higher-dimensional systems, it is not computationally tractable to quantify structure uncertainty by considering all DAGs, our framework can still be used as a compromise in between. Rather than solely relying on the full structure provided by experts, our method can be used to consider a set of plausible causal structures. This way practitioners can incorporate varying degrees of belief of experts in different parts of the structure by fixing parts with high confidence but still considering the remaining uncertainty over the rest. Along similar lines, instead of focusing on a single

output of causal learning algorithms, practitioners can use our method to a set of (most likely) causal structures and incorporate some degree of structure uncertainty in causal-effect estimates.

In summary, we consider our study of total causal-effect in linear causal models with equal error variances as the beginning of a new line of research on confidence sets under structure uncertainty. We anticipate fruitful generalizations in many directions including not only other model settings such as the linear non-Gaussian case, or additive noise models, but also different causal parameters other than total effects. From this perspective, while this article develops our ideas in a very concrete setting, it sets the scene for various follow-up works in providing rigorously justified confidence regions in causal-effect estimation when the causal structure itself is learned from data.

Funding information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 83818). Furthermore, this work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content.

Conflict of interest: The authors state no conflict of interest.

References

[1] Pearl J. Causality. Models, reasoning, and inference. 2nd ed. Cambridge: Cambridge University Press; 2009.
[2] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. Adaptive computation and machine learning. Cambridge, MA: MIT Press; 2000.
[3] Hoyer PO, Hyttinen A. Bayesian discovery of linear acyclic causal models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI'09. Arlington, Virginia, USA: AUAI Press; 2009. p. 240–8.
[4] Claassen T, Heskes T. A Bayesian approach to constraint based causal inference. In: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. UAI'12. Arlington, Virginia, USA: AUAI Press; 2012. p. 207–16.
[5] Cao X, Khare K, Ghosh M. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. Ann Statist. 2019;47(1):319–48.
[6] Peters J, Janzing D, Schölkopf B. Elements of causal inference. Adaptive Computation and Machine Learning. Foundations and learning algorithms. Cambridge, MA: MIT Press; 2017.
[7] Maathuis M, Drton M, Lauritzen S, Wainwright M, editors. Handbook of graphical models. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton, FL: CRC Press; 2019.
[8] Peters J, Bühlmann P. Identifiability of Gaussian structural equation models with equal error variances. Biometrika. 2014;101(1):219–28.
[9] Strieder D, Freidling T, Haffner S, Drton M. Confidence in causal discovery with linear causal models. In: Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. UAI'21. Vol. 161. PMLR; 2021. p. 1217–26.
[10] Chen W, Drton M, Wang YS. On causal discovery with an equal-variance assumption. Biometrika. 2019;106(4):973–80.
[11] Ghoshal A, Honorio J. Learning linear structural equation models in polynomial time and sample complexity. In: Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. Vol. 84. PMLR; 2018. p. 1466–75.
[12] Andrews DWK, Guggenberger P. Asymptotic size and a problem with subsampling and with the m out of n bootstrap. Econ Theory. 2010;26(2):426–68.
[13] Drton M, Williams B. Quantifying the failure of bootstrap likelihood ratio tests. Biometrika. 2011;98(4):919–34.
[14] Casella G, Berger RL. Statistical inference. The Wadsworth & Brooks/Cole Statistics/Probability Series. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1990.
[15] Drton M. Algebraic problems in structural equation modeling. In: The 50th anniversary of Gröbner bases. Vol. 77 of Adv. Stud. Pure Math. Math. Soc. Japan, Tokyo; 2018. p. 35–86.
[16] Silvapulle MJ, Sen PK. Constrained statistical inference. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons]; 2005.
[17] Wasserman L, Ramdas A, Balakrishnan S. Universal inference. Proc Natl Acad Sci USA. 2020;117(29):16880–90.
[18] Drton M. Likelihood ratio tests and singularities. Ann Statist. 2009;37(2):979–1012.
[19] Strieder D, Drton M. On the choice of the splitting ratio for the split likelihood ratio test. Electron J Stat. 2022;16(2):6631–50.
[20] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523–9.
[21] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res. 2006;7:2003–30.

Permission to include the article

DE GRUYTER

Last update: February 2021

Editorial Policies

OVERVIEW OF THE EDITORIAL PROCESS

For detailed submission guidelines please refer to the Instruction for Authors or contact the Editors of **Journal of Causal Inference**.

Submission

Each manuscript should be accompanied by a cover letter which should explicitly state that the authors have the authority to publish the work and that the manuscript (or one with substantially the same content, by any of the authors) has not been previously published in any language anywhere and that it is not under simultaneous consideration by another journal. All authors of the manuscript are responsible for its content; they must have agreed to its publication and have given the corresponding author the authority to act on their behalf in all matters pertaining to publication. The corresponding author is responsible for informing the co-authors of the manuscript status throughout the submission, review, and production process.

Authorship

Authorship should be limited to those who have made a significant contribution to the conception, design, execution, or interpretation of the reported study. All those who have made significant contributions should be listed as co-authors. Those who do not meet that criteria should be acknowledged (see Instructions for Authors for details). It is the sole responsibility of contributors to determine the authors of the manuscript submitted to the journal.

Authors must ensure that anyone named in the acknowledgments agrees to being so named. Editors of **Journal of Causal Inference** may require that the corresponding author obtain written permission to be acknowledged from all acknowledged individuals.

Addition or Removal of Authors

The authors' request for addition or removal of an author should be properly justified. Please note that a change in authorship (order of listing, addition or deletion of a name, or corresponding author designation) after submission of the manuscript will be implemented only after receipt of signed statements of agreement from all parties involved (all listed authors and the author to be removed or added).

Peer Review process

Each manuscript after uploading to Editorial Manager receives an individual identification code that is used in all correspondence regarding the publication process. However, a submission may be declined by the Editors without review, if the studies reported are not sufficiently novel or important to merit publication in the journal. Manuscripts deemed unsuitable (insufficient originality or of limited interest to the target audience) are returned to the author(s) without review.

Choice of reviewers

The Editor seeks advice from experts in the appropriate field. Research articles and communications are refereed by a minimum of two reviewers, review papers by at least three.

Suggestions from authors

Authors are requested to suggest persons competent to review their manuscript. However, please note that this will be treated only as a suggestion, and the final selection of reviewers is exclusively the Editor's decision. The authors' names are revealed to the referees, but not vice versa.

The reviewers make an objective, impartial evaluation of scientific merits of the manuscript. Reviewers operate under guidelines set forth in the Guidelines for reviewers and are asked to comment on the following aspects of submitted manuscripts:

- novelty and originality of the work;
- broad interest to the community of researchers;
- significance to the field, potential impact of the work, conceptual or methodological advances described;
- study design and clarity;
- substantial evidence supporting claims and conclusions;

Permission to include the article

DE GRUYTER

- rigorous methodology

If a manuscript is believed to not meet the standards of the journal or is otherwise lacking in scientific rigor or contains major deficiencies, the reviewers will attempt to provide constructive criticism to assist the authors in ultimately improving their work. If a manuscript is believed to be potentially acceptable for publication but needs to be improved, it is invited for reconsideration with the expectation that the authors will fully address the reviewer's suggestions.

Once all reviews have been received and considered by the Editor, a decision letter to the author is drafted. There are several types of decisions possible:

- Accept without revision
- Minor revision
- Major revision
- Reject

Revised manuscript submission

When revision of a manuscript is requested, authors should return the revised version of their manuscript as soon as possible. Prompt action may ensure fast publication, if a paper is finally accepted for publication. The authors need to return their revised manuscript within the time specified by the Editor in the decision letter (usually 28 days for the first revision and 14 days for subsequent revision). If these deadlines are not met, and no specific arrangements for completion have been made with the Editor, the manuscript will be treated as a new one and will receive a new identification code along with a new registration date.

The final decision is made by the main Editor of the journal.

Final proofreading

Authors will receive a pdf file with the edited version of their manuscript for final proofreading. This is the last opportunity to view an article before its publication on the Journal web site. No changes or modifications can be introduced once it is published. Thus, authors are requested to check their proof pages carefully against the manuscript within 3 working days and prepare a separate document containing all changes that should be introduced. Authors are sometimes asked to provide additional comments and explanations in response to remarks and queries from technical editors.

Immediate publication

Manuscripts ready for publication are promptly posted online. The manuscripts are considered to be ready for publication when the final proofreading has been performed by authors, and all concerns have been resolved. Authors should notice that no changes can be made to the articles after online publication.

Erratum

If any errors are detected in the published material, they should be reported to the Editor of the journal. The corresponding authors should send the appropriate corrected material to the Editor via email. This material will be considered for publication as soon as feasible.

Reprints

Because the journal is published in an Open Access model, and has no printed version, the authors receive no reprints.

Copyright

All authors retain copyright, unless – due to their local circumstances – their work is not copyrighted. The copyrights are governed by the Creative-Commons Attribution Only license (CC-BY) which is compliant with Plan-S. Scanned copy of license should be sent to the journal, as soon as possible.

SCIENTIFIC MISCONDUCT AND OTHER FRAUD

Scientific misconduct is defined by the Office of Research Integrity as "fabrication, falsification, plagiarism, or other practices that seriously deviate from those that are commonly accepted within the academic community for proposing, conducting, or reporting research". In cases where there is a suspicion or allegation of scientific misconduct or fraudulent research in manuscripts submitted or published, the Editors reserve the right to impose sanctions on the authors, such as:

Permission to include the article

DE GRUYTER

- an immediate rejection of the manuscript;
- banning author(s) from submitting manuscripts to the journal for a certain period of time;
- retracting the manuscript;
- alerting editors of other journals and publishers;
- bringing the concerns to the authors' sponsoring or funding institution or other appropriate authority for investigation

This journal publishes only original manuscripts that are not also published or going to be published elsewhere. Multiple submissions/publications, or redundant publications (re-packaging in different words of data already published by the same authors) will be rejected. If they are detected only after publication, the journal reserves the right to publish a Retraction Note. In each particular case Editors will follow COPE's Code of Conduct and implement its advice.

Plagiarism

As a member of Cross Ref, De Gruyter provides plagiarism detection software CrossCheck to all its journals. When plagiarism in the submitted manuscript is identified, Editors will follow COPE guidelines on plagiarism.

Retraction Policy

Serious errors in a published manuscript and infringements of professional ethical codes will result in an article being retracted. This will occur where the article is clearly defamatory, or infringes others' legal rights, or where the article is, or there is good reason to expect it will be, the subject of a court order, or where the article, if acted upon, might pose a serious health risk. In any of these cases all coauthors will be informed about a retraction. A Retraction Note detailing the reason for retraction will be linked to the original article.

CONFLICT OF INTEREST

In order to encourage transparency without impeding publication, all authors, referees and editors must declare any association that poses a conflict of interest in connection with the manuscript. There should be no contractual relations or proprietary considerations that would affect the publication of information contained in a submitted manuscript. A competing interest for a scholarly journal is anything that interferes with, or could reasonably be perceived as interfering with, the full and objective presentation, review, or publication of research findings, or of articles that comment on or review research findings. Potential conflicts of interest exist when an author, editor or reviewer has financial, personal or professional interests in a publication that might influence their scientific judgment.

Examples of such conflicts include, but are not limited to:

- Financial conflicts: stock ownership; patents; paid employment or consultancy; board membership; research grants; travel grants and honoraria for speaking or participation at meetings; gifts
- Personal conflicts: relationship with editors, editorial board members, or with possible reviewers who have had recent or ongoing collaborations with the authors, have commented on drafts of the manuscript, are in direct competition, have a history of dispute with the authors
- Professional conflicts: public associations with institutions or corporations whose products or services are related to the subject matter of the article; membership of a government advisory council/committee; relationship with organizations and funding bodies

Authors should declare whether they have any conflicts of interests that could have influenced the reporting of the experimental data or conclusions in their paper. Such a statement should list all potential interests or, if appropriate, should clearly state that there are none. The editors may decide not to publish papers when we believe the competing interests are such that they may have compromised the work or the analyses or interpretations presented. Upon submission of a manuscript, authors may suggest to exclude any specific editors or reviewers from the peer review of their article. It is the responsibility of authors to disclose in the Acknowledgments section any funding sources for the project or other relationships that are relevant. Authors are suggested to fill in the Conflicts of Interest Form and send the electronic version to the Editor.

Editors should consider whether any of the above competing interests are relevant to them and the manuscript under consideration. Editor who believes that the conflict will preclude an impaired judgment should disclose to the main Editor of the journal the nature of the conflict and decline to handle the paper.

Permission to include the article

DE GRUYTER

Reviewers should consider whether any of the above applies to them and declare any such competing interests. If they feel they cannot review a paper because of any competing interest, they should tell us. They should also declare any association with the authors of a paper.

DISTRIBUTION OF MATERIALS AND DATA

The publication of an article in the journal is subject to the understanding that authors will make all data and associated protocols available to readers on request. The Methods section should include details of how materials and information may be obtained. In cases of dispute, authors may be required to make any primary data available to the main Editor.

In the case of new software, source code should ideally be made available, for example as supporting information with the rest of the paper, or by deposition at a publicly accessible resource such as sourceforge.net. For a new algorithm, a detailed description should be published in the paper. In cases where the software/algorithm is not central to the paper, we nevertheless encourage authors to make all relevant materials freely available. Software can be provided under license where necessary, but any restrictions on the availability or on the use of materials might be judged to diminish the significance of a paper, and therefore influence the decision about whether a paper should be published subject to those conditions.

POLICY ON COMMENTING ARTICLES

Readers are free to submit comments, questions or criticism about all articles published in the journal.

De Gruyter reserves the right not to post comments deemed to be discourteous, inaccurate or libelous and the right to remove comments already posted.

Comments may also be declined if they:

- are irrelevant to the article
- are lacking cogency
- are incomprehensible
- apear to be advertising

Authors of all comments are requested to reveal all competing interests they might have with respect to the article. For more details see Conflict of Interest.

APPEALS AND COMPLAINTS

Appeals

Authors who want may appeal on the rejection of their manuscript should contact the Editors of the journal. Appeals should refer to scientific content of the manuscript and its suitability for publication. The decision made by the Editor is final.

Complaints

Authors who want to make complaints should, in first instance, contact the main Editor of the journal. In case, the Editor is not able to resolve the complaint, the Authors should contact us directly

(Editorial Office; jci_editorial@degruyter.com).

Permission to include the article

Permission to include the article

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- Learn more about our work
- **Learn more about CC Licensing**
- Support our work
- Use the license for your own material.
- Licenses List
- Public Domain List

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- More info

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indication of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- Marking guide
- More info

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WIPO Copyright Treaty.

- More info

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- More info

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- More info

A.2 On the Choice of the Splitting Ratio for the Split Likelihood Ratio Test

Electronic Journal of Statistics
Vol. 16 (2022) 6631–6650
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2099

On the choice of the splitting ratio for the split likelihood ratio test*

David Strieder and Mathias Drton

Technical University of Munich; TUM School of Computation, Information and Technology, Munich Center for Machine Learning (MCML), Munich Data Science Institute (MDSI)
e-mail: david.strieder@tum.de; mathias.drton@tum.de

Abstract: The recently introduced framework of universal inference provides a new approach to constructing hypothesis tests and confidence regions that are valid in finite samples and do not rely on any specific regularity assumptions on the underlying statistical model. At the core of the methodology is a split likelihood ratio statistic, which is formed under data splitting and compared to a cleverly selected universal critical value. As this critical value can be very conservative, it is interesting to mitigate the potential loss of power by careful choice of the ratio according to which data are split. Motivated by this problem, we study the split likelihood ratio test under local alternatives and introduce the resulting class of non-central split chi-square distributions. We investigate the properties of this new class of distributions and use it to numerically examine and propose an optimal choice of the data splitting ratio for tests of composite hypotheses of different dimensions.

MSC2020 subject classifications: Primary 62E20, 62F03.
Keywords and phrases: Chi-square distribution, likelihood ratio test, local alternatives, universal inference.

Received March 2022.

Contents

1 Introduction 6632
2 Background on the split likelihood ratio test 6633
3 Asymptotic theory for the SLRT 6634
3.1 Asymptotic distribution 6635
3.2 Optimal splitting ratio 6639
4 Simulations 6643
4.1 Power of the SLRT in regular setting 6643
4.2 Influence of the splitting ratio 6644
4.3 Power of the SLRT in irregular setting 6646
5 Extension to the cross-fit SLRT 6647
6 Conclusion 6649
References 6649

*This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 83818). Further, this work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content.

1. Introduction

Likelihood ratio tests provide powerful solutions to a broad range of hypothesis testing problems. However, their implementation generally relies on asymptotic approximations whose validity requires the underlying statistical models to satisfy a number of regularity conditions. When these conditions are not met, the needed distribution-theoretic insights may be difficult to obtain; see, e.g., [1, 3]. Recent work of Wasserman, Ramdas and Balakrishnan [8] addresses this challenge by providing a *split likelihood ratio test* that is universally applicable to problems with i.i.d. samples. In this *universal inference* methodology, the data are split into two parts: one part is used to form a maximum likelihood estimate of a distribution under the full model, and the remaining data are used to compare the likelihood under the estimate versus the null hypothesis. Crucially, the independence of the split data allows one to apply a universal critical value, which merely depends on the chosen significance level and is guaranteed to be conservative even for finite samples. The resulting methodology makes it possible to conduct rather simple analyses of complicated composite hypotheses. For example, Strieder et al. [5] recently used the approach to construct hypothesis tests for causal effects in a setting with unknown causal structure.

The initial work in [8] and the follow-up paper by Dunn et al. [2] investigate the performance of the universal inference framework in the Gaussian case and under consideration of point hypotheses/construction of confidence regions. Unsurprisingly, the universal framework is rather conservative. To cite the authors: "our methods may not be optimal, though we do not yet fully understand how close to optimal they are beyond special cases (uniform, Gaussian)." During the review period, the independent work of Tse and Davison [6] was announced to appear, which covers different aspects of related problems.

The goal of the present paper is to expand our insights about the behavior of the split likelihood ratio test, as introduced in more detail in Section 2. In particular, we seek to shed light on the impact of the dimensionality of the tested null and alternative hypotheses. To this end, we study the case of smooth hypotheses in regular parametric models that are differentiable in quadratic mean. Under similar conditions, Wasserman, Ramdas and Balakrishnan [8] studied in their initial work the diameter of confidence sets corresponding to the inverted split likelihood ratio tests. We extend our insights about the limit behaviour of the split likelihood ratio test by deriving precise limit theory and calculating the large-sample asymptotic distribution allowing for local alternatives. This distribution belongs to a "split-version" of noncentral chi-square distributions, for which moments may be derived explicitly. We then use this new class of *noncentral split chi-square distributions* to propose a new routine for calculating the optimal splitting ratio for the split likelihood ratio test based on the dimensionality of the tested null and alternative hypotheses (Section 3). Furthermore, we use this new class of distributions to conduct numerical experiments that analyze the power and the optimal choice of the splitting ratio for the split likelihood ratio (Section 4). The simulations suggest, in particular, that while in lower dimensional settings an even split performs well, in higher dimensions a lower

splitting ratio is advantageous and our proposed new splitting ratio significantly improves power. In an experiment with factor analysis models, we demonstrate that our proposal may also lead to a power gain in irregular settings. This and other findings are further discussed in the concluding Section 6.

Notation In the remainder, the symbols $\xrightarrow{P}$, $\xrightarrow{a.s.}$ and $\xrightarrow{\mathcal{D}}$ stand for convergence in probability, almost sure convergence and convergence in distribution, respectively. The stochastic Landau symbol $o_P(1)$ indicates convergence to zero in probability. If not stated otherwise, the limits refer to $n \to \infty$. With Id we denote the $d \times d$ identity matrix, and $\mathcal{N}_d(0, \mathrm{Id})$ is the standard normal distribution in $\mathbb{R}^d$. For a given vector $X \in \mathbb{R}^d$ we denote the vector of its first p components by $X_{[p]}$.

2. Background on the split likelihood ratio test

Let $\{P_\theta : \theta \in \Theta\}$ be a given (parametric) statistical model, with parameter space $\Theta \subset \mathbb{R}^d$. The distributions P_θ are assumed to be dominated by a measure μ, and we write p_θ for the μ-density of P_θ. Given an i.i.d. sample $X_1, \ldots, X_n$ from an unknown distribution P_θ in the model, we are interested in testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta \setminus \Theta_0 \tag{1}$$

for a subset $\Theta_0 \subset \Theta$. Universal inference solves this problem by appealing to a likelihood ratio, however, one that is built using data splitting.

To split the data, one chooses a fraction $m_0 \in (0, 1)$ and partitions the n data points into two disjoint subsets $D_0 = \{X_{1,0}, \ldots, X_{\lfloor m_0 n \rfloor, 0}\}$ and $D_1 = \{X_{1,1}, \ldots, X_{\lceil m_1 n \rceil, 1}\}$, where $m_1 \equiv 1 - m_0$. In order to lighten notation in subsequent derivations, we simply write $m_0 n$ for $\lfloor m_0 n \rfloor$ and $m_1 n$ for $\lceil m_1 n \rceil$. Let

$$\ell_k(\theta) = \sum_{i=1}^{m_k n} \log p_\theta(X_{i,k}), \quad k = 0, 1,$$

be the log-likelihood functions based on D_0 and D_1, respectively. Let $\widehat{\theta}_{n,0} := \operatorname{argmax}_{\theta \in \Theta_0} \ell_0(\theta)$ be the maximum likelihood estimator (MLE) of θ under H_0 and based on D_0. Furthermore, let $\widehat{\theta}_{n,1} := \operatorname{argmax}_{\theta \in \Theta} \ell_1(\theta)$ be the MLE of θ under the full model and based on D_1. Now the split likelihood ratio statistic is defined as

$$\Lambda_n := 2 \left(\ell_0(\widehat{\theta}_{n,1}) - \ell_0(\widehat{\theta}_{n,0}) \right). \tag{2}$$

As shown in [8], an application of Markov's inequality yields for any $\alpha \in (0, 1)$ that under $H_0 : \theta \in \Theta_0$ we have

$$\begin{aligned} P_\theta\big(\Lambda_n > -2\log(\alpha)\big) &\le \alpha \mathbb{E}_\theta \left[\frac{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,1}}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,0}}(X_{i,0})} \right] \\ &\le \alpha \mathbb{E}_\theta \left[\mathbb{E}_\theta \left[\frac{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,1}}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_\theta(X_{i,0})} \middle| D_1 \right] \right] \le \alpha. \end{aligned}$$

Here, we used the fact that $\widehat{\theta}_{n,1}$ is fixed when we condition on D_1, and for any fixed $\theta^* \in \Theta$ it holds that

$$\mathbb{E}_\theta\left[\frac{\prod_{i=1}^{m_0 n} p_{\theta^*}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_\theta(X_{i,0})}\right] \leq \int \prod_{i=1}^{m_0 n} p_{\theta^*}(X_{i,0}) = 1.$$

Therefore, the decision rule

$$\text{reject } H_0 \quad \text{if} \quad \Lambda_n > -2\log(\alpha) \tag{3}$$

constitutes a valid level α test. Notably, this *split likelihood ratio test* (SLRT) holds level α in finite samples and without any regularity conditions.

Remark 2.1. The MLE $\widehat{\theta}_{n,1}$ could be replaced by any other estimator and the test would continue to be valid. While this may be of interest for computational reasons, we focus in the following on the asymptotically efficient maximum likelihood estimator. However, the analysis can be extended in a similar fashion for any asymptotic linear estimator, see Remark 3.2.

In the following, we derive the asymptotic distribution of the split likelihood ratio Λ_n and use it to study the power of the SLRT and the impact of the splitting ratio m_0. Our calculation of the limiting distribution of Λ_n is couched in the classical framework of local alternatives in models that are differentiable in quadratic mean.

3. Asymptotic theory for the SLRT

Let θ_0 be a point in the interior of Θ. Assume that $\theta_0 \in \Theta_0$ and define the sequence of parameters $\theta_n = \theta_0 + h/\sqrt{n}$ for a choice of $h \in \mathbb{R}^d$. Suppose then that for each (large) n we are given an i.i.d. sample of size n from the local alternative P_{θ_n}. Suppose further that the considered model possesses the usual smoothness properties that lead to chi-square limits for the ordinary likelihood ratio, see, e.g., [7]. Specifically, we assume that:

(A1) The model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ_0, with derivative (i.e., score function) $\dot{\ell}_{\theta_0}$. Its Fisher information $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T] = I(\theta_0)$ is nonsingular, and there exists a measurable function $\dot{\ell}$ with $\mathbb{E}_{\theta_0}[\dot{\ell}^2] < \infty$ such that

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|$$

for all θ_1, θ_2 in a neighborhood of θ_0.

(A2) The maximum likelihood estimators $\widehat{\theta}_{n,0}$ and $\widehat{\theta}_{n,1}$ are consistent estimators for θ_0 under P_{θ_0}.

(A3) The local parameter spaces $H_n := \sqrt{n}(\Theta_0 - \theta_0)$ converge to a set H_0.

Assumption (A3) uses the notion of convergence of sets in the sense of [7], that is, the local parameter spaces H_n converge to the limit hypothesis H_0, if the set H_0

a) is the set of all limits $\lim h_n$ of converging sequences with $h_n \in H_n$ and
b) contains all limits $\lim_{i\to\infty} h_{n_i}$ of converging sequences with $h_{n_i} \in H_{n_i}$.

Further, we note that the set convergence is guaranteed to hold when Θ_0 is defined by polynomial equations and inequalities, in which case the limit H_0 is the tangent cone of Θ_0 at θ_0 [1].

3.1. Asymptotic distribution

Theorem 3.1 (Asymptotic distribution of the split likelihood ratio statistic)**.** *Suppose the considered statistical model* $\{P_\theta : \theta \in \Theta\}$ *satisfies assumptions (A1)-(A3). Then under* P_{θ_n} *with* $\theta_n = \theta_0 + h/\sqrt{n}$*, the split likelihood ratio statistic from* (2) *satisfies*

$$\Lambda_n \xrightarrow{\mathcal{D}} \|X + \sqrt{m_0} I(\theta_0)^{1/2} h - I(\theta_0)^{1/2} H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}} Y\|^2,$$

where $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ *independent and* $\|x - H_0\| = \inf_{h\in H_0} \|x - h\|$.

Proof. The proof is based on classical local asymptotic normality results. As shown in Theorem 7.12 of [7], our assumption (A1) implies the existence of the Fisher information and the uniform approximation

$$\sup_{\|h\|\le M_n} \left| \log \prod_{i=1}^{n} \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}}(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \dot{\ell}_{\theta_0}(X_i) + \frac{1}{2} h^T I(\theta_0) h \right| = o_{P_{\theta_0}}(1) \tag{4}$$

for M_n a slowly diverging sequence in $\mathbb{R}$. Via the results collected in [7], assumption (A2) implies that consistent MLEs are $\sqrt{n}$-consistent, which entails that both our split sample MLEs $\hat{\theta}_{n,0}$ and $\hat{\theta}_{n,1}$ are $\sqrt{n}$-consistent under P_{θ_0}.

Define $\widehat{\psi}_{n,1} := \sqrt{m_1 n}(\widehat{\theta}_{n,1} - \theta_0)$, $G_{n,0} := \frac{1}{\sqrt{m_0 n}} \sum_{i=1}^{m_0 n} \dot{\ell}_{\theta_0}(X_{i,0})$ and $G_{n,1} := \frac{1}{\sqrt{m_1 n}} \sum_{i=1}^{m_1 n} \dot{\ell}_{\theta_0}(X_{i,1})$. Let $B(M_n) = \{h \in \mathbb{R}^d : \|h\| \le M_n\}$ be the ball of radius M_n. Similarly to the proof of Theorem 16.7 in [7] but accounting for the split sample, we obtain from (4) and the $\sqrt{n}$-consistency of $\hat{\theta}_{n,0}$ and $\hat{\theta}_{n,1}$ that

$$\begin{aligned}
\Lambda_n &= 2\left(\ell_0(\widehat{\theta}_{n,1}) - \ell_0(\widehat{\theta}_{n,0})\right) \\
&= 2\left(\log \prod_{i=1}^{m_0 n} \frac{p_{\theta_0 + \widehat{\psi}_{n,1}/\sqrt{m_1 n}}}{p_{\theta_0}}(X_{i,0}) - \sup_{h\in H_{m_0 n}} \log \prod_{i=1}^{m_0 n} \frac{p_{\theta_0 + h/\sqrt{m_0 n}}}{p_{\theta_0}}(X_{i,0})\right) \\
&= 2\Bigg(\sqrt{\frac{m_0}{m_1}} \widehat{\psi}_{n,1}^T G_{n,0} - \frac{m_0}{2m_1} \widehat{\psi}_{n,1}^T I(\theta_0) \widehat{\psi}_{n,1} \\
&\qquad - \sup_{h\in H_{m_0 n}\cap B(M_n)} \left(h^T G_{n,0} - \frac{1}{2} h^T I(\theta_0) h\right)\Bigg) + o_{P_{\theta_0}}(1) \\
&= \|I(\theta_0)^{-1/2} G_{n,0} - I(\theta_0)^{1/2}[H_{m_0 n} \cap B(M_n)]\|^2 \\
&\qquad - \|I(\theta_0)^{-1/2} G_{n,0} - \sqrt{\frac{m_0}{m_1}} I(\theta_0)^{1/2} \widehat{\psi}_{n,1}\|^2 + o_{P_{\theta_0}}(1).
\end{aligned}$$

By Theorem 5.39 in [7], the MLE $\hat{\theta}_{n,1}$ is asymptotically linear with $\widehat{\psi}_{n,1} = I(\theta_0)^{-1}G_{n,1} + o_{P_{\theta_0}}(1)$. Hence,

$$\begin{aligned}\Lambda_n = &\|I(\theta_0)^{-1/2}G_{n,0} - I(\theta_0)^{1/2}[H_{m_0 n} \cap B(M_n)]\|^2 \\ &- \|I(\theta_0)^{-1/2}G_{n,0} - \sqrt{\frac{m_0}{m_1}}I(\theta_0)^{-1/2}G_{n,1}\|^2 + o_{P_{\theta_0}}(1).\end{aligned} \tag{5}$$

Now we use Le Cam's Lemmas to show contiguity of P_{θ_0} and P_{θ_n}. Applying (4),

$$\begin{aligned}\log \frac{\mathrm{d}P_{\theta_n}^{\otimes n}}{\mathrm{d}P_{\theta_0}^{\otimes n}} &= \log\left(\prod_{i=1}^{m_0 n}\frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_{i,0})\prod_{i=1}^{m_1 n}\frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_{i,1})\right) \\ &= \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{m_0 n}h^T\dot{\ell}_{\theta_0}(X_{i,0}) + \sum_{i=1}^{m_1 n}h^T\dot{\ell}_{\theta_0}(X_{i,1})\right) - \frac{1}{2}h^T I(\theta_0)h + o_{P_{\theta_0}}(1).\end{aligned}$$

The central limit theorem yields that under P_{θ_0},

$$\begin{pmatrix} G_{n,0} \\ G_{n,1} \\ \log\frac{\mathrm{d}P_{\theta_n}^{\otimes n}}{\mathrm{d}P_{\theta_0}^{\otimes n}}\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_{2d+1}(\mu, \Sigma),$$

where

$$\mu := \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2}h^T I(\theta_0)h\end{pmatrix}, \qquad \Sigma := \begin{bmatrix} I(\theta_0) & 0 & \sqrt{m_0}I(\theta_0)h \\ 0 & I(\theta_0) & \sqrt{m_1}I(\theta_0)h \\ \sqrt{m_0}I(\theta_0)h & \sqrt{m_1}I(\theta_0)h & h^T I(\theta_0)h\end{bmatrix}.$$

By Le Cam's first lemma, the probability measures P_{θ_0} and P_{θ_n} are thus mutually contiguous and therefore $o_{P_{\theta_0}}(1)$ and $o_{P_{\theta_n}}(1)$ interchangeable. By Le Cam's third lemma, it follows that under P_{θ_n} we have

$$\begin{pmatrix} G_{n,0} \\ G_{n,1}\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_{2d}\left(\begin{pmatrix}\sqrt{m_0}I(\theta_0)h \\ \sqrt{m_1}I(\theta_0)h\end{pmatrix}, \begin{bmatrix} I(\theta_0) & 0 \\ 0 & I(\theta_0)\end{bmatrix}\right). \tag{6}$$

We may now use this joint convergence in (5) and arrive at our claim by observing that for any converging sequence of random vectors $X_n \xrightarrow{\mathcal{D}} X$ and any sequence of converging sets $H_n \to H$ it holds that

$$\|X_n - H_n\| \xrightarrow{\mathcal{D}} \|X - H\|;$$

see Lemma 7.13 in [7]. Indeed, our assumption (A3) implies the convergence $H_{m_0 n} \cap B(M_n) \to H_0$ and our claim follows. □

Remark 3.2. Throughout this work we use the MLE $\widehat{\theta}_{n,1}$ to solve the estimation task on data set D_1. Suppose we employ instead a suboptimal, asymptotically linear estimator $\tilde{\theta}_{n,1}$, that is,

$$\sqrt{m_1 n}(\tilde{\theta}_{n,1} - \theta_0) = \tfrac{1}{\sqrt{m_1 n}}I(\theta_0)^{-1}\sum_{i=1}^{m_1 n}\tilde{g}(X_{i,1}) + o_{P_{\theta_0}}(1),$$

with $\mathrm{Var}[\tilde{g}(X_{i,1})] \succeq I(\theta_0)$, where $\succeq$ denotes the Loewner order. Tracing the proof of Theorem 3.1, one obtains that

$$\tilde{\Lambda}_n \xrightarrow{\mathcal{D}} \|X + \sqrt{m_0} I(\theta_0)^{1/2} h - I(\theta_0)^{1/2} H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}\tilde{Y}\|^2, \tag{7}$$

where $X, \tilde{Y}$ are independent, $X \sim \mathcal{N}_d(0, \mathrm{Id})$ and $\tilde{Y} \sim \mathcal{N}_d(\mu_{\tilde{Y}}, V_{\tilde{Y}})$ with $\mu_{\tilde{Y}} \neq 0$ and $V_{\tilde{Y}} \succeq \mathrm{Id}$. In fact, we may assume that $V_{\tilde{Y}}$ is diagonal with diagonal entries $(V_{\tilde{Y}})_{jj} \geq 1$. (Otherwise, apply an orthogonal transformation to $X - \sqrt{\frac{m_0}{m_1}}\tilde{Y}$ to diagonalize $V_{\tilde{Y}}$.) The second part of the representation of the asymptotic distribution is thus

$$\|X - \sqrt{\tfrac{m_0}{m_1}}\tilde{Y}\|^2 \overset{\mathcal{D}}{=} \sum_{i=1}^{d} \big(1 + \tfrac{m_0}{m_1}(V_{\tilde{Y}})_{ii}\big) Z_i^2 \geq \big(1 + \tfrac{m_0}{m_1}\big) \sum_{i=1}^{d} Z_i^2,$$

where $Z_i \sim \mathcal{N}(-\sqrt{\frac{m_0}{m_1}}(\mu_{\tilde{Y}})_i, 1)$ are independent. Now, $\sum_{i=1}^{d} Z_i^2 \sim \chi_d^2(\lambda)$ with noncentrality parameter $\lambda > 0$. Since $\chi_d^2(\lambda)$ is stochastically larger than a (central) χ_d^2-distribution, we obtain from (7) that the limiting distribution when using a suboptimal estimator is stochastically smaller than the limit distribution of Theorem 3.1 with the MLE. Using a suboptimal estimator on D_1 thus leads to a decrease in power for the SLRT.

In the sequel, we investigate properties of the limiting distribution of the split likelihood ratio statistic in the smooth case, where the original null hypothesis is a k-dimensional smooth manifold and the limiting set H_0 is thus a k-dimensional tangent space. We start by introducing the arising new class of distributions, *noncentral split chi-square distributions*, that depends on four parameters, the dimension of the parameter space, the dimension of the null hypothesis, the splitting ratio, and a noncentrality parameter.

Definition 3.3 (Noncentral split chi-square distribution)**.** Let $d \in \mathbb{N}$, $p \in \{0, \ldots, d\}$, $\delta \geq 0$, and $m_0 \in (0, 1)$. The d-dimensional *noncentral split chi-square distribution* with p degrees of freedom, noncentrality parameter δ and splitting ratio m_0, denoted $\mathrm{split}_{m_0}\text{-}\chi_{p,d}^2(\delta)$, is the distribution of

$$\|X_{[p]} + \sqrt{m_0} h\|^2 - \|X - \sqrt{\tfrac{m_0}{1-m_0}} Y\|^2 \sim \mathrm{split}_{m_0}\text{-}\chi_{p,d}^2(\delta),$$

where $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ independent and $h \in \mathbb{R}^p$ such that $h^T h = \delta$.

We emphasize that the noncentral split chi-square distribution is well-defined in that it depends on the vector h only through its norm δ. This follows from the invariance of the standard normal distribution of X and Y under orthogonal rotations analogously to the classical noncentral chi-square distribution.

While the classical noncentral chi-square distribution is the distribution of the squared distance from a standard normal vector to some fixed point in the space, the noncentral split chi-square distribution is the distribution of the difference of two squared distances. The first part is the squared distance from

a standard normal vector to a fixed point in the space. However, a second part arises from splitting the data into two subsets, namely, the squared distance of two independent standard normal vectors scaled according to the splitting ratio. Notice that the two parts are not independent.

In the following we calculate the first moments of this new class of distributions. In Section 3.2 we use the calculated moments to approximate the noncentral split chi-square distribution and thus the asymptotic behavior of the SLRT.

Corollary 3.4 (Moments of the noncentral split chi-square distribution)**.** *Let* $Z \sim \text{split}_{m_0}\text{-}\chi^2_{p,d}(\delta)$. *Then*

1. $\mathbb{E}[Z] = p - d - d\frac{m_0}{1-m_0} + m_0\delta$,
2. $\text{Var}[Z] = 2(d-p) + 4d\frac{m_0}{1-m_0} + 2d\frac{m_0^2}{(1-m_0)^2} + 4m_0\delta$.

Proof. Let

$$\epsilon \sim \mathcal{N}_{2d}\left(0, \begin{bmatrix} m_0^{-1}\text{Id} & 0 \\ 0 & m_1^{-1}\text{Id} \end{bmatrix}\right),$$

and define

$$\mu := \begin{pmatrix} h \\ 0 \\ h \\ 0 \end{pmatrix}, \quad A := \begin{bmatrix} 0 & 0 & \text{Id}_\text{p} & 0 \\ 0 & -\text{Id}_\text{k} & 0 & \text{Id}_\text{k} \\ \text{Id}_\text{p} & 0 & -\text{Id}_\text{p} & 0 \\ 0 & \text{Id}_\text{k} & 0 & -\text{Id}_\text{k} \end{bmatrix},$$

with $h \in \mathbb{R}^p$ such that $h^T h = \delta$. Then the quadratic form $m_0(\epsilon+\mu)^T A(\epsilon+\mu)$ follows a $\text{split}_{m_0}\text{-}\chi^2_{p,d}(\delta)$ distribution and we can use properties of quadratic forms to calculate moments of the noncentral split chi-square distribution.

Using $\mathbb{E}[(\epsilon+\mu)^T A(\epsilon+\mu)] = \text{tr}[A\Sigma] + \mu^T A\mu$, a short calculation yields the claim for the expectation and the claimed variance follows via $\text{Var}[(\epsilon+\mu)^T A(\epsilon+\mu)] = 2\,\text{tr}[A\Sigma A\Sigma] + 4\mu^T A\Sigma A\mu$. □

Remark 3.5. Higher moments can be calculated via the cumulants $\kappa_n(\epsilon^T A\epsilon) = 2^{n-1}(n-1)!\,\text{tr}[A^n]$ with the following formulas for moments of quadratic forms:

1. $\mathbb{E}[(\epsilon^T A\epsilon)^1] = \kappa_1$.
2. $\mathbb{E}[(\epsilon^T A\epsilon)^2] = \kappa_1^2 + \kappa_2$.
3. $\mathbb{E}[(\epsilon^T A\epsilon)^3] = \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3$.
4. $\mathbb{E}[(\epsilon^T A\epsilon)^4] = \kappa_1^4 + 6\kappa_1^2\kappa_2 + 3\kappa_2^2 + 4\kappa_1\kappa_3 + \kappa_4$.

Formulas for moments up to order ten can be found in [4].

Due to the rotational invariance of the standard normal distribution, we may study the limit of the SLRT in the smooth case, where the limiting hypothesis is a k-dimensional tangent space, by simply assuming that $I(\theta_0)^{1/2}H_0$ is a coordinate subspace, i.e., $I(\theta_0)^{1/2}H_0 = \{0\}^p \times \mathbb{R}^k$ with $d = p + k$.

Corollary 3.6. *If the rotated limiting hypothesis* $I(\theta_0)^{1/2}H_0 = \{0\}^p \times \mathbb{R}^k$ *is a coordinate subspace, then the asymptotic distribution from Theorem 3.1 follows a*

d-dimensional noncentral split chi-square distribution with p degrees of freedom, noncentrality parameter $\tilde{h}^T\tilde{h}$ *and splitting ratio* m_0*. That is*

$$\Lambda_\infty \overset{\mathcal{D}}{=} \|X_{[p]} + \sqrt{m_0}\tilde{h}_{[p]}\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}Y\|^2 \;\sim\; \text{split}_{m_0}\text{-}\chi^2_{p,d}(\tilde{h}^T_{[p]}\tilde{h}_{[p]}),$$

with $X, Y \sim \mathcal{N}_d(0, \text{Id})$ *independent and* $\tilde{h} = [I(\theta_0)^{1/2}h]_{[p]}$.

Proof. We look at the first part of the limiting distribution from Theorem 3.1. With $X \sim \mathcal{N}_d(\sqrt{m_0}I(\theta_0)^{1/2}h, \text{Id})$ we have

$$\|X - I(\theta_0)^{1/2}H_0\|^2 = \inf_{\theta\in\mathbb{R}^k}\left(X - \begin{pmatrix}0\\ \theta\end{pmatrix}\right)^T\left(X - \begin{pmatrix}0\\ \theta\end{pmatrix}\right) = X^T_{[p]}X_{[p]},$$

and the claim follows immediately. □

Remark 3.7. Under the null hypothesis, the limiting distribution of the split likelihood ratio test statistic reduces to the following difference of dependent (scaled) chi-square distributions

$$\Lambda_n \;\overset{\mathcal{D}}{\to}\; \|X_{[p]}\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}Y\|^2, \tag{8}$$

with $X, Y \sim \mathcal{N}_d(0, \text{Id})$ independent.

The limiting null distribution in (8) clearly shows the asymptotic difference between the LRT and the SLRT. For the SLRT, a new second term arises in the limit that behaves like a scaled chi-square distributed random variable where the scaling factor depends only on the chosen splitting ratio. Furthermore, looking at Corollary 3.4, the limiting distribution has a negative expectation under the null hypothesis.

The SLRT uses the conservative critical value $-2\log(\alpha)$ that is universally valid but whose adoption may come with a loss of power. In Figures 1 and 2 we illustrate the source of this loss of power in different settings by comparing the universal threshold (`SLRT`) of the SLRT with (simulated) quantiles of a split chi-square distribution (`Asym`), the limiting distribution under the null hypothesis (8). The difference between the universal threshold and the quantile of the limiting distribution is smaller for lower significance level α and thus, the power loss, which stems from employing an universal threshold, is less noteworthy for low significance levels. Furthermore, the universal threshold is asymptotically more accurate for smaller splitting ratios. Moreover, we observe that using the universal threshold is asymptotically less precise for higher dimensions of the parameter space and for higher dimensional hypotheses. In Section 4.1 we further analyze the power loss from using the universal threshold in simulations.

3.2. Optimal splitting ratio

The main advantage of the SLRT over classical likelihood methods is its flexibility for settings where asymptotic distributions are difficult to obtain. This

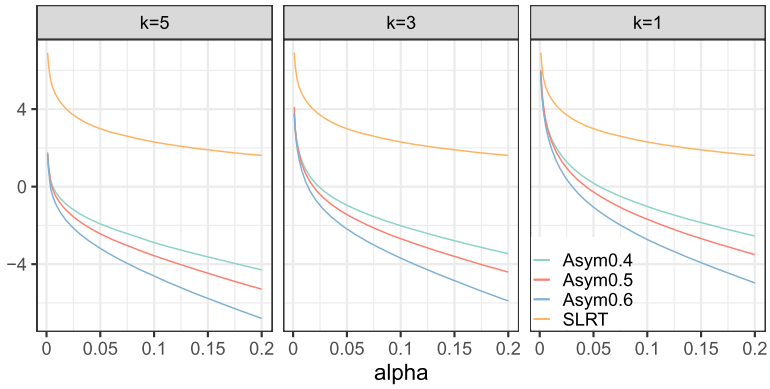


FIG 1. *Quantile of* split_{m_0}-$\chi^2_{p,6}$ *compared to the universal threshold.*

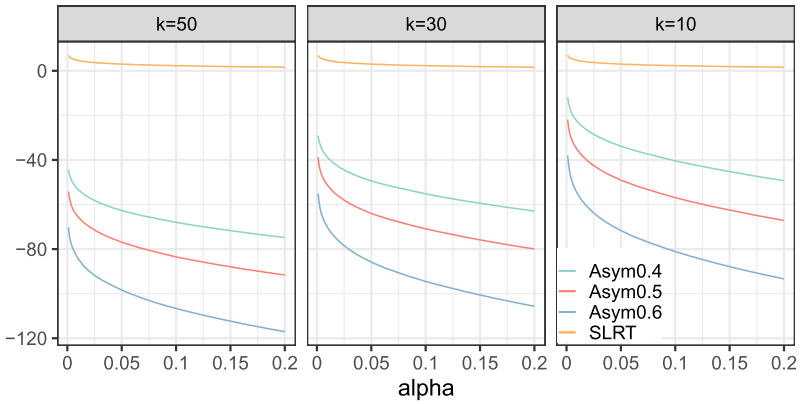


FIG 2. *Quantile of* split_{m_0}-$\chi^2_{p,60}$ *compared to the universal threshold.*

flexibility that stems from using only the general Markov inequality to control the type I error comes at the price of a potential loss of power. In the smooth setting of Theorem 3.1, we could improve the asymptotic power of the SLRT by using quantiles from the calculated asymptotic distribution, but such an asymptotic SLRT is not of practical relevance as the testing problem could then be better solved using the standard LRT, see Section 4.1.

Instead, our focus will remain on the SLRT with its conservative critical value $-2\log(\alpha)$, and our goal is to provide a new method for choosing the splitting ratio m_0 that helps retain power. The idea behind our proposed method is simple. Having access to the asymptotic distribution of the split likelihood ratio, the noncentral split chi-square distribution, we choose the splitting ratio that achieves the highest (asymptotic) power. Given both the dimensions of the null and alternative hypotheses and a significance level α, we minimize the cumulative distribution function of the noncentral split chi-square distribution at $-2\log(\alpha)$ with respect to the splitting ratio. To achieve a meaningful and comparable power, we propose to scale the unknown noncentrality parameter such that the best-performing method achieves a power of 0.8. We note that this

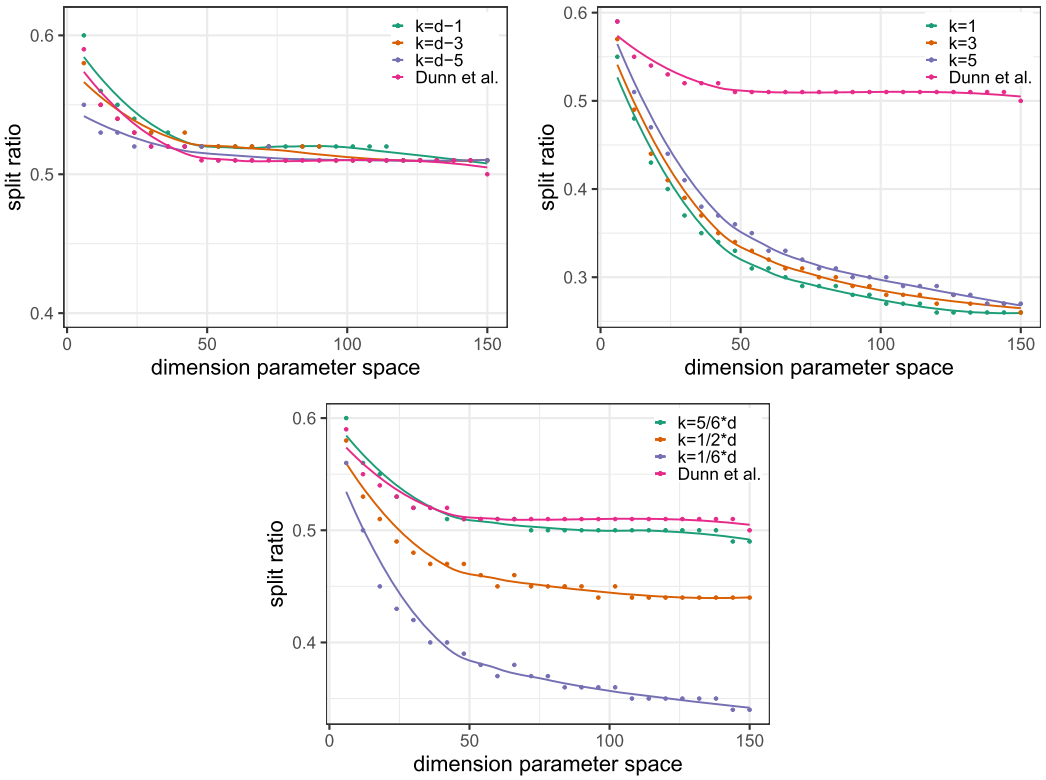


FIG 3. *Optimal splitting ratio against dimension of parameter space.*

tuning parameter of our method can easily be adapted if practitioners prefer to choose the best-performing method in different power levels. However, in our experience the effect of this tuning parameter on the performance is negligible across reasonable power levels.

As we are still lacking dedicated numerical routines to evaluate the cumulative distribution function of the noncentral split chi-square distribution, we use Monte Carlo approximations via repeated sampling from a noncentral split chi-square distribution. Using these approximations, we then choose the best-performing splitting ratio over a fine grid.

Figure 3 displays our results for the optimal choice of the splitting ratio against the dimension of the parameter space for different regimes of the null hypothesis space k. We observe that the underlying dimension of the null hypothesis k is crucial for the optimal choice of the splitting ratio. For comparison Figure 3 additionally shows the difference to the existing proposal for the choice of the splitting ratio by Dunn et al. [2]. To obtain a high power they propose to use the split ratio

$$m_0 = 1 - \frac{\sqrt{4d^2 + 8d\log(1/\alpha)} - 2d}{4\log(1/\alpha)}, \tag{9}$$

which minimizes the squared radius of the universal inference confidence set for the mean vector of a Gaussian distribution. In contrast to what our results

Algorithm 1 Optimal splitting ratio

Input: d, p, α
initialize δ small
power $\leftarrow 0.5$
while power< 0.8 **do**
  $\exp(m_0) \leftarrow p - d - d\frac{m_0}{1-m_0} + m_0\delta$ ▷ Define functions in split m_0
  $\mathrm{var}(m_0) \leftarrow 2(d-p) + 4d\frac{m_0}{1-m_0} + 2d\frac{m_0^2}{(1-m_0)^2} + 4m_0\delta$
  $\mathrm{target}(m_0) \leftarrow \texttt{pnorm}(-2\log(\alpha), \exp(m_0), \sqrt{\mathrm{var}(m_0)})$
  (min, value) $\leftarrow$ **minimize** target(m_0) **for** m_0 **in** $(0, 1)$ ▷ Optimization
  power $\leftarrow$ 1-value
  increase δ
end while
return min

suggest, their proposed split does not depend on the dimension of the null hypothesis. This dimensionality is, however, incorporated in our proposed choice of the splitting ratio.

Note that analytically the split (9), proposed by Dunn et al. [2], converges to 0.5 for high dimensions d. In a similar fashion, our proposed splitting ratio converges to 0.5 for high dimensions when testing a fixed number of parameters to be constant, that is $d - k$ is fixed. However, we can see that the optimal splitting ratio varies with the dimension of the tested hypothesis space, thus, e.g. in settings where the number of tested, fixed parameters in the hypothesis grows proportional with the dimension, a splitting ratios below 0.5 is beneficial, even in the limit.

Further, due to the impact of the noncentrality parameter and the complexity of the limit distribution, it is difficult to determine the optimal splitting ratio analytically. To avoid extensive Monte Carlo approximation for the calculation of the optimal splitting ratio, we additionally propose the following computationally fast alternative. Instead of using extensive simulations to approximate the noncentral split chi-square distribution, we employ normal approximations with the expectation and variance calculated based on Corollary 3.4. This then leads to Algorithm 1, which very quickly determines an optimal splitting ratio based on the dimension of the null hypothesis k, the dimension of the parameter space d, and the significance level α via repeated minimization of values of the standard Gaussian cdf (the `pnorm` function in `R`).

Our simple algorithm constitutes a fast solution to obtain an optimal splitting ratio based on normal approximations of the limit distribution. Without considering the variance, our algorithm can be interpreted as maximizing the expectation of the limit distribution of the SLRT, see Corollary 3.4. However, this expectation depends on the distance of the local alternative δ. Thus, to achieve a meaningful choice between the different splitting ratios, we first scaled δ to represent a setting where the test still has power against the local alternative but the task is nevertheless non-trivial. Furthermore, to get some more intuition for the behaviour of our proposed splitting ratio, Figure 4 provides a visual rule of thumb for the choice of the optimal splitting ratio based on the ratio of the

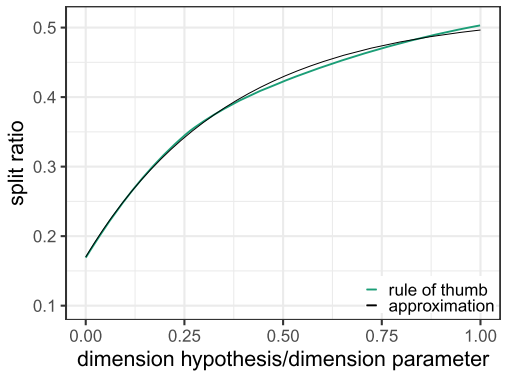


FIG 4. *Rule of thumb for the choice of the split ratio in high dimensions.*

dimension of the hypothesis space and the dimension of the parameter space derived in a high-dimensional setting. We found the impact of the significance level α in this high-dimensional setting to be negligible. For further simplicity and easy computation, this visual rule of thumb can be approximated with the function $m_0 = -\exp(-2.7\frac{k}{d} - 1.05) + 0.52$.

In Section 4.2 we analyze the performance of both proposed methods, the Monte Carlo method and the normal approximation variant numerically and show that it outperforms the existing proposal by Dunn et al. [2] in regular settings. Further, in Section 4.3 we show that this splitting ratio leads to optimal power even in the investigated irregular settings.

4. Simulations

We present the results of a simulation study that investigates the asymptotic behavior of the SLRT and compare its performance in different model settings, namely, regular and irregular settings, different dimensions of the parameter space d, different dimensions of the null hypothesis space k, and different splitting ratios m_0. All reported quantities are computed from simulations with 100,000 replications and if not stated otherwise we use the significance level $\alpha = 0.05$.

4.1. Power of the SLRT in regular setting

How much does using the universal threshold cost in terms of power asymptotically? In the following, we explore this question by comparing the power of the SLRT using the two different critical values, the standard universal threshold (`SLRT`) and the quantile of the asymptotic distribution (`Asym`). To this end we consider samples from a d-dimensional multivariate standard normal distribution with mean vector $(\theta, \dots, \theta)$ with $\theta = 0.1$ and test the hypothesis that the first $d - k$ entries of θ equal zero.

Figures 5 and 6 display the (simulated) power of both variants as well as that of the classical LRT against the sample size. In this regular setting where the

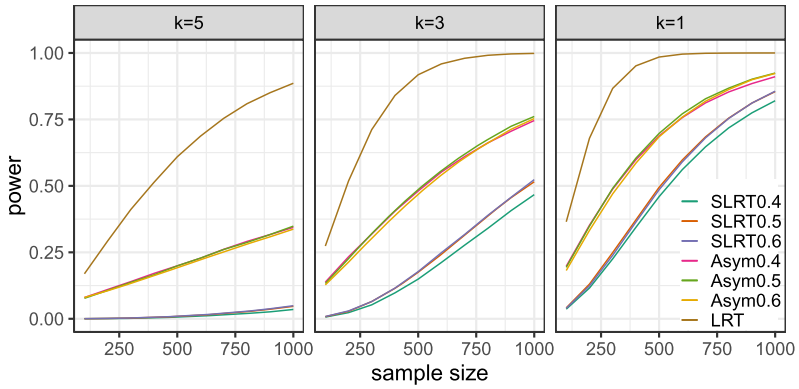


FIG 5. *Power against sample size in Gaussian setting with* $\theta = 0.1$, $d = 6$.

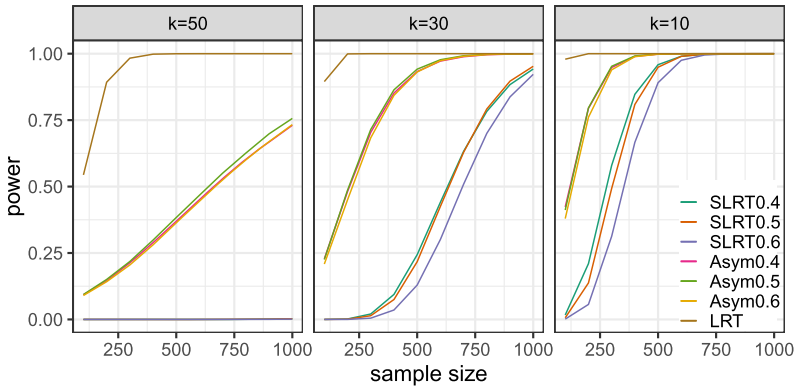


FIG 6. *Power against sample size in Gaussian setting with* $\theta = 0.1$, $d = 60$.

classical asymptotic distribution theory holds, the LRT outperforms the SLRT also when using the asymptotically correct quantiles. Furthermore, we see again that the power loss from using the universal threshold is larger in higher dimensions and higher dimensional null hypothesis settings. The simulations show that the choice of the splitting ratio plays an important role in the performance of the SLRT, especially in higher dimensional settings. In the following, we further examine the impact of the optimal choice of the splitting ratio in simulations.

4.2. Influence of the splitting ratio

In the following experiments, we analyze the influence of the splitting ratio on the asymptotic power of the SLRT. To this end, we sample data from a noncentral split chi-square distribution and calculate the power for testing the hypothesis of a zero noncentrality parameter δ. Figures 7 and 8 show the (simulated) power against the splitting ratio for the two different critical values, the universal threshold (`SLRT`) and the asymptotic quantile (`Asym`). We can see that in the lower dimensional setting a splitting ratio above 0.5 performs best

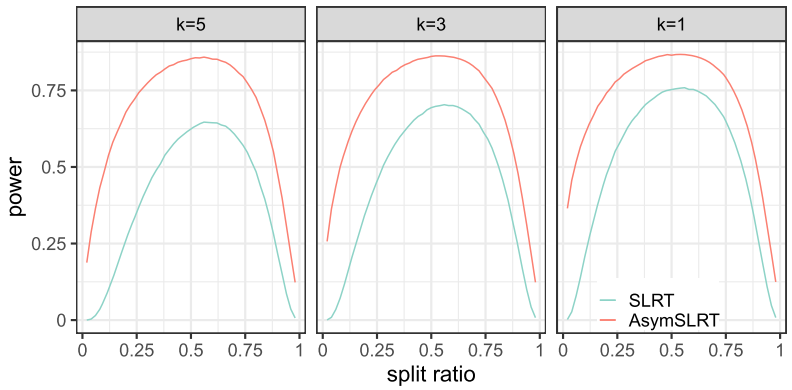


FIG 7. *Power of SLRT against splitting ratio,* $d = 6$, $\delta = 40$.

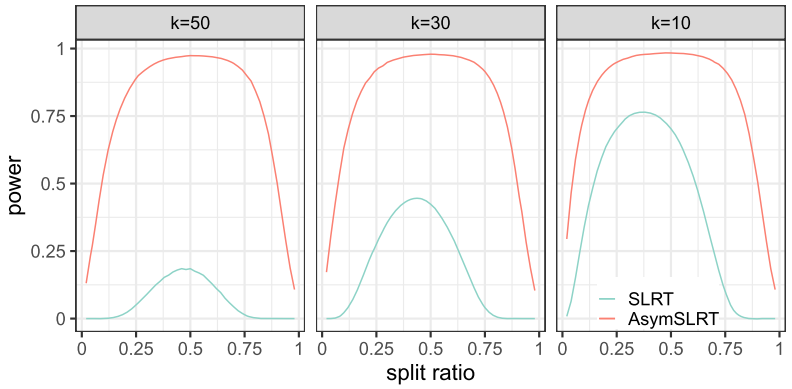


FIG 8. *Power of SLRT against splitting ratio,* $d = 60$, $\delta = 180$.

while in the higher dimensional setting a smaller splitting ratio below 0.5 seems beneficial, especially for a lower dimensional null hypothesis.

Figure 9 quantifies the improvement in power that can be achieved with our proposed (empirical) optimal splitting ratio (`emp.optim`) and the fast estimation routine (`est.optim`) that uses a normal approximation instead of extensive simulations to approximate the power of the SLRT. We compare the power for different noncentrality parameters $\delta \in \{100, 250\}$ plotted as 'dashed' and 'solid' lines respectively. Figure 9 displays that our fast estimation routine of the optimal split leads to valid approximations with a similarly good performance as the empirical optimal splitting ratio and further that there is a notable gain in power by using our new proposed optimal splitting ratios compared to the split (9) by Dunn et al. [2], especially in higher dimensions.

This is even more apparent in Figure 10, where we calculated the power for two different regimes of increasing noncentrality parameter δ. For each dimension of the parameter space, we chose the smallest δ such that the test with our new proposed optimal splitting ratio achieves a power of 0.8 and 0.65 respectively. While our methods, therefore, keep the power level, the split from (9) leads to a rapid loss of power in higher dimensions.

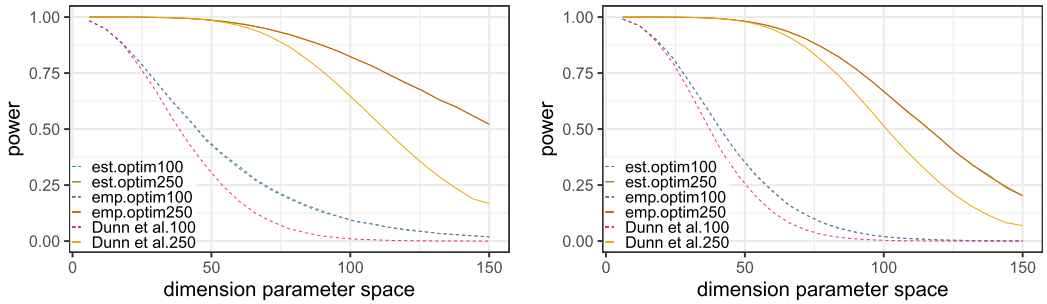


FIG 9. *Power for fixed noncentrality parameter and* $k = 5$ *(left);* $k = \frac{1}{6}\,d$ *(right).*

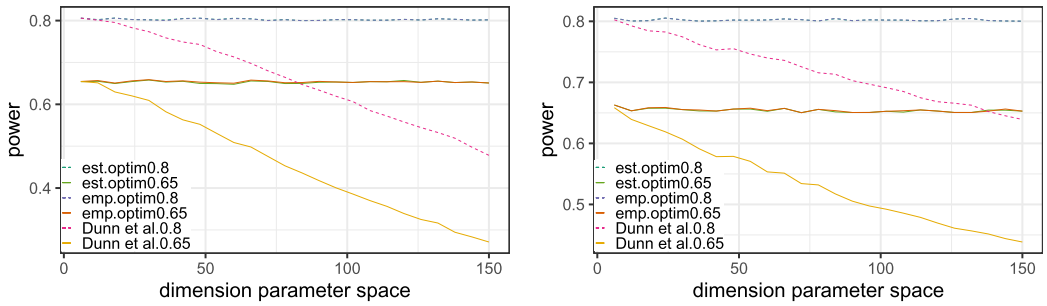


FIG 10. *Power for increasing noncentrality parameter and* $k = 5$ *(left);* $k = \frac{1}{6}\,d$ *(right)*

4.3. Power of the SLRT in irregular setting

Our proposed optimal splitting ratio is based on the limit distribution of the SLRT obtained under regularity conditions. In the following, we show that this optimal choice of the splitting ratio leads to an improvement in power even in irregular settings. To this end, we investigate the performance of the SLRT in the one-factor analysis setting with 12 observed variables. Assuming zero means, the one-factor model is given by the family of normal distributions $\mathcal{N}_{12}(0, \Sigma)$ with $\Sigma \in F_{12,1} := \{\Omega + \Gamma\Gamma^T : \Omega \in \mathbb{R}^{12\times 12}_{>0}$ diagonal, $\Gamma \in \mathbb{R}^{12}\}$. In the following experiment, we consider testing the one-factor model against the saturated alternative, that is, the entire cone of positive definite matrices $PD(12)$. Note that the hypothesis defines a 24-dimensional subset of the 78-dimensional parameter space, thus, our algorithm suggests using the optimal splitting ratio 0.41 while (9) suggests using the splitting ratio 0.51. Drton [1] shows that in this one-factor analysis setting the hypothesis space has singularities at loadings Γ with less than 3 nonzero values and thus, at those points, classic asymptotic distribution theory for likelihood ratio tests is not valid.

Figure 11 displays the power of the SLRT under alternatives around irregular and regular points using the two different splitting ratios. More specifically, we set Ω as the diagonal matrix with all diagonal entries equal 1/5 and $\Gamma = (5, 5, 0, ..., 0)$ for the irregular and $\Gamma = (5, 5, 5, 0, ..., 0)$ for the regular setting, respectively. Then we sampled $n = 2000$ data points from a two-factor alterna-

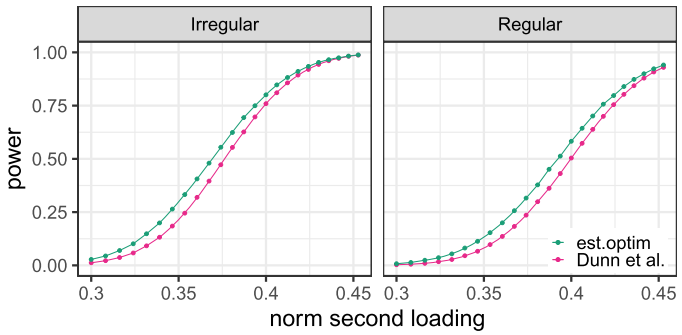


FIG 11. *Power of SLRT in one-factor analysis.*

tive $\mathcal{N}_{12}(0,\Sigma)$ with $\Sigma = \Omega + \Gamma\Gamma^T + \Gamma_2\Gamma_2^T$, where $\Gamma_2 = (h/\sqrt{12}, ..., h/\sqrt{12})$ with varying values of the norm h of the second factor loading Γ_2. We observe, that the optimality of our proposed splitting ratio extends to this irregular setting and our splitting ratio outperforms the splitting ratio suggested by Dunn et al. [2] in all settings.

5. Extension to the cross-fit SLRT

The SLRT and the universal inference framework starts with randomly splitting the available data. Thus, the value of the split likelihood ratio statistic still varies given a fixed data set, depending on the random split. It is natural to think about reducing this randomness by aggregating the results of different splits at the cost of performing more computations. Already in their initial work Wasserman, Ramdas and Balakrishnan [8] propose the cross-fit SLRT, a variant of the SLRT where the test statistic is calculated twice with alternating roles of the two data sets. Then, both results are averaged to obtain the cross-fit split likelihood ratio statistic $\frac{1}{2}(\Lambda_n + \Lambda_n^{swap})$, where Λ_n^{swap} is defined by (2) with the roles of D_0 and D_1 swapped.

Similar to the proof of Theorem 3.1 we can derive the asymptotic distribution of the cross-fit split likelihood ratio statistic.

Corollary 5.1. *Under assumptions (A1)-(A3), the cross-fit split likelihood ratio statistic satisfies under* P_{θ_n} *with* $\theta_n = \theta_0 + h/\sqrt{n}$

$$\begin{aligned}\Lambda_n + \Lambda_n^{swap} \xrightarrow{\mathcal{D}} &\|X + \sqrt{m_0} I(\theta_0)^{1/2} h - I(\theta_0)^{1/2} H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}} Y\|^2 \\ &+ \|Y + \sqrt{m_1} I(\theta_0)^{1/2} h - I(\theta_0)^{1/2} H_0\|^2 - \|Y - \sqrt{\tfrac{m_1}{m_0}} X\|^2,\end{aligned}$$

where $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ *independent and* $\|x - H_0\| = \inf_{h \in H_0} \|x - h\|$.

Since this cross-fit variant of the SLRT only splits the data once and then uses the same data sets twice with alternating roles, the limit distribution is still generated by two independent random variables. Analogous to the SLRT, we can thus calculate the expectation and variance of the arising limit distribution in

the smooth setting of Corollary 3.6, where the limiting hypothesis is a coordinate subspace, in the same way as the proof of Corollary 3.4 by exploiting properties of quadratic forms.

Corollary 5.2. *In the smooth setting of Corollary 3.6 the expectation and variance of the limit distribution of the cross-fit split likelihood ratio test statistics is given by*

1. $\mathbb{E}[\frac{1}{2}(\Lambda_\infty + \Lambda_\infty^{swap})] = p - d - \frac{1}{2}d\big(\frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big) + \frac{1}{2}\delta,$
2. $\mathrm{Var}[\frac{1}{2}(\Lambda_\infty + \Lambda_\infty^{swap})] = (d-p)\big(1 + \frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big) + d\big(2 + \frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big)$
$+\frac{1}{2}d\Big(\frac{m_0^2}{(1-m_0)^2} + \frac{(1-m_0)^2}{m_0^2}\Big) + \delta,$

where $\delta = [I(\theta_0)^{1/2}h]_{[p]}^T[I(\theta_0)^{1/2}h]_{[p]}$.

Furthermore, analogously to Section 3.2, we can employ the limit distribution to determine the optimal splitting ratio and obtain the intuitive result that for the cross-fit SLRT an even split of $m_0 = 0.5$ is optimal in all dimensions.

Remark 5.3. For an equal splitting ratio $m_0 = 0.5$ the cross-fit split likelihood ratio statistic has the same expectation but a lower variance than the split likelihood ratio statistic in the limit. Thus, in situations where the method has power, the cross-fit SLRT further improves upon the SLRT.

The cross-fit SLRT employs equal weights $w_0 = 0.5$ to combine both test statistics and obtain the cross-fit split likelihood ratio test statistic, that is $w_0\Lambda_n + (1-w_0)\Lambda_n^{swap}$. Using equal weights is intuitive for an even splitting ratio $m_0 = 0.5$, since both test statistics are expected to perform similarly. However, considering the results from the previous section, for different splitting ratios, it might be beneficial to vary the weights and emphasize the better-performing test statistic. Furthermore, the idea of the cross-fit SLRT of swapping the roles of the two data sets allows us to derive properties of the limit distribution and is conceptually simple. Nevertheless, in view of our previous analysis, subsampling provides a promising alternative. Instead of using the same split data set with swapped roles to calculate the second test statistic, we repeat the process of randomly splitting the available data. Such a method has a similar computational burden as the cross-fit SLRT, however, we can make use of the optimal splitting ratio for both test statistics. Note that this subsampling procedure can be extended to multiple repeats and thus, further decrease the randomness of the splits at the cost of computation time.

Figure 12 compares the performance of the different extensions of the SLRT in the one-factor analysis setting introduced in Section 4.3. We display the power of the cross-fit SLRT with different splitting ratios m_0 and different weights w_0 as well as the subsampling alternative with our proposed optimal splitting ratio $m_0 = 0.41$. As previously mentioned, under the assumption of fixed, equal weights $w_0 = 0.5$, the best-performing splitting ratio for the cross-fit SLRT is an even split $m_0 = 0.5$. However, for an uneven splitting ratio, we can emphasize the better-performing test statistic by adjusting the weights to achieve similar performance. Furthermore, Figure 12 shows that using our proposed optimal

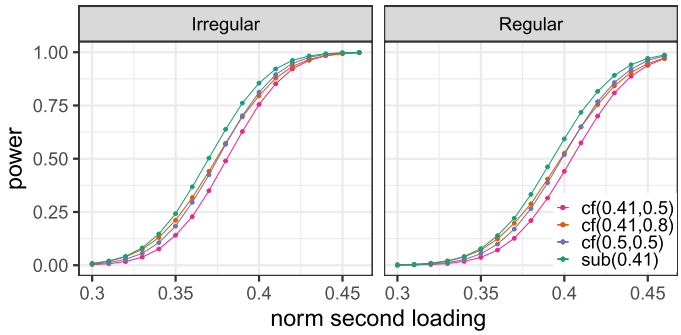


FIG 12. *Power of cross-fit(m_0, w_0) and subsampling SLRT in one-factor analysis.*

splitting ratio for both test statistics in a subsampling procedure outperforms the competition in all settings.

6. Conclusion

The split likelihood ratio test (SLRT) is a flexible tool that provides valid level α tests in finite samples even when classical regularity conditions are not satisfied. The underlying universal approach of splitting the data allows one to conduct rather simple analyses even in complicated settings. In general, this flexibility leads to a rather conservative method and, thus, it is of interest to carefully choose the splitting ratio in order to mitigate possible loss of power.

In order to provide new insights about the performance of the SLRT we studied its asymptotic behavior in the setting of smooth hypotheses. Our study gives rise to a new class of distributions, noncentral split chi-square distributions, that appear as limiting distributions of the SLRT. The split chi-square distribution depends on the dimensions of both null and alternative hypotheses and not only the difference of the dimensions. Naturally, it also depends on the chosen data splitting ratio. Using the new class of distributions, we analyzed the power of the SLRT in extensive simulations, and we proposed a new routine for calculating the optimal splitting ratio for the SLRT that significantly boosts power, especially in higher dimensions.

References

[1] DRTON, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658

[2] DUNN, R., RAMDAS, A., BALAKRISHNAN, S. and WASSERMAN, L. (2021). Gaussian Universal Likelihood Ratio Testing.

[3] HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser.* 807–810. Wadsworth, Belmont, CA. MR822066

[4] Kendall, M. G. and Stuart, A. (1969). *The advanced theory of statistics. Vol. 1: Distribution theory*, Third ed. Hafner Publishing Co., New York. MR0246399 MR0243648

[5] Strieder, D., Freidling, T., Haffner, S. and Drton, M. (2021). Confidence in causal discovery with linear causal models. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* (C. de Campos and M. H. Maathuis, eds.). *Proceedings of Machine Learning Research* **161** 1217–1226. PMLR.

[6] Tse, T. and Davison, A. C. A Note on Universal Inference. *Stat* e501.

[7] van der Vaart, A. W. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247 MR1652247

[8] Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890.

Permission to include the article

Institute of Mathematical Statistics
Fostering the development and dissemination of the theory and applications of statistics and probability
Renew / Join IMS

Electronic Journal of Statistics Copyright

Electronic Journal of Statistics has chosen to apply the Creative Commons Attribution License (CCAL) to all articles we publish in this journal (click here to read the full-text legal code). Under the CCAL, authors retain ownership of the copyright for their article, but authors allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in EJS, so long as the original authors and source are credited. This broad license was developed to facilitate open access to, and free use of, original works of all types. Applying this standard license to your work will ensure your right to make your work freely and openly available.

Summary of the Creative Commons Attribution License

You are free:

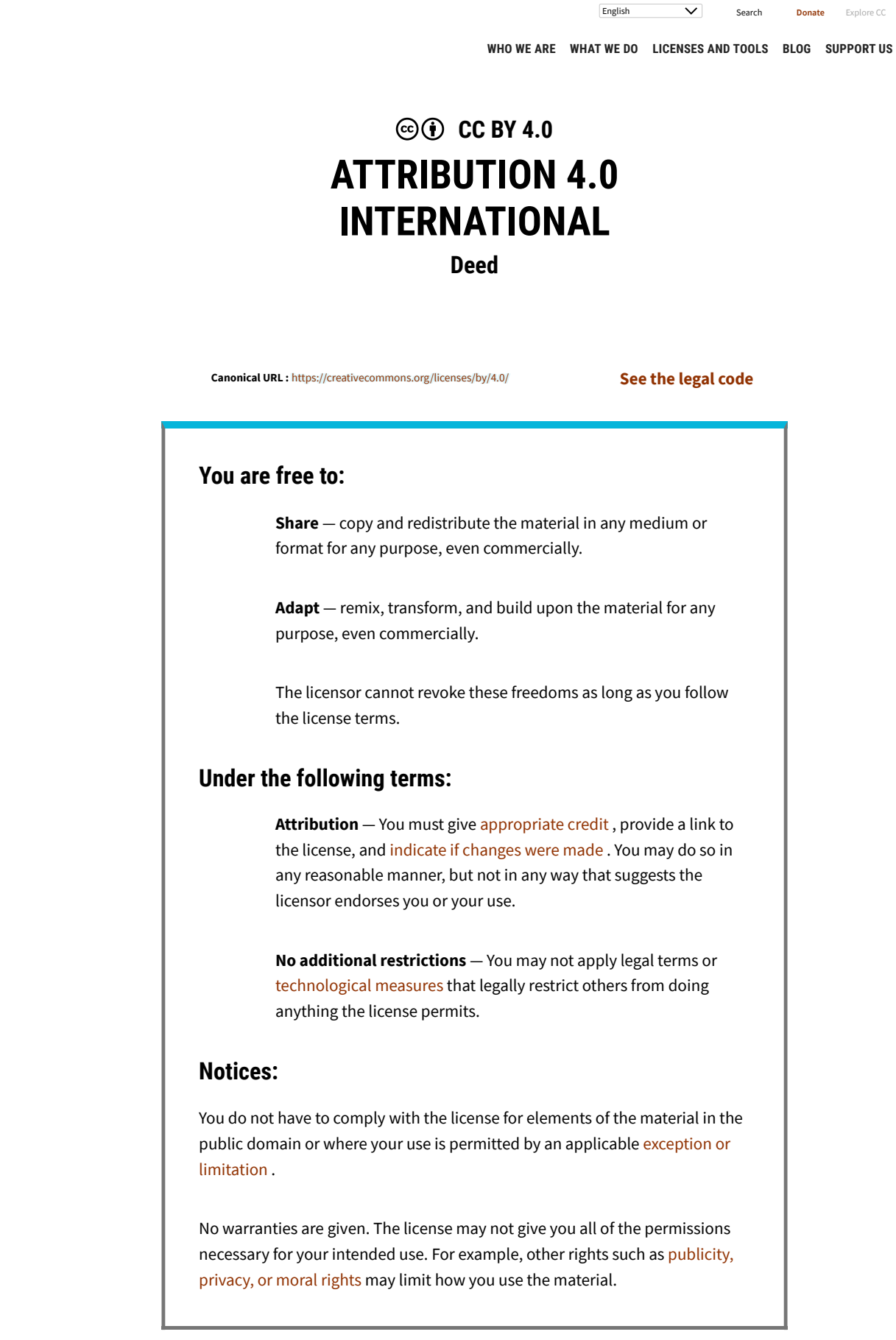
to copy, distribute, display, and perform the work
to make derivative works
to make commercial use of the work

under the following condition of Attribution: others must attribute the work if displayed on the web or stored in any electronic archive by making a link back to the website of the Journal via its Digital Object Identifier (DOI), or if published in other media by acknowledging prior publication in this Journal with a precise citation including the DOI. For any further reuse or distribution, the same terms apply. Any of these conditions can be waived by permission of the Corresponding Author.

We are using the same license as the Public Library of Science (PLoS).
http://www.plos.org/

Institute of Mathematical Statistics
toll free (US): 877.557.4674
tel: 216.295.2340
fax: 216.295.5661
email: ims@imstat.org

Permission to include the article



English Search Donate Explore CC

WHO WE ARE WHAT WE DO LICENSES AND TOOLS BLOG SUPPORT US

CC BY 4.0

ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL : https://creativecommons.org/licenses/by/4.0/

See the legal code

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation .

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Permission to include the article

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- Learn more about our work
- **Learn more about CC Licensing**
- Support our work
- Use the license for your own material.
- Licenses List
- Public Domain List

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- More info

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indication of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- Marking guide
- More info

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WIPO Copyright Treaty.

- More info

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- More info

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- More info

A.3 Dual Likelihood for Causal Inference under Structure Uncertainty

Dual Likelihood for Causal Inference under Structure Uncertainty

David Strieder DAVID.STRIEDER@TUM.DE and **Mathias Drton** MATHIAS.DRTON@TUM.DE
Munich Center for Machine Learning and Department of Mathematics, TUM School of Computation, Information and Technology, Technical University of Munich, Germany

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Knowledge of the underlying causal relations is essential for inferring the effect of interventions in complex systems. In a widely studied approach, structural causal models postulate noisy functional relations among interacting variables, where the underlying causal structure is then naturally represented by a directed graph whose edges indicate direct causal dependencies. In the typical application, this underlying causal structure must be learned from data, and thus, the remaining structure uncertainty needs to be incorporated into causal inference in order to draw reliable conclusions. In recent work, test inversions provide an ansatz to account for this data-driven model choice and, therefore, combine structure learning with causal inference. In this article, we propose the use of dual likelihood to greatly simplify the treatment of the involved testing problem. Indeed, dual likelihood leads to a closed-form solution for constructing confidence regions for total causal effects that rigorously capture both sources of uncertainty: causal structure and numerical size of nonzero effects. The proposed confidence regions can be computed with a bottom-up procedure starting from sink nodes. To render the causal structure identifiable, we develop our ideas in the context of linear causal relations with equal error variances.

Keywords: linear structural causal models, graphical models, dual likelihood, causal effects, uncertainty quantification

1. Introduction

Reasoning about causal relations and inferring the effect of interventions in complex systems is a central task of scientific research. The field of causal discovery addresses this challenge with much fundamental research being done over the last decades (Pearl, 2009; Spirtes et al., 2000). In particular, in many applied settings, access to interventional data is limited, and thus, performing classical controlled experiments to investigate causal relations is not feasible. To tackle this problem, various identifiability results clarify under which conditions interventional distributions can theoretically be identified from observational data alone. Furthermore, many structure learning algorithms have been proposed that estimate causal structures using only observed data. On the other hand, when the underlying causal structure is known, causal inference results enable researchers to estimate causal quantities in complex systems and assess remaining uncertainty. However, when the underlying causal structure is learned from data, the remaining structure uncertainty needs to be incorporated into causal inference in order to draw reliable conclusions.

Recently, Strieder et al. (2021) introduced a first ansatz for constructing confidence regions for causal effects that account for both types of uncertainty, uncertainty about the causal structure and uncertainty about the numerical size of the effect. They consider bivariate linear structural causal models (SCMs) and develop a solution based on test inversion. Their strategy was improved and generalized to higher-dimensional linear SCMs in Strieder and Drton (2023). In this work, we

take up their idea of using a test inversion ansatz to construct confidence regions for total causal effects (see Section 2). However, in contrast to the earlier work, we suggest here to employ dual likelihood for the arising testing problem. An important consequence of the use of dual likelihood is that we are able to provide a closed-form solution for our proposed confidence regions (Section 3) and can, thus, avoid the numerical optimization algorithms deployed by Strieder and Drton (2023). Furthermore, while the earlier methods rely on a top-down procedure starting from the source node, by employing dual likelihood, we obtain a bottom-up procedure starting from the sink node. Thus, our proposed dual likelihood method provides an interesting alternative perspective on strategies to save on computation. Finally, in Section 4.2, we present computational details and analyze the performance in a simulation study. Further simulation results can be found in Appendix A. In the concluding Section 5, we discuss extensions and possible future work.

2. Background

A common tool for studying causal relationships among interacting variables are structural causal models (Peters et al., 2017; Maathuis et al., 2019), where each variable is represented as a function of other variables (its causes) and a random error term. The causal perspective views these relationships as assignments rather than mathematical equations, and thus, changes in the causes result in changes in the effects but not vice versa, reflecting the inherent asymmetry in cause-effect relationships. In this section, we review a restricted class of causal models, namely linear structural causal models with Gaussian errors and equal error variances (Peters and Bühlmann, 2014), as well as the definition of the total causal effect, which is the causal target of interest in this study. Further, we review the test inversion ansatz introduced by Strieder et al. (2021) to construct confidence regions for causal effects that account for structure uncertainty.

2.1. Causal Effects in Linear Structural Causal Models

We assume access to observational data in the form of n independent copies of a random vector $X = (X_1, ..., X_d)$ with zero mean. To ensure unique identifiability, we follow a line of research introduced by Peters and Bühlmann (2014) that focuses on linear relations and normal distributed errors with equal variances, represented by the equation system

$$X_j = \sum_{i \neq j} \beta_{j,i} X_i + \varepsilon_j, \quad j = 1, \dots, d, \tag{1}$$

where $B := [\beta_{j,i}]_{j,i=1}^{d}$ represents the direct causal effects between variables, and ε_j are independent, normally distributed error terms with common variance σ^2. Such a linear structural causal model is naturally represented by a directed graph, where a missing edge $i \to j$ indicates no direct effects, that is, $b_{j,i} = 0$. Further, we assume the underlying graph to be acyclic, which entails the unique solution $X = (I_d - B)^{-1}\varepsilon$ of the equation system (1), where I_d denotes the $d \times d$ identity matrix. The corresponding covariance matrix is given by $\Sigma = \sigma^2(I_d - B)^{-1}(I_d - B)^{-T}$.

In the remainder of the article, we use the following notation and graphical concepts. We write $i <_G j$ if node i precedes node j in a causal ordering of the corresponding directed acyclic graph (DAG). If the DAG contains an edge from node i to node j, then node i is called a parent of node j, and we denote the set of all parents of node j with $p(j)$. Further, if the DAG contains a directed path from node i to node j, then node j is called a descendant of node i, and we denote the set of

all descendants of node i with $d(i)$. Finally, we write $\Sigma_{j,i|p(i)}$ for the conditional covariance matrix, that is,

$$\Sigma_{j,i|p(i)} := \Sigma_{j,i} - \Sigma_{j,p(i)}(\Sigma_{p(i),p(i)})^{-1}\Sigma_{p(i),i}.$$

Our aim is to estimate the total causal effect $\mathcal{C}(i \to j)$ in linear SCMs and rigorously quantify the remaining uncertainty. Formally, the total causal effect is defined as the unit change in the expectation of X_j with respect to an intervention in X_i

$$\mathcal{C}(i \to j) := \frac{\mathrm{d}}{\mathrm{d}x_i}\mathbb{E}[X_j|\ \mathrm{do}(X_i = x_i)] = \Sigma_{j,i|p(i)}(\Sigma_{i,i|p(i)})^{-1}.$$

While this parameter of interest is given as a simple function of the covariance matrix, difficulties arise in practice when the underlying causal structure is unknown. In such situations, the conditioning set is unknown and has to be inferred from data, which introduces structure uncertainty. We emphasize that due to the identifiability of Gaussian SCMs with equal error variances (e.g., Chen et al., 2019; Ghoshal and Honorio, 2018), estimating the total causal effect is nevertheless a well-defined problem.

A naive two-step approach that treats the two tasks of causal discovery and causal inference separately by learning the causal structure first and then calculating confidence intervals in the inferred model does not account for uncertainty in the structure. Furthermore, classical bootstrapping methods also fail to correctly account for model uncertainty due to singularities at the intersection of models. To address this issue, Strieder et al. (2021) introduced a framework for constructing confidence intervals that rigorously account for structure uncertainty, which was then generalized in Strieder and Drton (2023). However, thoroughly accounting for structure uncertainty proves a challenging task, given the enormous number of possible structures, even in moderate dimensions.

2.2. Confidence Intervals via Test Inversion

The main idea is to leverage the classical duality between statistical hypothesis tests and confidence regions (Casella and Berger, 1990). To this end, we consider all possible causal structures in the following way. If we perform valid hypothesis tests for all attainable causal effects, then all values that cannot be rejected form a confidence region for the total causal effect. This shifts the task to constructing suitable hypothesis tests for all attainable causal effects.

Under the assumption of an underlying linear structural causal model with equal error variances, every possible distribution $N(0, \Sigma)$ is given by a set of covariance matrices $\mathcal{M} := \bigcup_{G\in\mathcal{G}(d)} \mathcal{M}(G)$, where $\mathcal{G}(d)$ is the set of all possible complete DAGs on d nodes and

$$\mathcal{M}(G) = \Big\{\Sigma \in \mathrm{PD}(d) : \exists\sigma^2 > 0 \text{ with } \sigma^2 = \Sigma_{k,k|p(k)} \quad \forall\, k = 1, \dots, d\Big\}.$$

Here, we write $\mathrm{PD}(d)$ for the cone of positive definite $d \times d$ matrices. The hypothesis of a fixed causal effect $\mathcal{C}(i \to j)$ of size ψ further restricts the set of possible distributions, where the specific structure of the constraint depends on the DAG G. Thus, with

$$\mathcal{M}_\psi(G) := \Big\{\Sigma \in \mathrm{PD}(d) : \exists\sigma^2 > 0 \text{ with } \psi\sigma^2 = \Sigma_{j,i|p(i)} \text{ and } \sigma^2 = \Sigma_{k,k|p(k)} \quad \forall\, k = 1, \dots, d\Big\}, \tag{2}$$

and $\mathcal{M}_\psi := \bigcup_{G\in\mathcal{G}(d)} \mathcal{M}_\psi(G)$, the task is to invert the statistical testing problem

$$H_0^{(\psi)} : \Sigma \in \mathcal{M}_\psi \quad \text{against} \quad H_1 : \Sigma \in \mathcal{M}\backslash\mathcal{M}_\psi \tag{3}$$

for all $\psi \in \mathbb{R}$.

3. Testing for Total Causal Effects with the Dual Likelihood

Earlier methods (Strieder and Drton, 2023) propose to use constrained likelihood ratio tests for the testing problem (3). However, the direct maximization of the Gaussian likelihood with constraints on the total causal effect has no closed-form solution, which leads to algorithms that rely on numerical optimization routines and grid searches. Instead, our proposal is to employ dual likelihood theory for Gaussian models (e.g., Brown, 1986; Kauermann, 1996) to solve the testing problem (3) and, thus, obtain a simple closed-form solution.

To obtain our main result, we first introduce the dual likelihood for Gaussian graphical models. Maximizing the Gaussian likelihood corresponds to the minimization problem

$$\operatorname*{arg\,min}_{\Sigma\in\mathcal{M}} \mathrm{KL}(P_{\widehat{\Sigma}}, P_{\Sigma}) = \operatorname*{arg\,min}_{\Sigma\in\mathcal{M}} \big(\mathrm{tr}(\Sigma^{-1}\widehat{\Sigma}) + \log\det(\Sigma)\big),$$

where KL is the Kullback–Leibler divergence with sample covariance $\widehat{\Sigma}$. As the Kullback–Leibler divergence is not symmetrical in its arguments, the minimization problem

$$\begin{aligned}\operatorname*{arg\,min}_{\Sigma\in\mathcal{M}} \mathrm{KL}(P_{\Sigma}, P_{\widehat{\Sigma}}) &= \operatorname*{arg\,min}_{\Sigma\in\mathcal{M}} \big(\mathrm{tr}(\widehat{\Sigma}^{-1}\Sigma) - \log\det(\Sigma)\big) \\ &= \operatorname*{arg\,min}_{\Sigma\in\mathcal{M}} \big(\mathrm{tr}(\Sigma\widehat{\Sigma}^{-1}) + \log\det(\Sigma^{-1})\big)\end{aligned}$$

yields a different estimator, the dual maximum likelihood estimator. Along similar lines, the dual (log-)likelihood of Σ for the observed parameter $\widehat{\Sigma}^{-1}$ is defined as

$$\tfrac{2}{n}\ell_n^{dual}(\Sigma|\widehat{\Sigma}^{-1}) := -\log\det(2\pi\Sigma^{-1}) - \mathrm{tr}(\Sigma\widehat{\Sigma}^{-1}).$$

We note that this dual likelihood is reciprocal to the Gaussian likelihood in the sense that for observed sample covariance $\widehat{\Sigma}^{-1}$, we have

$$\ell_n^{dual}(\Sigma|\widehat{\Sigma}^{-1}) = \ell_n(\Sigma^{-1}|\widehat{\Sigma}^{-1}), \tag{4}$$

where $\frac{2}{n}\ell_n(\Sigma|\widehat{\Sigma}) := -\log\det(2\pi\Sigma) - \mathrm{tr}(\Sigma^{-1}\hat{\Sigma})$.

3.1. Dual Likelihood Ratio Test

In order to construct confidence intervals for the total causal effects, our idea is to invert dual likelihood ratio tests of the testing problem (3) with the dual likelihood ratio test statistic

$$\begin{aligned}\text{dual-}\check{\lambda}_n^{(\psi)} &:= 2\Big(\sup_{\Sigma\in\mathcal{M}} \ell_n^{dual}(\Sigma|\widehat{\Sigma}^{-1}) - \sup_{\Sigma\in\mathcal{M}_\psi} \ell_n^{dual}(\Sigma|\widehat{\Sigma}^{-1})\Big) \\ &= 2\Big(\sup_{\Sigma\in\mathcal{M}} \ell_n(\Sigma^{-1}|\widehat{\Sigma}^{-1}) - \sup_{\Sigma\in\mathcal{M}_\psi} \ell_n(\Sigma^{-1}|\widehat{\Sigma}^{-1})\Big).\end{aligned} \tag{5}$$

First, we observe that by employing equation (4), this dual likelihood ratio test statistic equals the classical likelihood ratio test statistic for a modified problem. More precisely, we redefine the model space $\mathcal{M}^{-1} := \bigcup_{G\in\mathcal{G}(d)} \mathcal{M}^{-1}(G)$ with

$$\mathcal{M}^{-1}(G) := \Big\{\Omega \in \mathrm{PD}(d) : \exists\tilde{\sigma}^2 > 0 \text{ with } \tilde{\sigma}^2 = \Omega_{k,k|d(k)} \quad \forall\, k = 1,\dots,d\Big\},$$

where $\Omega = \Sigma^{-1}$. Similarly, we redefine the hypothesis space $\mathcal{M}_\psi^{-1} = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_\psi^{-1}(G)$ with

$$\mathcal{M}_\psi^{-1}(G) := \Big\{ \Omega \in \mathrm{PD}(d) : \exists \tilde{\sigma}^2 > 0 \text{ with } \tilde{\sigma}^2 = \Omega_{k,k|d(k)} \quad \forall\, k = 1, \dots, d \\ \text{and } \psi = \tau_j\big(\Omega_{i,d(i)}(\Omega_{d(i),d(i)})^{-1}\big) \Big\},$$

where τ_j projects the $|d(i)|$-dimensional vector onto the component corresponding to j if $j \in d(i)$ and zero otherwise. Then, we obtain the following Lemma.

Lemma 1 *Let $\Sigma \in PD(d)$. Then the following equivalences hold.*

1. *$\Sigma \in \mathcal{M}$ if and only if $\Sigma^{-1} \in \mathcal{M}^{-1}$.*
2. *$\Sigma \in \mathcal{M}_\psi$ if and only if $\Sigma^{-1} \in \mathcal{M}_\psi^{-1}$.*

Proof This result follows from straightforward calculations using the following observation. If $\Sigma = \sigma^2(I_d - B)^{-1}(I_d - B)^{-T}$ for some DAG G with edge weights B and equal error variance $\sigma^2 > 0$, then $\Sigma^{-1} = \sigma^{-2}(I_d - T)^{-1}(I_d - T)^{-T}$, where T represents the negative total causal effect between variables. Thus, Σ^{-1} corresponds to the covariance matrix of a DAG G', where the parent sets $p(i)$ in G' are the descendants $d(i)$ in G for all $i = 1, \dots, d$, the edge weights are given by T and equal error variance σ^{-2}. Constraints on the total causal effects in Σ and G thus correspond to constraints on direct effects in Σ^{-1} and G'. ■

Using Lemma 1, we immediately obtain that the dual likelihood ratio test statistic (5) corresponds to

$$\text{dual-}\check{\lambda}_n^{(\psi)} = 2\Big(\sup_{\Omega \in \mathcal{M}^{-1}} \ell_n(\Omega|\widehat{\Sigma}^{-1}) - \sup_{\Omega \in \mathcal{M}_\psi^{-1}} \ell_n(\Omega|\widehat{\Sigma}^{-1}) \Big),$$

which is the classical likelihood ratio test statistic for the modified testing problem

$$\tilde{H}_0^{(\psi)} : \Omega \in \mathcal{M}_\psi^{-1} \quad \text{against} \quad \tilde{H}_1 : \Omega \in \mathcal{M}^{-1} \backslash \mathcal{M}_\psi^{-1}, \tag{6}$$

with observed sample covariance $\widehat{\Sigma}^{-1}$. Thus, we can employ similar strategies to Strieder and Drton (2023) in order to obtain confidence regions for total causal effects.

3.2. Confidence Intervals with Dual Likelihood

In this subsection, we state our main result, a confidence region for total causal effects $\mathcal{C}(i \to j)$ that captures structure uncertainty as well as uncertainty about the numerical size of the effect. Our first step to tackle the testing problem (6) is to relax the alternative $\mathcal{M}^{-1}$ and employ the theory of intersection union tests, see e.g. Casella and Berger (1990), to obtain a simple upper bound on the distribution of the dual likelihood ratio test statistic via

$$\text{dual-}\check{\lambda}_n^{(\psi)} \leq \text{dual-}\lambda_n^{(\psi)}(G) := 2\Big(\sup_{\Omega \in \mathrm{PD}(d)} \ell_n(\Omega|\widehat{\Sigma}^{-1}) - \sup_{\Omega \in \mathcal{M}_\psi^{-1}(G)} \ell_n(\Omega|\widehat{\Sigma}^{-1}) \Big).$$

Further, we define $\text{dual-}\check{\lambda}_n^{(\psi)}(i <_G j) := \min_{G \in \mathcal{G}(d) \,:\, i <_G j} \text{dual-}\check{\lambda}_n^{(\psi)}(G)$, where

$$\text{dual-}\check{\lambda}_n^{(\psi)}(G) := 2\Big(\sup_{\Omega \in \mathcal{M}^{-1}} \ell_n(\Omega|\widehat{\Sigma}^{-1}) - \sup_{\Omega \in \mathcal{M}_\psi^{-1}(G)} \ell_n(\Omega|\widehat{\Sigma}^{-1}) \Big).$$

Then we obtain our following main result via test inversions of asymptotically conservative hypothesis test based on the asymptotic distribution of the upper bound dual-$\lambda_n^{(\psi)}(G)$ under every single hypothesis $\mathcal{M}_\psi^{-1}(G)$, see Lemma 3, as well as the closed-form solution of the dual likelihood estimation under constraints on the total causal effect, see Section 3.3.

Theorem 2 *Let $\alpha \in (0,1)$. Then an asymptotic $(1-\alpha)$-confidence set for the total causal effect $\mathcal{C}(i \to j)$ is given by*

$$C := \{\psi \in \mathbb{R} : \text{dual-}\check{\lambda}_n^{(\psi)}(i <_G j) \leq \chi^2_{d,1-\alpha}\} \cup \{0 : \text{dual-}\check{\lambda}_n^{(0)}(j <_G i) \leq \chi^2_{d-1,1-\alpha}\}.$$

This confidence set can be constructed explicitly as follows. Define

$$K := \min_{G \in \mathcal{G}(d)} \sum_{k=1}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)}, \qquad Z := \min_{G \in \mathcal{G}(d): j <_G i} \sum_{k=1}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)},$$

and for all $G \in \mathcal{G}(d)$ with $i <_G j$,

$$D_G := (\widehat{\Sigma}^{-1})^2_{i,j|d(i)\setminus\{j\}} - (\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}} \Big(\sum_{k \neq i}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)} + (\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} - K \exp\big(\tfrac{\chi^2_{d,1-\alpha}}{dn}\big) \Big).$$

Moreover, for $D_G \geq 0$, define

$$L_G := \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} - \sqrt{D_G}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}, \qquad U_G := \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + \sqrt{D_G}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}.$$

Then the closed-form solution for the confidence set C is given by

$$C = \bigcup_{G \in \mathcal{G}(d): D_G \geq 0} [L_G, U_G] \bigcup \{0 : Z \leq K \exp\big(\tfrac{\chi^2_{d-1,1-\alpha}}{dn}\big)\}.$$

Proof Let $\psi \in \mathbb{R}$ and $\Sigma \in \mathcal{M}_\psi$. Since $\mathcal{M}_\psi = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_\psi(G)$ there exists a complete graph G, such that $\Sigma \in \mathcal{M}_\psi(G)$. If $i <_G j$, it follows with the introduced upper bound and Lemma 3

$$P_\Sigma\Big(\text{dual-}\check{\lambda}_n^{(\psi)} > \chi^2_{d,1-\alpha}\Big) \leq P_\Sigma\Big(\text{dual-}\lambda_n^{(\psi)}(G) > \chi^2_{d,1-\alpha}\Big) \to \alpha.$$

Furthermore, plugging in the dual likelihood estimates (7) and (8) derived in Section 3.3, we view

$$\text{dual-}\check{\lambda}_n^{(\psi)}(G) - \chi^2_{d,1-\alpha} = 2\Big(\sup_{\Omega \in \mathcal{M}^{-1}} \ell_n(\Omega|\widehat{\Sigma}^{-1}) - \sup_{\Omega \in \mathcal{M}_\psi^{-1}(G)} \ell_n(\Omega|\widehat{\Sigma}^{-1}) \Big) - \chi^2_{d,1-\alpha}$$

as a strictly convex quadratic polynomial in ψ. With $K := \min_{G \in \mathcal{G}(d)} \sum_{k=1}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)}$ and

$$D_G := (\widehat{\Sigma}^{-1})^2_{i,j|d(i)\setminus\{j\}} - (\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}} \Big(\sum_{k \neq i}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)} + (\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} - K \exp\big(\tfrac{\chi^2_{d,1-\alpha}}{dn}\big) \Big),$$

this quadratic polynomial has real roots if $D_G \geq 0$. Thus, the inequality dual-$\check{\lambda}_n^{(\psi)}(G) \leq \chi^2_{d,1-\alpha}$ holds if and only if

$$\psi \in \left[\frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} - \sqrt{D_G}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}, \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + \sqrt{D_G}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}} \right].$$

We obtain an analogous result for $j <_G i$, and the test inversion approach yields the claim. ■

Note that for non-zero effects, we do not need to test among graphs with $j <_G i$. Thus, our confidence regions may consist of a non-zero interval and an isolated zero, reflecting the remaining uncertainty about the direction of the causal relation. In contrast, for zero-sized effects, we need to consider all causal orderings since multiple paths might cancel.

Our main result employs the following asymptotic distribution of the upper bound dual-$\lambda_n^{(\psi)}(G)$ under every single hypothesis $\mathcal{M}_\psi^{-1}(G)$.

Lemma 3 *Let $G \in \mathcal{G}(d)$ be a DAG and let $\psi \in \mathbb{R}$. Then, the relaxed dual likelihood ratio test statistic* dual-$\lambda_n^{\psi}(G)$ *satisfies one of the following properties.*

a) If $i <_G j$, then under the hypothesis $H_0^{(\psi)}(G)$

$$\text{dual-}\lambda_n^{(\psi)}(G) \xrightarrow{\mathcal{D}} \chi^2_d.$$

b) If $j <_G i$, then under the hypothesis $H_0^{(0)}(G)$

$$\text{dual-}\lambda_n^{(0)}(G) \xrightarrow{\mathcal{D}} \chi^2_{d-1}.$$

Proof We assume $i <_G j$. Without loss of generality, let $(1, 2, \ldots, d)$ be the causal ordering of G. Then $\mathcal{M}_\psi^{-1}(G) = \{\Omega \in \text{PD}(d) : f_\psi(\Omega|G) = 0\}$, with

$$f_\psi(\Omega|G) := \begin{pmatrix} \tau_j(\Omega_{i,\{i+1,\ldots,d\}}(\Omega_{\{i+1,\ldots,d\},\{i+1,\ldots,d\}})^{-1}) - \psi \\ \Omega_{1,1|\{2,\ldots,d\}} - \Omega_{d,d} \\ \vdots \\ \Omega_{d-1,d-1|d} - \Omega_{d,d} \end{pmatrix}.$$

The Jacobian of f_ψ has full rank d, which follows due to its (nonzero) triangular structure in derivatives for $\Omega_{k,k}$ for all $k = 1, \ldots, d-1$, while the first row has nonzero derivative for $\Omega_{i,j}$ but zero derivatives for $\Omega_{k,k}$ for all $k = 1, \ldots, i$. Thus, $\mathcal{M}_\psi^{-1}(G)$ defines a $\frac{1}{2}(d^2 - d)$-dimensional submanifold of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$. The case $j <_G i$ follows similarly, and thus, the claim follows. For details, we refer to Drton (2009). ■

Note that the constraint of a zero-sized effect does not restrict the submanifold $\mathcal{M}_0^{-1}(G)$ for $j <_G i$. All edge weights in the corresponding linear SCMs can vary freely, and thus, the two different degrees of freedom for the chi-square limits arise.

3.3. Dual Likelihood Estimation

Maximizing the classical Gaussian likelihood under constraints on total causal effects leads to complex optimization problems with polynomial constraints that need to be solved with numerical optimization routines. In contrast to that, we can explicitly maximize the dual likelihood under constraints on total causal effects. Thus, as an important consequence of the use of dual likelihood, we are able to provide a closed-form solution for our proposed confidence regions in Theorem 2.

In order to derive this closed-form solution, we need to calculate the dual likelihood ratio test statistic (5). This involves maximizing the dual likelihood in two cases, the unrestricted case $\Sigma \in \mathcal{M}$, and under the constraint of a fixed-size total causal effect $\Sigma \in \mathcal{M}_\psi$. In the following, we solve the dual likelihood equations for a fixed graph G. In order to consider full uncertainty over all possible causal structures, we subsequently need to cleverly optimize over the space of DAGs, see Section 4.1. In the case $\Sigma \in \mathcal{M}(G)$, we obtain

$$\begin{aligned}\frac{2}{n} \sup_{\Sigma \in \mathcal{M}(G)} \ell_n^{dual}(\Sigma|\widehat{\Sigma}^{-1}) &= \sup_{\Omega \in \mathcal{M}^{-1}(G)} -\log\det(2\pi\Omega) - \operatorname{tr}(\Omega^{-1}\hat{\Sigma}^{-1}) \\ &= \sup_{T \in \mathbb{R}^{-G}, \sigma^2 > 0} -d\log(2\pi\sigma^{-2}) - \sigma^2 \operatorname{tr}\big((I-T)^T(I-T)\hat{\Sigma}^{-1}\big),\end{aligned}$$

where $\mathbb{R}^{-G} := \{T \in \mathbb{R}^{d\times d} : t_{k,l} = 0 \text{ if } l \notin d(k)\}$. This immediately follows if we recapitulate

$$\Sigma = \sigma^2(I_d - B)^{-1}(I_d - B)^{-T} = \sigma^2(I-T)^T(I-T),$$

where $T \in \mathbb{R}^{-G}$ represents the negative total causal effect between variables. Maximizing over the equal error variance σ^2 then leads to a linear least squares problem with the optimal solution

$$\min_{T \in \mathbb{R}^G} \operatorname{tr}\big((I-T)^T(I-T)\hat{\Sigma}^{-1}\big) = \sum_{k=1}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)}. \tag{7}$$

In the second case of an additional constraint of a fixed size total causal effect $\mathcal{C}(i \to j) = \psi$, we can similarly first optimize over the equal error variance σ^2 to derive a linear least squares problem. That is, for $\Sigma \in \mathcal{M}_\psi(G)$, maximizing the dual likelihood then corresponds to the linear least squares problem

$$\min_{T \in \mathbb{R}^{-G,\psi}} \operatorname{tr}\big((I-T)^T(I-T)\hat{\Sigma}^{-1}\big).$$

Here, the search space is additionally restricted by the fixed size total causal effect $\mathcal{C}(i \to j) = \psi$ and given by $\mathbb{R}^{-G,\psi} := \{T \in \mathbb{R}^{d\times d} : t_{k,l} = 0 \text{ if } l \notin d(k), t_{i,j} = -\psi\}$. Solving this linear least squares problem with the constraint of one fixed parameter leads to the solution

$$\sum_{k \neq i}^{d} (\widehat{\Sigma}^{-1})_{k,k|d(k)} + (\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} + \psi^2(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}} + 2\psi(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}}. \tag{8}$$

Note that maximizing dual likelihood under constraints on total causal effects corresponds with the reciprocal property (4) to maximizing classical Gaussian likelihood under modified constraints on direct causal effects. Thus, while directly maximizing the classical Gaussian likelihood under constraints on the total causal effect is complicated due to the arising polynomial constraint on the parameters, we are able to obtain a closed-form solution for the maximum dual likelihood estimate since the constraint only pertains to one parameter.

4. Computation of Confidence Regions for Total Causal Effects

In the following section, we present important computational shortcuts as well as the results of a simulation study that analyzes the performance and computation times of our proposed confidence regions.

4.1. Computational Shortcuts

Our proposal of employing the dual likelihood to construct confidence regions for total causal effects resolves the need for numerical optimization routines of earlier methods and leads to a closed-form solution that significantly reduces computation times (see Figure 3). However, the bottleneck of the task itself is the superexponential growth of the number of possible causal structures with the number of nodes. Without any knowledge about the underlying causal relation, we have to consider all possible causal structures in order to make calibrated confidence statements that include the structure uncertainty. This, already at the scale of systems with 12 involved variables, entails accounting for more than 10^{26} possible graphs. Using combinatorial shortcuts that quickly reduce the search space of plausible causal structures is thus extremely important to compute confidence regions that solve this task.

First, we highlight that with our methodology, we only need to consider complete DAGs, thus reducing the search space to d factorial structures for d involved variables. Namely, we have to search through all permutations of d nodes, which then correspond to all possible topological orderings of the involved variables. Furthermore, we employ a branch and bound type search algorithm to quickly reduce the space of plausible causal orderings before performing the testing procedure for all possible causal effects $\psi \in \mathbb{R}$. The main idea of this procedure in order to immediately reject implausible causal orderings is the following. We quickly approximate the maximum dual likelihood under the alternative $\Omega \in \mathcal{M}^{-1}$ with the dual likelihood estimate (7) under the causal ordering obtained via recursively minimizing $(\widehat{\Sigma}^{-1})_{k,k|d(k)}$ starting from the sink node. Then, recursively starting from the sink node, we search through the space of causal orderings in reverse and reject a partial ordering if the unrestricted partial dual likelihood estimate (7) already exceeds the threshold of the alternative. This is possible because the dual likelihood estimate (7) collapses into subproblems which only depend on descendants. Moreover, looking at (8), the specific ordering among the descendants of i is not relevant for the subsequent testing procedure for the precise total causal effects $\psi \in \mathbb{R}$. Therefore, in our branch and bound type search algorithm, it suffices to proceed with the partial ordering that achieves the highest dual likelihood among all plausible orderings with the same set of descendants $d(i)$. More specifically, at each step of our algorithm, we reject all partial orderings that do not achieve the highest dual likelihood among all plausible partial orderings of the same set of nodes and do not include i.

Combined, all these computational shortcuts vastly improve computation times and lead to a bottom-up procedure that allows us to compute the proposed confidence region in a reasonable time for moderate dimensions, see Section 4.2. We emphasize that considering the fast superexponential growth of the number of possible DAGs with the number of nodes, rigorously accounting for structure uncertainty over all DAGs is an intrinsically difficult task already in moderate dimensions. Further, we highlight the interesting distinction to the `LRT`-method of Strieder and Drton (2023), that by employing dual likelihood, our method leads to a bottom-up procedure, starting from the sink node, while their method is a top-down procedure, starting from the source node.

4.2. Simulation Study

In the following simulation study, we compare the width of the proposed confidence region as well as the computation times to the LRT-method by Strieder and Drton (2023). For more simulation results, we refer the reader to the Appendix A. Our experiments are designed as follows. We generate 1000 synthetic data sets based on linear SCMs corresponding to randomly selected DAGs. The edge weights are sampled from a normal distribution $N(\beta, 0.1)$, and the error terms are sampled from a standard normal distribution. Then we compute the confidence regions for the total causal effect $\mathcal{C}(1 \to 2)$ with confidence level $\alpha = 0.05$. We repeat this procedure for a range of different average direct effect strengths β, sample sizes n, dimensions d, sparse or dense graphs, and for true non-zero total effects as well as for no total effect.

First, we note that the theoretical coverage guarantees of our proposed confidence regions are based on (conservative) asymptotic critical values. Nevertheless, the confidence regions achieve the desired coverage in all our simulation settings, even in low sample sizes. We explore this in Table 1 in the Appendix A, which shows the empirical coverage frequencies against the sample size. Furthermore, while the employed dual likelihood estimators are asymptotically efficient in the sense of having the same asymptotic variance as ordinary maximum likelihood estimates, differences in finite samples are possible. Thus, we compare the performance of our confidence regions to the existing maximum likelihood approach. In Figure 1, we compare the average width of the non-zero part of the confidence regions, reflecting the remaining uncertainty about the numerical size of the effect. We show the results for $d = 10$ and $\beta = 0.5$ against the sample size. The main observation is that the difference in performance between both methods seems negligible. Similar observations can be made for other performance measures, which we defer to the Appendix A. For example, we investigate how often the zero is included in the confidence regions when there is a true nonzero effect and observe no significant differences. Thus, both methods seem to be equally conclusive about the existence as well as the numerical size of the total causal effects. Further, the two methods also perform similarly under model misspecification of equal error variances and linearity.

However, the big advantage of the proposed dual likelihood method is the significantly reduced computation time. Employing the dual likelihood avoids time-consuming numerical optimization routines, and we immensely benefit from the closed-form solution calculated in Theorem 2. Figure 2 shows the computation times for $\beta = 0.5$ and sample size 1000 against the dimension. Our proposed confidence regions are faster to compute by a factor of up to 10^2, without any compromises in terms of information value. This is even more apparent in Figure 3, which compares the computation times for $\beta = 0.1$ and sample size 1000 against the dimension. In this setting, the generated data encompasses more structure uncertainty, given the lower average edge strength. Thus, the time difference between repeatedly applying numerical optimization algorithms versus directly computing the closed-form solution is more fatal, such that we are not able to compute the LRT-confidence regions in a reasonable time in dimensions beyond $d = 8$. In contrast, the computation times of our proposed method seem to vary less with the amount of structure uncertainty inherent in the data set.

Furthermore, we highlight the observation that dense systems generally lead to wider confidence regions compared to sparse systems. Intuitively, for dense graphs more paths contribute to the total causal effects on average, and thus, more uncertainty about the numerical size of the effect remains in the data. In contrast, data generated by sparse DAGs generally exhibits more uncertainty about the underlying causal structure, which explains the observed increased computation times.

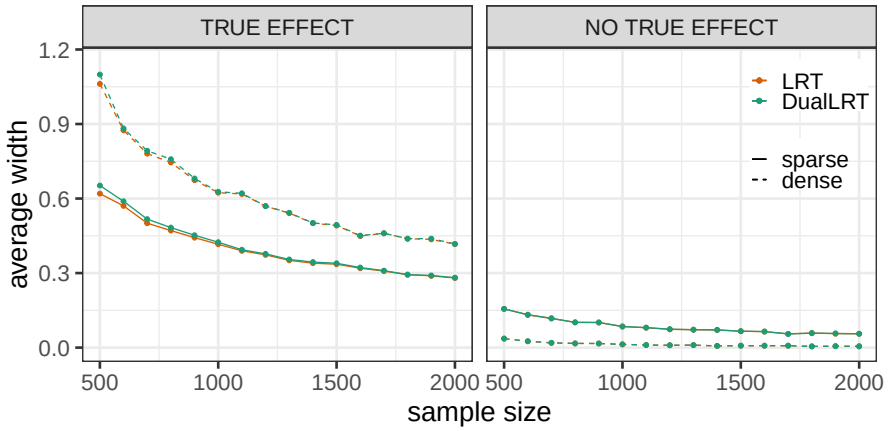


Figure 1: Mean width of 95%-confidence regions for the total causal effect in randomly generated 10-dim. DAGs (1000 replications).

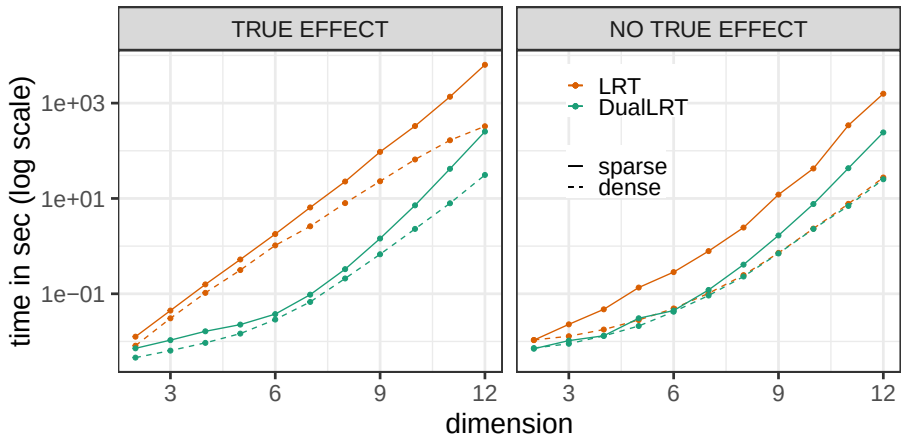


Figure 2: Mean computation times of 95%-confidence regions for the total causal effect in randomly generated DAGs with average edge strength $\beta = 0.5$ (1000 replications).

In summary, our simulation study validates that our proposed confidence regions based on the dual likelihood successfully pick up on the direction and the numerical size of the total causal effect while correctly quantifying the remaining uncertainty in structure as well as effect size. Furthermore, the computation times of our proposed confidence regions, computed with the proposed bottom-up procedure, are significantly lower than the competition. Here, we immensely benefit from the available closed-form solution due to employing dual likelihood. Especially under high structure uncertainty, avoiding repeated numerical optimization algorithms is crucial in order to compute the confidence regions in a reasonable time.

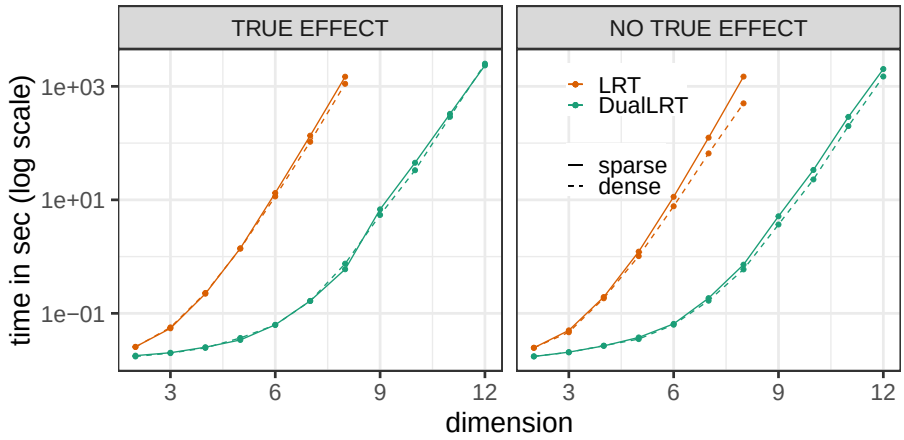


Figure 3: Mean computation times of 95%-confidence regions for the total causal effect in randomly generated DAGs with average edge strength $\beta = 0.1$ (1000 replications).

5. Conclusion

In this article, we address the important problem of rigorously accounting for uncertainty in causal inference when the causal structure itself has to be learned from data. We provide a closed-form solution for constructing confidence regions for the total causal effect that captures structure uncertainty as well as uncertainty about the numerical size of the effect. The confidence regions are developed for the concrete context of linear SCMs with equal error variances, where the underlying assumption of homoscedasticity across all interacting variables renders the causal structure identifiable. Thus, causal inference of the total causal effect without knowledge of the underlying structure is a well-defined task. An interesting conclusion of our work is that even in this specialized setting, we already observe high levels of uncertainty about causal directions and size of effects, which highlights the importance of methods that properly account for structure uncertainty in causal inference.

One thing to note is that our general approach of leveraging test inversions of joint tests for causal structure and effect size is generalizable to other settings. In particular, if one deals with parametric models for which the causal DAG is identifiable, then likelihood ratio tests may be deployed with the same stochastic approximation strategies as in this paper, and we anticipate that the matrix inversion interplay between direct and total effects that is behind our use of dual likelihood can be similarly exploited.

While unique structure identifiability is required in order to consistently resolve uncertainty about nonzero causal effects (e.g., when facing full Markov equivalence in linear Gaussian models without variance assumption, the confidence region would always include zero), our framework could in principle be adjusted to target all effects that correspond to graphs within a Markov equivalence class. However, in cases where the causal structure is merely identifiable up to some faithfulness-type assumption, more research needs to be done to fully understand which causal effects are entailed by a given (possibly unfaithful) distribution.

The observed reciprocal correspondence between dual likelihood with total effects and classical likelihood with direct effects could be exploited for other causal inference tasks. The idea is to leverage tools developed for direct effects and classical likelihood to causal inference results for total effects via the dual likelihood. In our simulation study, the differences in performance and explanatory power that stem from employing dual likelihood ratio tests instead of classical likelihood ratio tests are negligible and heavily outweighed by the significantly reduced computation times due to the available closed-form solution.

Finally, given the fast superexponential growth of the number of possible causal structures with the dimension of the system, it is essential to quickly reject implausible orderings and reduce the search space in order to compute confidence regions that account for uncertainty over all structures. The introduced computational shortcuts lead to a bottom-up procedure that, starting from sink nodes, recursively reduces the search space by immediately rejecting implausible partial orderings. In contrast, using the classical likelihood leads to a top-down procedure that starts with source nodes. Cleverly alternating between both procedures in a first step to reduce the search space might even further decrease the computation times of our proposed confidence regions.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 83818). Further, this work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content.

References

Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986.

George Casella and Roger L. Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Mathias Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1466–1475. PMLR, 09–11 Apr 2018.

Göran Kauermann. On a dualization of graphical Gaussian models. *Scand. J. Statist.*, 23(1):105–116, 1996.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. Models, reasoning, and inference.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2000.

David Strieder and Mathias Drton. Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *J. Causal Inference*, 11(1), 2023.

David Strieder, Tobias Freidling, Stefan Haffner, and Mathias Drton. Confidence in causal discovery with linear causal models. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. UAI'21*, volume 161, pages 1217–1226. PMLR, 27–30 Jul 2021.

Appendix A. Additional Simulation Results

In this section, we present additional results of our simulation study. The data generation process and general simulation setup are outlined in Section 4.2. Table 1 shows the observed coverage probabilities for 10-dimensional graphs. Both methods seem to be conservative and achieve the desired coverage frequency. Considering that we employ critical values based on an upper bound as well as the theory of intersection union tests for the arising testing problem, it is not surprising that the resulting confidence regions are conservative. Further, note that under high structure uncertainty, it is not possible to compute the competitor `LRT` in a reasonable time, which we denoted with NA.

As an additional performance measure, in Figure 4, we compare the proportions of times zero is included in the confidence regions when there is a true nonzero effect. The inclusion of zero in the confidence regions reflects the remaining uncertainty about the existence of the effect. We show the results for $d = 10$ and $\beta = 0.5$ against the sample size n as well as for $d = 10$ and sample size $n = 1000$ against the average direct effect strength β. In terms of performance, we observe no significant differences. However, we note again that for high structure uncertainty, which is the case for low average direct effect strength β, it is not possible to compute the competitor `LRT` in a reasonable time.

Furthermore, with practical applications in mind, Figure 5 investigates the robustness of the methods towards small deviations from equal error variances. Here, to generate data, we sample the error variances uniformly from $[1 - 0.5v, 1 + 0.5v]$, where v indicates the degree of deviation from homoscedasticity across all variables. We show the empirical coverage probabilities for $d = 10, \beta = 0.5$, and sample size $n = 1000$ against the degree of deviation v. Once again, we only observe negligible differences, and both methods indicate some robustness to small deviations from equal error variances.

Finally, to further explore aspects of model misspecification, we also investigate the robustness of the methods towards deviations from linear relations. We generate data as outlined before, however, with an increasing emphasis on an additional quadratic dependence structure via the relation $X_j = \sum_{i \in p(j)} \beta_i((1+3v)X_i + 0.02vX_i^2) + \varepsilon_j$, for all $j = 1, \dots, d$. Thus, v indicates the degree of deviation from linearity. The scaling ensures similar magnitudes of true causal effects across all investigated simulation settings. Note that for nonlinear relations, generally, the true total causal effect is not given by a single parameter but rather by a functional relation. However, when applying linear models it is natural to consider as a parameter of interest the slope parameter determining the best linear approximation within the true causal structure. In Figure 6, we consider this 'pseudo-linear' total effect and show the empirical coverage probabilities for $d = 10, \beta = 0.5$, and sample size $n = 1000$ against the degree of deviation v. We observe that even under nonlinearity, the methods pick up on the direction of the causal relation, and thus, for no causal effect, zero is always covered by the confidence intervals. Note that in the considered nonlinear additive equal variance setting, the causal ordering is still implied by ordering conditional variances. Further, the performance differences between both methods are negligible, and for nonzero effects, both methods indicate some robustness to small deviations from linearity.

Table 1: Empirical coverage of 95%-confidence regions for the total causal effect in randomly generated 10-dim. DAGs (1000 replications).

		TRUE EFFECT				NO TRUE EFFECT			
		SPARSE		DENSE		SPARSE		DENSE	
method	$n \backslash \beta$	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
DualLRT	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2000	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
LRT	500	NA	0.99	NA	1.00	NA	1.00	NA	1.00
	1000	NA	1.00	NA	1.00	NA	1.00	NA	1.00
	2000	NA	1.00	NA	1.00	NA	1.00	NA	1.00

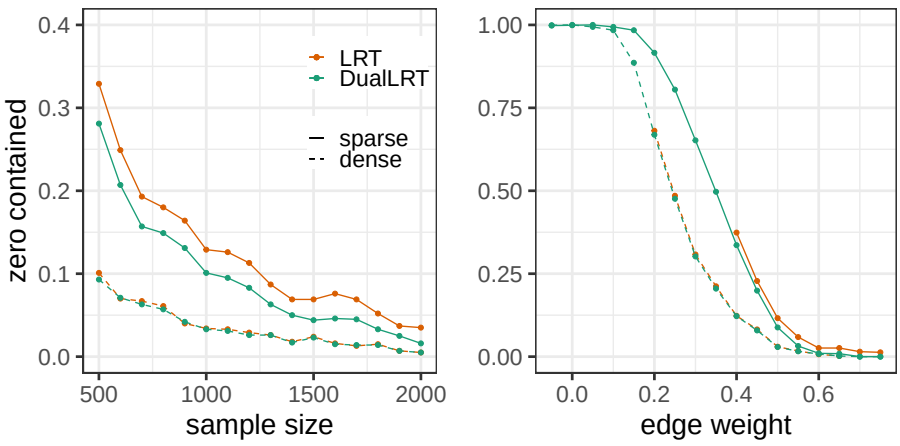


Figure 4: Proportion of times zero contained in 95%-confidence regions for the total causal effect in randomly generated 10-dim. DAGs with true non-zero effect (1000 replications).

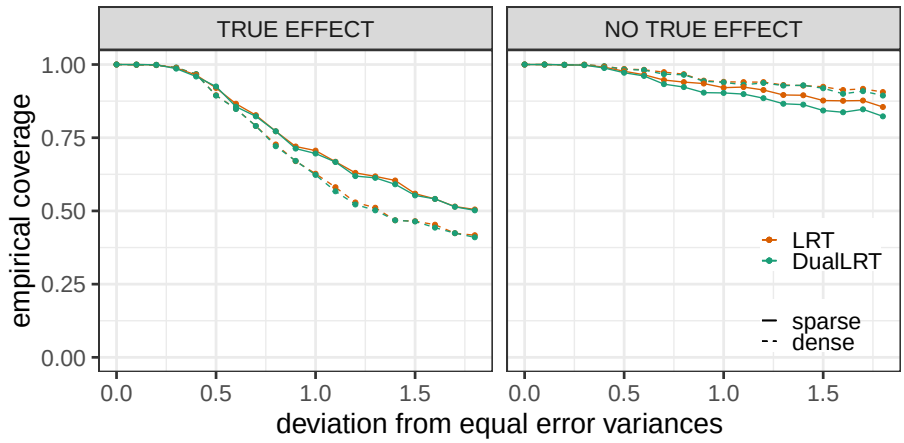


Figure 5: Empirical coverage of 95%-confidence regions for the total causal effect in randomly generated 10-dim. DAGs under departure from equal error variances (1000 replications).

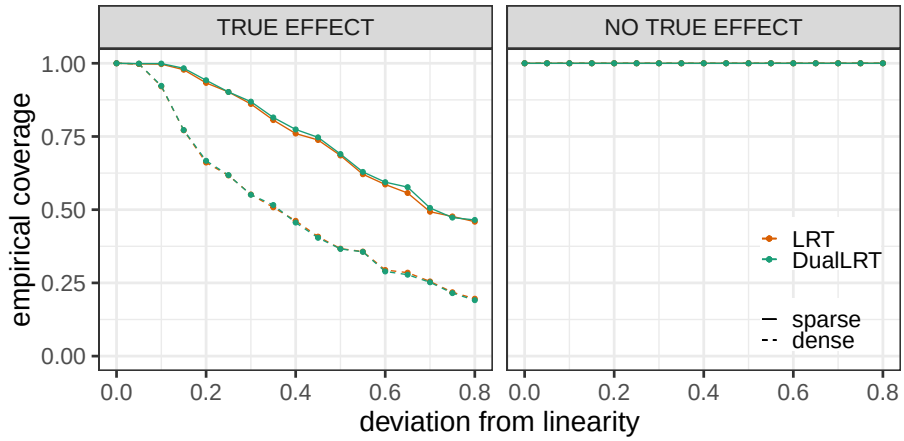


Figure 6: Empirical coverage of 95%-confidence regions for the total causal effect in randomly generated 10-dim. DAGs under departure from linearity (1000 replications).

Permission to include the article

PMLR Proceedings of Machine Learning Research

[edit]

Proceedings of Machine Learning Research

ISSN: 2640-3498

The Proceedings of Machine Learning Research is a series that publishes machine learning research papers presented at conferences and workshops. Each volume is separately titled and associated with a particular workshop or conference and will be published online on the PMLR web site. Authors retain copyright.

Editors

The Series Editor is Neil Lawrence. Please send proposals for new volumes under this series to us via e-mail: proceedings@mlr.press. Each proposal should include:

- A brief description of the event's scope and topics to be covered.
- A description of the review process for the proceedings.
- The names and short CVs (a few lines) of the proposed Proceedings Editors.

For frequently asked questions on publishing proceedings please see the FAQ.

For details on how to prepare a proceedings for publication please see the Specification.

Reissue Series

We have launched a "reissue series" for republishing volumes of papers that are no longer on line. Our first reissues are early editions of AISTATS, AISTATS 1995, AISTATS 1997, AISTATS 1999, AISTATS 2001, AISTATS 2003 and AISTATS 2005. To suggest volumes for the reissue

Permission to include the article

series contact the editors. You will need to provide the information provided in the Proceedings Specification to provide the volume.

Volume R6-draft Reissue of UAI 1998
Volume R0 Pre-proceedings of AISTATS 1995
Volume R5 Proceedings of AISTATS 2005
Volume R4 Proceedings of AISTATS 2003
Volume R3 Proceedings of AISTATS 2001
Volume R2 Proceedings of AISTATS 1999
Volume R1 Proceedings of AISTATS 1997

Proceedings

Volume 256 Proceedings of AutoML 2024
Volume 255 Proceedings of ML Meets Differential Equations
Volume 244 Proceedings of UAI 2024
Volume 230 Proceedings of COPA 2024
Volume 246 Proceedings of PGM 2024
Volume 249 Proceedings of the 1st ContinualAI Unconference
Volume 257 AI for Education Workshop 2024
Volume 228 Proceedings of NeuReps 2023
Volume 245 Proceedings of MLIC 2024
Volume 253 AABI 2024 Proceedings
Volume 248 Proceedings of CHIL 2024
Volume 235 Proceedings of ICML 2024
Volume 247 Proceedings of COLT 2024
Volume 242 Proceedings of L4DC 2024
Volume 241 Proceedings of LIDTA 2023
Volume 243 Proceedings of UniReps
Volume 226 Proceedings of the NeurIPS 2023 Gaze Meets ML Workshop
Volume 239 "I Can't Believe It's Not Better" NeurIPS Workshop 2023
Volume 238 Proceedings of AISTATS 2024
Volume 231 Proceedings of Learning on Graphs 2023
Volume 236 Proceedings of CLeaR 2024
Volume 240 Proceedings of MLCB 2024
Volume 237 Proceedings of ALT 2024
Volume 222 Proceedings of ACML 2023
Volume 227 Proceedings of Medical Imaging with Deep Learning 2023
Volume 233 Proceedings of the Northern Lights Deep Learning Workshop 2024
Volume 234 Proceedings of Conference on Parsimony and Learning 2024
Volume 219 Proceedings of Machine Learning for Healthcare 2023
Volume 225 Proceedings of the 3rd Machine Learning for Health Symposium

A.4 Identifying Total Causal Effects in Linear Models under Partial Homoscedasticity

Proceedings of Machine Learning Research 246:213–230, 2024 Probabilistic Graphical Models (PGM)

Identifying Total Causal Effects in Linear Models under Partial Homoscedasticity

David Strieder DAVID.STRIEDER@TUM.DE
Mathias Drton MATHIAS.DRTON@TUM.DE
Munich Center for Machine Learning (MCML) and Department of Mathematics,
TUM School of Computation, Information and Technology, Technical University of Munich

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

A fundamental challenge of scientific research is inferring causal relations based on observed data. One commonly used approach involves utilizing structural causal models that postulate noisy functional relations among interacting variables. A directed graph naturally represents these models and reflects the underlying causal structure. However, classical identifiability results suggest that, without conducting additional experiments, this causal graph can only be identified up to a Markov equivalence class of indistinguishable models. Recent research has shown that focusing on linear relations with equal error variances can enable the identification of the causal structure from mere observational data. Nonetheless, practitioners are often primarily interested in the effects of specific interventions, rendering the complete identification of the causal structure unnecessary. In this work, we investigate the extent to which less restrictive assumptions of partial homoscedasticity are sufficient for identifying the causal effects of interest. Furthermore, we construct mathematically rigorous confidence regions for total causal effects under structure uncertainty and explore the performance gain of relying on stricter error assumptions in a simulation study.

Keywords: Causal inference; linear structural causal models; equal error variances; structure uncertainty; confidence intervals.

1. Introduction

Graphical models are a valuable tool for studying statistical dependencies among complex systems of random variables (Lauritzen, 1996; Maathuis et al., 2019). In this paper, we study (directed) graphical models in their intuitive interpretation as structural causal models; edges indicate causal dependencies among interacting variables. Specifically, we consider statistical models specified by a recursive system of structural equations corresponding to a directed (acyclic) graph. The causal perspective views these structural equations as assignments rather than mathematical equations, reflecting the inherent asymmetry in cause-effect relationships.

A fundamental challenge of scientific research is inferring causal relations based on available data (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). In particular, knowledge of the underlying causal structure is crucial to correctly estimate the effect of external interventions or to reason about counterfactual questions. However, expert knowledge of the exact causal mechanism governing the data-generating process is missing in many applied settings. Furthermore, performing classical controlled experiments to study the causal dependencies is often also not feasible due to ethical concerns or cost considerations. The field

of causal discovery addresses this challenge by developing structure learning algorithms that estimate causal structures using only observed data (Drton and Maathuis, 2017).

A well-known result states that this causal structure, represented by a directed acyclic graph, can, at best, be identified up to a Markov equivalence class of indistinguishable models, unless one has access to interventional data or makes additional structural assumptions (see, e.g., Drton, 2018). Thus, an essential aspect of causal discovery is clarifying under which conditions the task is well-defined in that the causal quantities of interest can theoretically be identified from observational data alone. In this paper, we follow a line of research introduced by Peters and Bühlmann (2014) that focuses on linear relations and Gaussian errors with a common variance (see Section 2). In this setting, the underlying causal mechanisms are fully identifiable from observational data alone, as the causal order is implied by ordering conditional variances (Ghoshal and Honorio, 2018; Chen et al., 2019). However, practitioners are typically interested in specific causal effects, where identifying the entire causal structure seems excessive. This raises the question of whether less stringent modeling assumptions suffice to identify and estimate the effect of interest. Specifically, we investigate the identifiability of total causal effects in a setting that lies between the general case of unrestricted error variances and the fully homoscedastic case: We assume a shared variance only for the potential cause and the response of interest (Section 3).

Identifiability aside, an important aspect of (causal) inference is appropriately accounting for uncertainty to draw reliable conclusions from causal estimates. With only access to observational data, this includes not only classical statistical uncertainty about the numerical size of involved effects but also causal structure uncertainty, stemming from the data-driven structure learning approach and the equivalence of multiple plausible models. In Section 4, we follow an ansatz first introduced in Strieder et al. (2021) and explicitly construct rigorous confidence regions for causal effects under partial homoscedasticity of error variances as well as for the general case. We conclude our work with a simulation study in Section 5 that compares the confidence regions under different error restrictions and thus exemplifies the performance gained by relying on stricter structural assumptions.

2. Background

In this section, we review structural causal models and the definition of the total causal effect, which is the causal target of interest in this study. Further, we recapitulate the notions of identifiability and Markov equivalence of models.

2.1. Structural Causal Models and Total Causal Effects

Structural Causal Models (SCMs) are a common tool to model noisy functional relations among a set of interacting variables $\{X_i : i = 1, \dots, d\}$. In an SCM, each variable X_i is described as a function of a subset of other variables and a stochastic noise term ε_i. In this work, we focus on linear relations, that is, written in terms of random vectors, the data-generating process solves the linear equation system

$$\boldsymbol{X} = B\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_d)$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)$, and $B := [\beta_{j,i}]_{j,i=1}^{d}$ is a matrix that represents direct causal dependencies. We further assume that the errors are normal distributed, that is, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Omega)$, where $\Omega = \mathrm{diag}(\boldsymbol{\omega})$ with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$. Note that the errors are assumed to be uncorrelated, which implies causal sufficiency. The causal perspective arises when viewing these equation equations as making assignments rather than mathematical equations. Thus, the underlying causal structure of SCMs is naturally represented by a (minimal) directed graph G, where each node i in the graph corresponds to a variable X_i and edges $i \to j$ indicate direct causal dependencies, that is, $\beta_{j,i} \neq 0$. Throughout this work, we focus on acyclic causal relations, which entails that B is permutation similar to a lower triangular matrix and yields the unique solution $\boldsymbol{X} = (I_d - B)^{-1}\boldsymbol{\varepsilon}$ of the equation system (1), where I_d denotes the $d \times d$ identity matrix. Thus, $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where the covariance matrix is given by

$$\Sigma = (I_d - B)^{-1}\Omega(I_d - B)^{-T}.$$

We use the following notation and graphical concepts in the remainder of the article. We write $i <_G j$ if node i precedes node j in a causal ordering of the corresponding directed acyclic graph (DAG). If the DAG contains an edge from node i to node j, then node i is called a *parent* of node j, and we denote the set of all parents of node j with $p(j)$. Further, if the DAG contains a directed path from node i to node j, then node j is called a *descendant* of node i, and we denote the set of all descendants of node i with $d(i)$. Finally, we write $\Sigma_{j,i|p(i)}$ for the conditional covariance matrix, that is,

$$\Sigma_{j,i|p(i)} := \Sigma_{j,i} - \Sigma_{j,p(i)}(\Sigma_{p(i),p(i)})^{-1}\Sigma_{p(i),i}.$$

The target of interest in this work is the *total causal effect* $\mathcal{C}(i \to j)$, that is, the effect of an external intervention on X_i onto X_j. Within the framework of linear SCMs, the total causal effect is formally defined as the unit change in the expectation of X_j with respect to an intervention in X_i

$$\mathcal{C}(i \to j) := \frac{\mathrm{d}}{\mathrm{d}x_i}\mathbb{E}[X_j|\ \mathrm{do}(X_i = x_i)].$$

If the underlying DAG G is known, this parameter of interest can be expressed as a simple function of the covariance matrix $\varphi(G, \Sigma) := \Sigma_{j,i|p(i)}(\Sigma_{i,i|p(i)})^{-1}$, which is the regression coefficient of X_i when regressing X_j on $(X_i, X_{p(i)})$. However, difficulties arise in practice when the underlying causal structure is unknown, and thus, a valid adjustment set is unknown.

2.2. Markov Equivalence and Identifiability

An important question is, thus, given an observational distribution P_Σ from a set of possible distributions $\{P_\Sigma : \Sigma \in \mathcal{M}\}$, can one uniquely recover the (causal) target of interest $\varphi(G, \Sigma)$? Each DAG G may generate a (sub-)set of possible distributions from the model space $\mathcal{M}$ with entailing effect $\varphi(G, \Sigma)$, and thus, geometrically, this question is about the set of distributions that lie in the intersection of multiple plausible (causal) models.

Generally, an underlying linear Gaussian SCM naturally encodes a set of conditional independence constraints, given graphically by the set of d-separation relations in the underlying (true) DAG, which correspond algebraically to polynomial constraints on the set

of possible covariance matrices. Two DAGs are called *Markov equivalent* if they imply the same set of conditional independence constraints and thus generate the same set of distributions. However, these Markov equivalent DAGs do not necessarily entail the same causal effect of interest and, therefore, without additional (structural) assumptions, identification is at most possible up to a Markov equivalence class, that is, a set of potential values of the target of interest, each corresponding to a DAG within the same Markov equivalence class.

Additional faithfulness-type assumptions commonly guarantee at least being able to identify the Markov equivalence class by assuming that the given distribution does not satisfy additional conditional independence constraints, which are not implied by d-separation relations in the (true) DAG. We note that to recover our target of interest, the total causal effect, it is not necessary to fully identify the underlying DAG. We do not need to be able to distinguish a generating (minimal) DAG from its supergraphs to identify the target effect since all supergraphs yield a valid adjustment set for the total causal effect. To see that, note that every 'additional' parent of node i in a supergraph has to be conditionally independent of node i given the (minimal) parent set, which follows from d-separation in the minimal DAG. Thus, while assuming sparsity might improve computational aspects (by reducing the size of the model space), from a theoretical point of view, it suffices to only consider complete DAGs, that is, (causal) topological orderings of the involved variables, which includes all (sparse) DAGs as subgraphs.

In this article, we will focus our attention on the concept of *generic identifiability*. We call the causal parameter $\mathcal{C}(i \to j)$ generically identifiable if it can be recovered uniquely for almost all $\Sigma \in \mathcal{M}$ from a given observational distribution P_Σ out of a set of possible distributions $\{P_\Sigma : \Sigma \in \mathcal{M}\}$. Put differently, the set of covariance matrices corresponding to distributions for which the target of interest is not identifiable form a Lebesgue measure zero subset of our model space $\mathcal{M}$.

3. Homoscedasticity in SCMs

In this section, we investigate how introducing homoscedasticity constraints among the different error variances leads to identifiability of the (causal) parameters of interest. As mentioned before, in general, identification is at most possible up to a set of indistinguishable effects corresponding to Markov equivalent models. To be precise, in our setting under the assumption of an underlying linear Gaussian SCM, every possible distribution $\mathcal{N}(0, \Sigma)$ belongs to a covariance matrix from the set $\mathcal{M} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}(G)$, where $\mathcal{G}(d)$ is the set of all complete DAGs on d nodes and

$$\mathcal{M}(G) = \left\{ \Sigma \in \mathrm{PD}(d) : \exists B \in \mathbb{R}^G, \boldsymbol{\omega} \in \mathbb{R}^d \text{ with } \Sigma = (I_d - B)^{-1} \mathrm{diag}(\boldsymbol{\omega}) (I_d - B)^{-T} \right\} \quad (2)$$

is the (causal) model given by a DAG G, where $\mathbb{R}^G = \{B \in \mathbb{R}^{d \times d} : \beta_{j,i} = 0 \text{ if } j <_G i\}$. Note that without restrictions on the error variances, each set $\mathcal{M}(G)$ corresponds to the entire cone of positive definite matrices $\mathrm{PD}(d)$. In other words, all complete DAGs are Markov equivalent, and any complete DAG could have generated every possible given observational distribution. However, the total causal effect given by $\varphi(G, \Sigma)$ is not the same within all complete DAGs G that could have generated Σ. Thus, the total causal effect is not identifiable without any additional order or structure constraints.

Theorem 1 (see, e.g., Pearl (2009)) *The total causal effect $\mathcal{C}(i \to j)$ is not generically identifiable under the assumption of an underlying linear Gaussian SCM.*

Nevertheless, given an observational distribution P_Σ, we can recover a finite set of possible values of the total causal effects by enumerating all effects corresponding to all equivalent complete DAGs (Maathuis et al., 2009). We emphasize that while the number of permutations of d nodes is d factorial, many of these permutations lead to the same effects. For example, half of the permutations correspond to orderings with $j <_G i$ and, thus, entail a zero-sized effect. Similarly, the ordering among the parent set $p(i)$ or among the descendant set $d(i)$ is irrelevant to the specific value of the total causal effect. In fact, the size of the set of possible values of the total causal effect is upper bound by $2^{(d-2)} + 1$. However, under a general linear Gaussian SCM, this set of indistinguishable effects contains zero and non-zero values and, thus, is not informative about the existence or direction of a total causal effect without additional assumptions.

3.1. Equal Error Variance Constraint

A frequently used structural assumption that resolves this issue is introducing homoscedasticity among the error variances, that is, assuming $\omega_i = \omega_j$ for all $i, j = 1, \ldots, d$. This assumption restricts the set of possible distributions to a proper subset of the cone of positive definite matrices. Working with this assumption of equal error variances does not seem unreasonable for exploratory analyses in applications involving variables from similar domains or measurements obtained by similar machines and applications with concrete knowledge about external error variances.

From a causal perspective, this restriction refines the Markov equivalence classes and yields a straightforward criterion to identify the underlying causal ordering. Since the external error variances are equal, source nodes in the DAG have to be of minimal variance. To be precise, the homoscedasticity assumption corresponds to the algebraic constraint that the variance of each node conditioned on its parents has to be equal. Thus, for a given (complete) DAG, the set of possible distributions is given by

$$\mathcal{M}_{EV}(G) = \Big\{\Sigma \in \mathrm{PD}(d) : \exists \omega > 0 \text{ with } \omega = \Sigma_{k,k|p(k)} \quad \forall\, k = 1, \ldots, d\Big\}.$$

Note that any given Σ has a minimal diagonal element, which 'fixes' the value of the equal error variance ω and has to correspond to a source node in an underlying (minimal) DAG. If this minimal element is not unique, all nodes corresponding to minimal elements must be source nodes. By recursively conditioning on previous nodes and selecting nodes with minimal conditional variances, we obtain a unique one-to-one correspondence between a distribution Σ and a generating underlying minimal DAG. Thus, two causal models $\mathcal{M}_{EV}(G)$ only coincide at distributions generated by subgraphs that admit both complete graphs as valid causal orderings. However, since the underlying minimal DAG is unique, all valid causal orderings entail the same total causal effect. As mentioned, this follows due to d-separation criteria in the (unique) minimal DAG. Every 'additional' parent of node i in a supergraph has to be conditionally independent of node i given the (minimal) parent set. In other words, we obtain the following result.

Theorem 2 (Peters and Bühlmann (2014)) *The total causal effect $\mathcal{C}(i \to j)$ is globally identifiable under the assumption of an underlying linear Gaussian SCM with full homoscedasticity.*

We emphasize that this holds globally without additional faithfulness-type assumptions.

3.2. Identifiability under Partial Homoscedasticity

In this section, we answer the question of what happens between the general case of arbitrary error variances and the fully homoscedastic case with one common error variance. Specifically, we study the case of partial homoscedasticity; that is, we restrict $\omega_i = \omega_j$ only for potential cause X_i and response X_j of interest. This extends the possible applications of the framework towards cases where (only) the potential cause and response are from similar domains and cases with concrete knowledge about external error variances of (only) the potential cause and response. Algebraically, this assumption of partial homoscedasticity corresponds to the constraint that the variance of node i conditioned on its parents equals the variance of node j conditioned on its parents. Thus, for a given (complete) DAG, the causal model under partial homoscedasticity is given by

$$\mathcal{M}_{PEV}(G) = \left\{\Sigma \in \mathrm{PD}(d) : \Sigma_{i,i|p(i)} = \Sigma_{j,j|p(j)}\right\}.$$

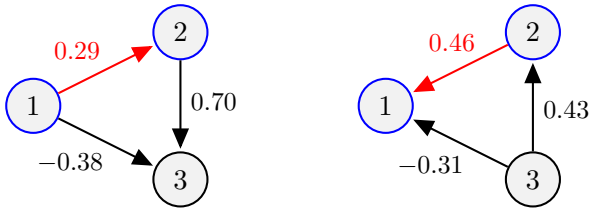
In previous work, Wu and Drton (2023) show that this assumption refines the Markov equivalence classes in such a way that two equivalent DAGs only generate the same causal model if and only if both postulate the same parent sets $p(i)$ and $p(j)$, respectively. This raises the question of whether this structural partial homoscedasticity assumption is enough to uniquely recover the total causal effect of interest $\mathcal{C}(i \to j)$ from a given observational distribution P_Σ. To answer that question, we study the set of possible values of the total causal effect entailed by a distribution that lies in the intersection of multiple plausible causal models.

Without loss of generality, we focus on the intersection between two (complete) graphs G^1 and G^2, that is, the intersection between two models $\mathcal{M}_{PEV}(G^1)$ and $\mathcal{M}_{PEV}(G^2)$. Both models can be parametrized similarly to (2), with $B \in \mathbb{R}^G$ and $\boldsymbol{\omega} \in \mathbb{R}^{d-1}$ under the partial homoscedasticity restriction $\omega_i = \omega_j$. Thus, as parameterized models, $\mathcal{M}_{PEV}(G^1)$ and $\mathcal{M}_{PEV}(G^2)$ are irreducible algebraic models (see, e.g., Cox et al., 2015). The intersection of two irreducible models is either of lower dimension or both models fully coincide. As already mentioned, Theorem 4.1 of Wu and Drton (2023) states that the two models fully coincide if and only if they imply the same parent sets for i and j, respectively, that is, $p_{G^1}(i) = p_{G^2}(i)$ and $p_{G^1}(j) = p_{G^2}(j)$. In this case however, both models would also entail the same total causal effect $\varphi(G^1, \Sigma) = \varphi(G^2, \Sigma) = \Sigma_{j,i|p(i)}(\Sigma_{i,i|p(i)})^{-1}$ since $p_{G^1}(i) = p_{G^2}(i)$. In the other case, if G^1 and G^2 do not imply the same parent sets for i and j, the intersection of $\mathcal{M}_{PEV}(G^1)$ and $\mathcal{M}_{PEV}(G^2)$ consequently has to be lower dimensional and, thus, a Lebesgue measure zero subset. Nevertheless, this intersection is not empty; see Example 1. Combining both cases, we obtain the following result.

Theorem 3 *The total causal effect $\mathcal{C}(i \to j)$ is generically identifiable under the assumption of an underlying linear Gaussian SCM with partial homoscedasticity among i and j.*

In other words, we can uniquely recover the total causal effect of interest for almost all $\Sigma \in \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}_{PEV}(G)$, that is, for almost all observational distributions from a linear Gaussian SCM under partial homoscedasticity. See Appendix A for an explicit example. Further, note that analogous additional faithfulness-type assumptions, i.e., assuming that the distribution satisfies only the equal error variance constraint between i and j as implied by the parent set of the true underlying DAG, would guarantee global identifiability. However, we emphasize that we can explicitly construct not only 'unfaithful' distributions in the intersection between two models that would imply effects of different magnitudes but also distributions in the intersection of two models that imply different directions for the considered effect. Thus, not even the causal ordering of the potential cause X_i and response X_j is globally identifiable without additional assumptions.

Example 1 *The two linear Gaussian SCMs with partial homoscedasticity $\omega_1 = \omega_2$, given by the underlying DAGs G^1 and G^2 generate the same observational distribution (with $\boldsymbol{\omega} = (1, 1, 1)$ and $\boldsymbol{\omega} = (0.81, 0.81, 1.52)$, respectively) but entail a different total causal effect $\mathcal{C}(1 \to 2)$, that is, $\varphi(G^1, \Sigma) = 0.29 \neq 0 = \varphi(G^2, \Sigma)$.*



(a) G^1 (b) G^2

We used computer algebra software to find a covariance matrix that satisfies both equal error variance constraints $\Sigma_{1,1} = \Sigma_{2,2|1}$ and $\Sigma_{1,1|\{2,3\}} = \Sigma_{2,2|3}$. Thus, such a covariance could have been generated by both causal orderings $1 <_G 2 <_G 3$ and $3 <_G 2 <_G 1$ and exemplifies that not even the direction of the effect is globally identifiable. Note that we rounded the displayed values; for more information see Appendix B.

4. Causal Inference under Partial Homoscedasticity

Theoretical identifiability aside, intuitively, stronger structural assumptions should lead to more informative inference results. Thus, in the following sections, we want to investigate how causal inference differs under these three error variance assumptions and how much performance can be gained by relying on stricter structural assumptions.

4.1. Maximum Dual Likelihood Estimation

Assume we have access to observational data in form of n independent copies $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ of a random vector $\boldsymbol{X}$ that follows a linear Gaussian SCM, given by an unknown DAG G. This section aims to estimate the set of indistinguishable total causal effects $\mathcal{C}(i \to j)$, corresponding to all equivalent models, similar to the IDA-framework (Maathuis et al., 2009). We employ dual likelihood methods (Brown, 1986; Kauermann, 1996) as an alternative to classical maximum likelihood estimation due to computational advantages (Strieder and

Drton, 2024). The main idea of dual likelihood is to swap the arguments in the usual classical minimization of the Kullback–Leibler divergence, which yields a different estimator and corresponds to maximizing

$$\ell_n^{dual}(\Sigma) := -\log\det(\Sigma^{-1}) - \operatorname{tr}(\Sigma\widehat{\Sigma}^{-1}),$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{l=1}^{n} \boldsymbol{X}^{(l)}(\boldsymbol{X}^{(l)})^T$ is the empirical covariance. We note that employing dual likelihood transfers the optimization problem from a covariance matrix to its inverse, simplifying constraints on total causal effects to facilitate the test inversion in Section 4.2.

Similar to the classical likelihood, $\widehat{\Sigma}$ is the unconstrained dual maximum likelihood estimate for $\Sigma \in \mathcal{M}$. Thus, in the general case without any restrictions on the error variances, the total causal effect can be estimated via $\{\varphi(G, \widehat{\Sigma}) \in \mathbb{R} : G \in \mathcal{G}(d)\}$, which is the set of all causal effects implied by all indistinguishable models that could have generated the maximum (dual) likelihood estimate $\widehat{\Sigma}$. Note that many complete DAGs necessarily entail the same effect as they imply the same adjustment set $p(i)$. Thus, to calculate this proposed estimate, we only need to compute 2^{d-2} values, that is, enumerate all distinct parent sets $p(i)$ that do not include j, and further append zero. Furthermore, note that the adjustment sets implied by complete DAGs are not necessarily efficient if the true underlying DAG is not complete (Henckel et al., 2022).

Under the additional structural assumption of partial homoscedasticity between potential cause i and response j, that is, assuming $\omega_i = \omega_j$, the maximum likelihood estimation is a complex combinatorial optimization problem. However, for a fixed DAG G, dual maximum likelihood estimation is equivalent to solving a sequence of linear regression problems for each node regressed on its descendants. Thus, we obtain the optimal value

$$\sup_{\Sigma\in\mathcal{M}_{PEV}(G)} \ell_n^{dual}(\Sigma) = -\sum_{k\neq i,j}^{d} \log\big((\widehat{\Sigma}^{-1})_{k,k|d(k)}\big) - 2\log\big(\tfrac{1}{2}\big((\widehat{\Sigma}^{-1})_{i,i|d(i)} + (\widehat{\Sigma}^{-1})_{j,j|d(j)}\big)\big) - d.$$

The corresponding total causal effect is the regression coefficient of node j when regressing node i on its descendants, that is, $\tau_j\big((\widehat{\Sigma}^{-1})_{d(i),d(i)}\big)^{-1}(\widehat{\Sigma}^{-1})_{d(i),i}\big)$, where τ_j projects the $|d(i)|$-dimensional vector onto the component corresponding to j if $j \in d(i)$ and zero otherwise. Subsequently, maximizing over the space of complete DAGs $G \in \mathcal{G}$ yields the dual maximum likelihood estimate under partial homoscedasticity. We emphasize that this dual maximum likelihood estimate may lie in the intersection of multiple causal models implying different total causal effects. Thus, analogously to the general case, we propose to estimate the total causal effect under structure uncertainty as the set of all causal effects corresponding to DAGs that achieve this optimal value. Using combinatorial shortcuts, we can compute this proposed estimate without checking all d factorial permutations, i.e., all complete DAGs implying the same parent sets $p(i)$ and $p(j)$ are equivalent and entail the same causal effect of interest. Thus, it suffices to compute and compare two times 3^{d-2} optimal values.

Further, under full homoscedasticity among all error variances and for a fixed DAG G, we obtain the optimal value

$$\sup_{\Sigma\in\mathcal{M}_{EV}(G)} \ell_n^{dual}(\Sigma) = -d\log\big(\tfrac{1}{d}\sum_{k=1}^{d}(\widehat{\Sigma}^{-1})_{k,k|d(k)}\big) - d$$

with corresponding total causal effect $\tau_j\big(\big((\widehat{\Sigma}^{-1})_{d(i),d(i)}\big)^{-1}(\widehat{\Sigma}^{-1})_{d(i),i}\big)$. Analogously, maximizing over the space of complete DAGs $G \in \mathcal{G}$ then yields the dual maximum likelihood estimate. This dual maximum likelihood estimate may also lie in the intersection of multiple causal models. However, in this case, this intersection has to correspond to distributions generated by a lower-dimensional subgraph. Thus, all complete DAGs that achieve this optimal value are valid causal orderings of the true underlying subgraph and entail the same total causal effect. We note that leveraging combinatorial shortcuts to calculate this total causal effect estimate is more complex, but evaluating all permutations can be avoided with ideas similar to Silander and Myllymäki (2006).

4.2. Confidence Intervals for Total Causal Effects

In order to draw reliable conclusions from causal estimates under structure uncertainty, we further want to construct a confidence interval, that is, in the technical sense, a region that (at least asymptotically) guarantees a desired frequentist coverage probability of the true total causal effect of interest. Recent proposals that rigorously account for uncertainty in causal inference rely on test-inversion approaches to construct confidence sets for causal quantities of interest (Strieder and Drton, 2023; Wang et al., 2023). Strieder and Drton (2023) explicitly construct a confidence region for the total causal effect by inverting joint tests for the underlying causal structure and effect size in the concrete setting of an underlying linear Gaussian SCM with equal error variances. We will take up their methodology and extend it to the partial homoscedastic and general cases. Again, we emphasize that the total causal effect is not globally identifiable when departing from the complete homoscedastic case. Thus, structure uncertainty is not only aleatoric due to the data-driven model choice but also epistemic due to multiple models being indistinguishable. We aim to construct reliable confidence intervals that consider all sources of uncertainty and cover all indistinguishable effects with the desired frequency.

In the general setting, without assuming equal error variance, the ansatz of inverting joint tests for causal structure and effect size leads to the following statistical testing problem for all $\psi \in \mathbb{R}$

$$H_0^{(\psi)} : \Sigma \in \mathcal{M}^{\psi} \quad \text{against} \quad H_1 : \Sigma \in \mathcal{M} \backslash \mathcal{M}^{\psi},$$

where $\mathcal{M}^{\psi} := \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}^{\psi}(G)$ with $\mathcal{M}^{\psi}(G) := \{\Sigma \in \mathrm{PD}(d) : \varphi(G, \Sigma) = \psi\}$. We solve the testing problem using the dual likelihood ratio test and the theory of intersection union tests (see, e.g., Casella and Berger, 1990) to obtain the following result.

Theorem 4 *Let $\alpha \in (0,1)$. For all $G \in \mathcal{G}(d)$ with $i <_G j$ define*

$$D(G) := (\widehat{\Sigma}^{-1})^2_{j,i|d(i)\setminus\{j\}} - (\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}\Big((\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} - (\widehat{\Sigma}^{-1})_{i,i|d(i)} \exp\big(\tfrac{\chi^2_{1,1-\alpha}}{n}\big)\Big).$$

Then an asymptotic $(1-\alpha)$-confidence set for the total causal effect $\mathcal{C}(i \to j)$ under the assumption of an underlying linear Gaussian SCM is given by

$$C := \bigcup_{G \in \mathcal{G}(d) \,:\, D(G) \geq 0} \big[L(G), U(G)\big] \bigcup \{0\},$$

where the closed-form solution for the limits is given by

$$L(G) := \frac{-(\widehat{\Sigma}^{-1})_{j,i|d(i)\setminus\{j\}} - \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}, \qquad U(G) := \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}.$$

The proof is given in Appendix C.

Without any additional structural assumption on the error variances, the total causal effect is not identifiable, and all complete DAGs could have generated every possible observational distribution. While the confidence regions still contain information and restrict the real line of all possible effects, they will always include the possibility of no effect.

Inducing additional structure by assuming equal error variances among the potential cause i and response j (potentially) resolves this ambiguity and leads to confidence regions via inverting the following statistical testing problem for all $\psi \in \mathbb{R}$

$$H_0^{(\psi)} : \Sigma \in \mathcal{M}_{PEV}^{\psi} \quad \text{against} \quad H_1 : \Sigma \in \mathcal{M}_{PEV} \setminus \mathcal{M}_{PEV}^{\psi},$$

where $\mathcal{M}_{PEV}^{\psi} := \bigcup_{G\in\mathcal{G}(d)} \mathcal{M}_{PEV}^{\psi}(G)$ with $\mathcal{M}_{PEV}^{\psi}(G) := \{\Sigma \in \mathcal{M}_{PEV}(G) : \varphi(G,\Sigma) = \psi\}$. Analogously, solving this testing problem with dual likelihood ratio tests and the theory of intersection union tests yields the following result.

Theorem 5 *Let $\alpha \in (0,1)$. Define $K := \min_{G\in\mathcal{G}(d)} K(G)$ and $Z := \min_{G\in\mathcal{G}(d):j<_G i} K(G)$, where*

$$K(G) := \prod_{k\neq i,j}^{d} \sqrt{(\widehat{\Sigma}^{-1})_{k,k|d(k)}}\Big((\widehat{\Sigma}^{-1})_{i,i|d(i)} + (\widehat{\Sigma}^{-1})_{j,j|d(j)}\Big).$$

Furthermore, for all $G \in \mathcal{G}(d)$ with $i <_G j$ define

$$\begin{aligned} D(G) :=& (\widehat{\Sigma}^{-1})^2_{j,i|d(i)\setminus\{j\}} \\ &- (\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}\left((\widehat{\Sigma}^{-1})_{j,j|d(j)} + (\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} - \frac{K\exp\left(\frac{\chi^2_{2,1-\alpha}}{2n}\right)}{\prod_{k\neq i,j}^{d}\sqrt{(\widehat{\Sigma}^{-1})_{k,k|d(k)}}}\right). \end{aligned}$$

Then an asymptotic $(1-\alpha)$-confidence set for the total causal effect $\mathcal{C}(i \to j)$ under the assumption of an underlying linear Gaussian SCM with (partially) equal error variance among i and j is given by

$$C := \bigcup_{G\in\mathcal{G}(d)\,:\,D(G)\geq 0} \big[L(G), U(G)\big] \bigcup \big\{0 : Z \leq K\exp\left(\tfrac{\chi^2_{1,1-\alpha}}{2n}\right)\big\},$$

where the closed-form solution for the limits is given by

$$L(G) := \frac{-(\widehat{\Sigma}^{-1})_{j,i|d(i)\setminus\{j\}} - \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}, \qquad U(G) := \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}.$$

The proof is given in Appendix D.

We highlight the overarching structure of the confidence regions, which consist of intervals that each represent a plausible causal ordering with $i <_G j$ and possibly a zero that

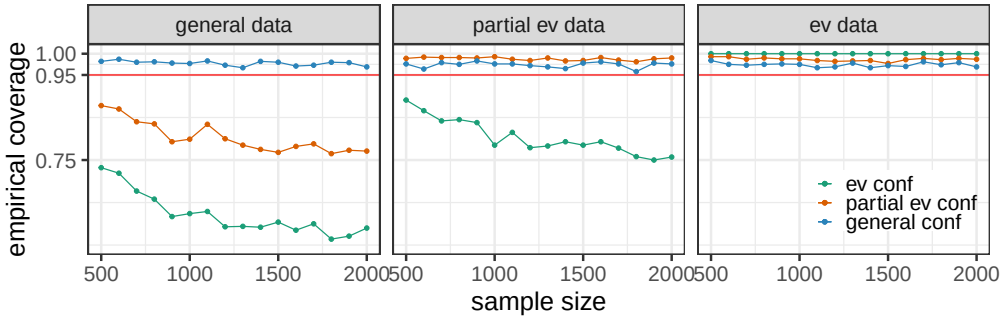


Figure 1: Empirical coverage of 95%-confidence intervals for the total causal effect.

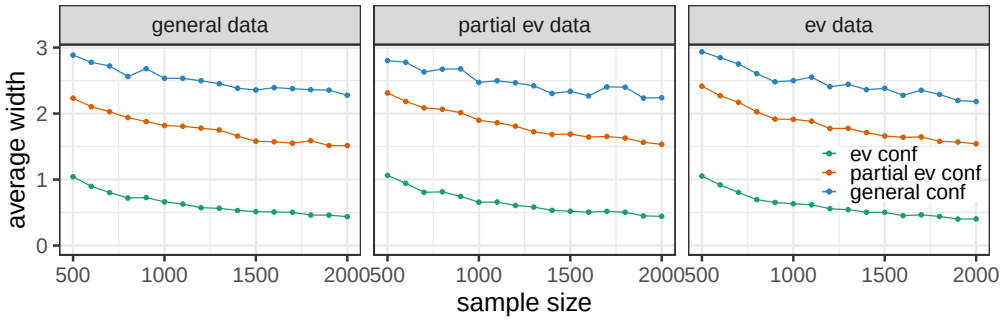


Figure 2: Mean width of 95%-confidence intervals for the total causal effect.

reflects remaining uncertainty about the direction of the effect. It is worth noting that any causal ordering with $i <_G j$ immediately implies no effect. Therefore, informally, the hypothesis of no effect for a DAG with $i <_G j$ is larger than that of a specific effect ψ in a DAG with $j <_G i$. This intuitively explains the lower degrees of freedom in the corresponding critical value.

5. Simulation Study

In this section, we present the results of a simulation study that compares the performance of our proposed confidence regions (`general conf` with no variance assumption and `partial ev conf` assuming partial homoscedasticity) and the confidence region proposed in Strieder and Drton (2024) under the (strict) assumption of all error variances being equal (`ev conf`). Thus, our simulations contrast the potential information gain of relying on stricter structural error variance assumptions. Our experiments were designed following Strieder and Drton (2024) with synthetic data from 10−dimensional sparse DAGs with true non-zero total causal effects under the three error variance regimes. For details on the data generation process and additional simulation results for data with no true effect, see Appendix E.

Figure 1 reports the empirical coverage frequencies. As expected, all methods achieve the desired coverage frequency of 0.95 under their respective error variance regime (and more restrictive cases); however, they fail under the more general regimes. To investigate

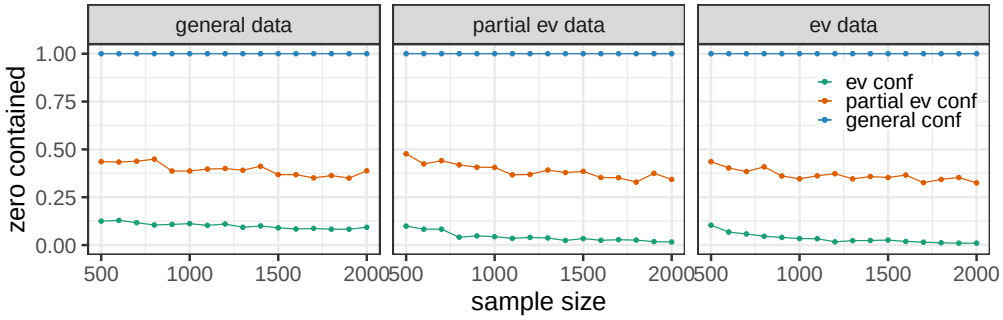


Figure 3: Proportion of times zero contained in 95%-confidence intervals for the total causal effect.

the potential upside of relying on stricter variance assumptions, we investigate how conclusive the proposed confidence regions are about the existence and size of causal effects in the form of the average width of the non-zero part of the confidence sets (Figure 2) as well as proportions of times zero is contained in the confidence sets (Figure 3). Note that the methods do not adapt to the complexity of the data; that is, each method's width and zero proportions do not vary across the different data generation settings. Furthermore, without any error variance assumption, the confidence region is never conclusive about the existence of an effect, highlighting that structural assumptions are necessary to obtain decisive results. Nevertheless, the validity of the confidence statements relies on the correctness of the error variance assumption, which thus, must be carefully chosen to make calibrated and informative confidence statements about total causal effects under structure uncertainty.

6. Conclusion

An essential aspect of causal discovery is clarifying under which conditions it is theoretically possible to identify causal quantities from observational data alone. In this paper, we investigate how different assumptions on the error variances lead to the (non-)identifiability of total causal effects of interest. While the classical setting with arbitrary error variances leads to a (finite) set of indistinguishable potential effects, the recently studied equal error variance setup with one common variance entails global identifiability. We demonstrate that in between those cases, assuming partial equal variance between the potential cause and response of interest is sufficient to generically identify the total causal effect.

Furthermore, appropriately accounting for uncertainty is crucial to draw reliable conclusions from causal estimates. With only access to observational data, this includes classical statistical uncertainty about the numerical size of involved effects and causal structure uncertainty, stemming from a data-driven structure learning approach as well as the equivalence of multiple plausible models. Via a test inversion approach, we construct rigorous confidence regions that account for all types of uncertainty and show that if practitioners are willing to rely on stricter error variance assumptions, then the corresponding confidence regions are more informative about the existence and size of total causal effects.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 83818). Further, this work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content.

References

L. D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986.

G. Casella and R. L. Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

W. Chen, M. Drton, and Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

D. A. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015.

M. Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

M. Drton. Algebraic problems in structural equation modeling. In *The 50th anniversary of Gröbner bases*, volume 77 of *Adv. Stud. Pure Math.*, pages 35–86. Math. Soc. Japan, Tokyo, 2018.

M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475. PMLR, 2018.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84(2):579–599, 2022.

G. Kauermann. On a dualization of graphical Gaussian models. *Scand. J. Statist.*, 23(1): 105–116, 1996.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164, 2009.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, UAI*. AUAI Press, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2000.

D. Strieder and M. Drton. Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *J. Causal Inference*, 11(1), 2023.

D. Strieder and M. Drton. Dual likelihood for causal inference under structure uncertainty. In *Causal Learning and Reasoning, CLeaR*, volume 236 of *Proceedings of Machine Learning Research*, pages 1–17. PMLR, 2024.

D. Strieder, T. Freidling, S. Haffner, and M. Drton. Confidence in causal discovery with linear causal models. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence, UAI*, volume 161 of *Proceedings of Machine Learning Research*, pages 1217–1226. AUAI Press, 2021.

Y. S. Wang, M. Kolar, and M. Drton. Confidence sets for causal orderings. *arXiv preprint arXiv:2305.14506*, 2023.

J. Wu and M. Drton. Partial homoscedasticity in causal discovery with linear models. *IEEE J. Sel. Areas Inf. Theory*, 4:639–650, 2023.

Appendix A. Identifiable Example

Under the assumption of an underlying linear Gaussian SCM with partial homoscedasticity $\omega_1 = \omega_2$, that is, $\Sigma_{1,1|p(1)} = \Sigma_{2,2|p(2)}$, the observational distribution P_Σ, given by

$$\Sigma = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 5.5 \end{pmatrix},$$

uniquely entails the total causal effect $\mathcal{C}(1 \to 2) = 1$. This follows from the observation that $\Sigma_{1,1|p(1)} = \Sigma_{2,2|p(2)}$ if and only if $p(1)$ equals the empty set and $p(2) = \{1\}$. Thus, $\mathcal{C}(1 \to 2) = \Sigma_{1,2}(\Sigma_{1,1})^{-1} = 1$.

Appendix B. Non-identifiable Example

In Example 1, we used computer algebra software to find a covariance matrix that satisfies both equal error variance constraints $\Sigma_{1,1} = \Sigma_{2,2|1}$ and $\Sigma_{1,1|\{2,3\}} = \Sigma_{2,2|3}$. The exact entries of the found covariance matrix are algebraic numbers, given as complex root expressions. Thus, we display in the following the found (exact) example up to a precision of 15 decimal digits, that is, the observational distribution P_Σ, given by

$$\Sigma = \begin{pmatrix} 1.00000000000000 & 0.294584930358565 & -0.176750958215139 \\ 0.294584930358565 & 1.08678028119436 & 0.648086846788844 \\ -0.176750958215139 & 0.648086846788844 & 1.52145794696742 \end{pmatrix},$$

could have been generate by two different linear Gaussian SCMs $\boldsymbol{X} = B\boldsymbol{X} + \boldsymbol{\varepsilon}$, that is, either

$$B_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.294584930358565 & 0 & 0 \\ -0.383006074698015 & 0.700155015505460 & 0 \end{pmatrix}$$

and $\mathrm{Var}(\boldsymbol{\varepsilon_1}) = \mathrm{diag}(1, 1, 1)$, represented by G_1 in Example 1, or

$$B_2 = \begin{pmatrix} 0 & 0.456230601077430 & -0.310510067542541 \\ 0 & 0 & 0.425964350891600 \\ 0 & 0 & 0 \end{pmatrix}$$

and $\mathrm{Var}(\boldsymbol{\varepsilon_2}) = \mathrm{diag}(0.810718388180567, 0.810718388180567, 1.52145794696742)$, given by G_2.

Appendix C. Proof of Theorem 4

Proof Let $\psi \in \mathbb{R}$ and $\Sigma \in \mathcal{M}^\psi$. Since $\mathcal{M}^\psi = \bigcup_{G \in \mathcal{G}(d)} \mathcal{M}^\psi(G)$ there exists a complete graph G such that $\Sigma \in \mathcal{M}^\psi(G)$. Without loss of generality, we assume $i <_G j$; otherwise, ψ has to be zero, which is always contained in the proposed confidence regions by construction.

The model space $\mathcal{M}$ equals the set of all positive definite matrices $\mathrm{PD}(d)$ and can be identified with an open subspace in $\mathbb{R}^{\frac{1}{2}(d^2+d)}$. Each (single) hypothesis of the union $\Sigma \in \mathcal{M}^\psi(G)$ then defines a $\frac{1}{2}(d^2 + d) - 1$-dimensional submanifold of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$. Therefore,

the asymptotic distribution of the dual likelihood ratio test statistic under every single hypothesis is given by χ^2_1 and, thus,

$$P_\Sigma\left(n\left(\sup_{\Sigma\in\mathcal{M}}\ell_n^{dual}(\Sigma)-\sup_{\Sigma\in\mathcal{M}^\psi(G)}\ell_n^{dual}(\Sigma)\right)>\chi^2_{1,1-\alpha}\right)\to\alpha.$$

For details on this classical result about the limit distribution of likelihood ratio test for submanifolds and the transformation to dual likelihood, we refer to Drton (2009) and Strieder and Drton (2024). Furthermore, dual likelihood estimation via solving a sequence of linear regression problems (see Section 4.1) yields the following optimal values

$$\begin{aligned}\sup_{\Sigma\in\mathcal{M}}\ell_n^{dual}(\Sigma)&=\sup_{\Sigma\in\mathcal{M}(G)}\ell_n^{dual}(\Sigma)=-\sum_{k=1}^{d}\log\big((\widehat{\Sigma}^{-1})_{k,k|d(k)}\big)-d,\\ \sup_{\Sigma\in\mathcal{M}^\psi(G)}\ell_n^{dual}(\Sigma)&=-\sum_{k\neq i}^{d}\log\big((\widehat{\Sigma}^{-1})_{k,k|d(k)}\big)\\ &\quad-\log\big((\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}}+\psi^2(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}+2\psi(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}}\big)-d.\end{aligned}$$

Plugging in these dual likelihood estimates, we view

$$n\left(\sup_{\Sigma\in\mathcal{M}}\ell_n^{dual}(\Sigma)-\sup_{\Sigma\in\mathcal{M}^\psi(G)}\ell_n^{dual}(\Sigma)\right)-\chi^2_{1,1-\alpha}$$

as a strictly convex quadratic polynomial in ψ, which has real roots if $D(G)\geq 0$, where

$$D(G):=(\widehat{\Sigma}^{-1})^2_{j,i|d(i)\setminus\{j\}}-(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}\Big((\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}}-(\widehat{\Sigma}^{-1})_{i,i|d(i)}\exp\big(\tfrac{\chi^2_{1,1-\alpha}}{n}\big)\Big).$$

Thus, the inequality $n\big(\sup_{\Sigma\in\mathcal{M}}\ell_n^{dual}(\Sigma)-\sup_{\Sigma\in\mathcal{M}^\psi(G)}\ell_n^{dual}(\Sigma)\big)>\chi^2_{1,1-\alpha}$ holds if and only if

$$\psi\in\left[\frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}}-\sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}},\frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}}+\sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}\right].$$

The test inversion approach and intersection union theory complete the proof. ■

Appendix D. Proof of Theorem 5

Proof Let $\psi\in\mathbb{R}$ and $\Sigma\in\mathcal{M}^\psi_{PEV}$. Since $\mathcal{M}^\psi_{PEV}=\bigcup_{G\in\mathcal{G}(d)}\mathcal{M}^\psi_{PEV}(G)$ there exists a complete graph G, such that $\Sigma\in\mathcal{M}^\psi_{PEV}(G)$. We assume $i<_G j$; the case $j<_G i$ can be constructed analogously.

The dual likelihood ratio test statistic is upper bounded by the larger test statistic of testing against the entire cone of positive definite matrices, that is,

$$n\left(\sup_{\Sigma\in\mathcal{M}_{PEV}}\ell_n^{dual}(\Sigma)-\sup_{\Sigma\in\mathcal{M}^\psi_{PEV}(G)}\ell_n^{dual}(\Sigma)\right)\leq n\left(\sup_{\Sigma\in\mathrm{PD}(d)}\ell_n^{dual}(\Sigma)-\sup_{\Sigma\in\mathcal{M}^\psi_{PEV}(G)}\ell_n^{dual}(\Sigma)\right).$$

Within the set of all positive definite matrices $\mathrm{PD}(d)$, each (single) hypothesis of the union $\Sigma \in \mathcal{M}^{\psi}_{PEV}(G)$ defines a $\frac{1}{2}(d^2+d)-2$-dimensional submanifold of $\mathbb{R}^{\frac{1}{2}(d^2+d)}$. Therefore, the asymptotic distribution of this stochastic upper bound of the dual likelihood ratio test statistic is given by χ^2_2 and, thus, with the conservative critical value $\chi^2_{2,1-\alpha}$ from the stochastic upper bound, we obtain

$$P_\Sigma\left(n\left(\sup_{\Sigma\in\mathcal{M}_{PEV}} \ell_n^{dual}(\Sigma) - \sup_{\Sigma\in\mathcal{M}^{\psi}_{PEV}(G)} \ell_n^{dual}(\Sigma)\right) > \chi^2_{2,1-\alpha}\right)$$
$$\leq P_\Sigma\left(n\left(\sup_{\Sigma\in\mathrm{PD}(d)} \ell_n^{dual}(\Sigma) - \sup_{\Sigma\in\mathcal{M}^{\psi}_{PEV}(G)} \ell_n^{dual}(\Sigma)\right) > \chi^2_{2,1-\alpha}\right) \to \alpha.$$

For details on the limit distribution of the likelihood ratio test for submanifolds and the transformation to dual likelihood, we refer to Drton (2009) and Strieder and Drton (2024). Furthermore, the dual likelihood estimation via solving a sequence of linear regression problems (see Section 4.1) yields

$$\sup_{\Sigma\in\mathcal{M}_{PEV}} \ell_n^{dual}(\Sigma) = -2\log(\tfrac{K}{2}) - d.$$

with

$$K := \min_{G\in\mathcal{G}(d)} \prod_{k\neq i,j}^{d} \sqrt{(\widehat{\Sigma}^{-1})_{k,k|d(k)}}\Big((\widehat{\Sigma}^{-1})_{i,i|d(i)} + (\widehat{\Sigma}^{-1})_{j,j|d(j)}\Big),$$

as well as

$$\sup_{\Sigma\in\mathcal{M}^{\psi}_{PEV}(G)} \ell_n^{dual}(\Sigma) = -\sum_{k\neq i,j}^{d} \log\left((\widehat{\Sigma}^{-1})_{k,k|d(k)}\right) - d - 2\log\Big(\tfrac{1}{2}\big((\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}}$$
$$+ \psi^2(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}} + 2\psi(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + (\widehat{\Sigma}^{-1})_{j,j|d(j)}\big)\Big).$$

Plugging in these dual likelihood estimates, we view

$$n\left(\sup_{\Sigma\in\mathcal{M}_{PEV}} \ell_n^{dual}(\Sigma) - \sup_{\Sigma\in\mathcal{M}^{\psi}_{PEV}(G)} \ell_n^{dual}(\Sigma)\right) - \chi^2_{2,1-\alpha}$$

as a strictly convex quadratic polynomial in ψ, which has real roots if $D(G) \geq 0$, where

$$D(G) := (\widehat{\Sigma}^{-1})^2_{j,i|d(i)\setminus\{j\}}$$
$$- (\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}\left((\widehat{\Sigma}^{-1})_{j,j|d(j)} + (\widehat{\Sigma}^{-1})_{i,i|d(i)\setminus\{j\}} - \frac{K\exp\left(\frac{\chi^2_{2,1-\alpha}}{2n}\right)}{\prod_{k\neq i,j}^{d}\sqrt{(\widehat{\Sigma}^{-1})_{k,k|d(k)}}}\right)$$

Thus, the inequality $n\big(\sup_{\Sigma\in\mathcal{M}_{PEV}} \ell_n^{dual}(\Sigma) - \sup_{\Sigma\in\mathcal{M}^{\psi}_{PEV}(G)} \ell_n^{dual}(\Sigma)\big) > \chi^2_{2,1-\alpha}$ holds if and only if

$$\psi \in \left[\frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} - \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}, \frac{-(\widehat{\Sigma}^{-1})_{i,j|d(i)\setminus\{j\}} + \sqrt{D(G)}}{(\widehat{\Sigma}^{-1})_{j,j|d(i)\setminus\{j\}}}\right].$$

The test inversion approach and intersection union theory complete the proof. ■

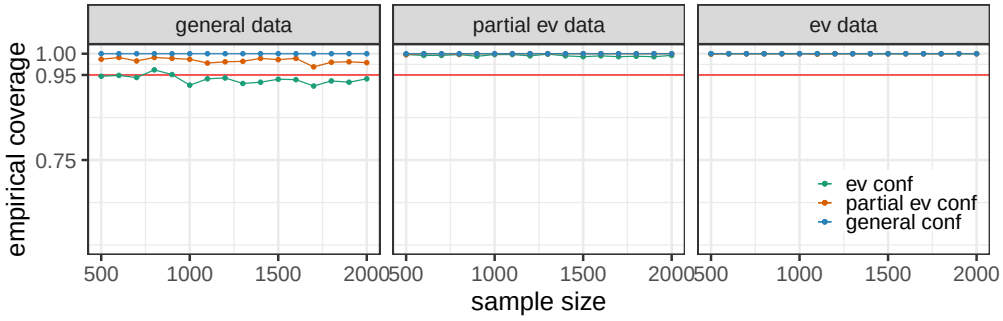


Figure 4: Empirical coverage of 95%-confidence intervals for the total causal effect.

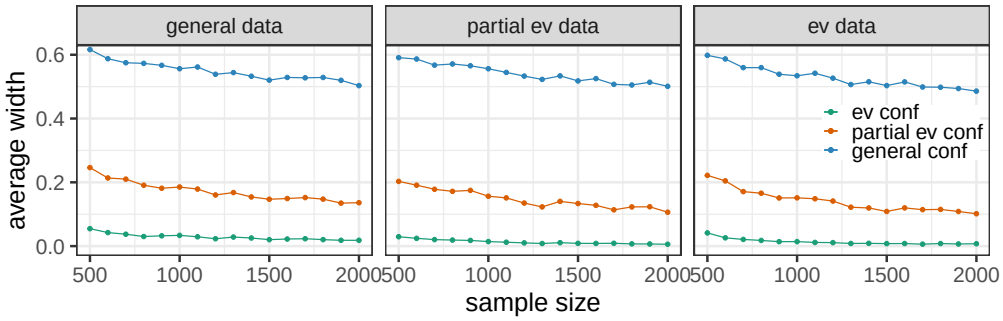


Figure 5: Mean width of 95%-confidence intervals for the total causal effect.

Appendix E. Additional Simulation Results

In this section, we recapitulate the synthetic data generation process adapted from Strieder and Drton (2024) and present further simulation results for true non-zero effects. To generate a synthetic data set, we first select a random permutation of 10 nodes. Second, we prune the corresponding complete graph by including each edge with a probability of 0.5. Then, we generate edge weights according to a normal distribution $\mathcal{N}(0.5, 0.1)$ and generate n samples according to the linear Gaussian SCM represented by the selected DAG. Here, we use three different error variance regimes, that is, `general data` where each error variance is sampled uniformly from $[0.5, 1.5]$, `partial ev data` with the two error variance of potential causal and effect being 1 and the rest sampled uniformly from $[0.5, 1.5]$, and `ev data` with all error variances being 1. We repeated this process to obtain 1000 independent data sets (twice, with true non-zero effect and no effect) to calculate the presented empirical quantities.

Figure 4 shows the empirical coverage frequencies when no true effect exists. Note that this task is 'easier' as no effect corresponds to the larger class of models being plausible where node j proceeds node i in the causal ordering. However, the information contained in the intervals still significantly differs, highlighted by the average width reported in Figure 5. Finally, note that, in this setting, the zero proportions are equal to the empirical coverage.

Permission to include the article

PMLR Proceedings of Machine Learning Research

[edit]

Proceedings of Machine Learning Research

ISSN: 2640-3498

The Proceedings of Machine Learning Research is a series that publishes machine learning research papers presented at conferences and workshops. Each volume is separately titled and associated with a particular workshop or conference and will be published online on the PMLR web site. Authors retain copyright.

Editors

The Series Editor is Neil Lawrence. Please send proposals for new volumes under this series to us via e-mail: proceedings@mlr.press. Each proposal should include:

- A brief description of the event's scope and topics to be covered.
- A description of the review process for the proceedings.
- The names and short CVs (a few lines) of the proposed Proceedings Editors.

For frequently asked questions on publishing proceedings please see the FAQ.

For details on how to prepare a proceedings for publication please see the Specification.

Reissue Series

We have launched a "reissue series" for republishing volumes of papers that are no longer on line. Our first reissues are early editions of AISTATS, AISTATS 1995, AISTATS 1997, AISTATS 1999, AISTATS 2001, AISTATS 2003 and AISTATS 2005. To suggest volumes for the reissue

Permission to include the article

series contact the editors. You will need to provide the information provided in the Proceedings Specification to provide the volume.

Volume R6-draft Reissue of UAI 1998
Volume R0 Pre-proceedings of AISTATS 1995
Volume R5 Proceedings of AISTATS 2005
Volume R4 Proceedings of AISTATS 2003
Volume R3 Proceedings of AISTATS 2001
Volume R2 Proceedings of AISTATS 1999
Volume R1 Proceedings of AISTATS 1997

Proceedings

Volume 256 Proceedings of AutoML 2024
Volume 255 Proceedings of ML Meets Differential Equations
Volume 244 Proceedings of UAI 2024
Volume 230 Proceedings of COPA 2024
Volume 246 Proceedings of PGM 2024
Volume 249 Proceedings of the 1st ContinualAI Unconference
Volume 257 AI for Education Workshop 2024
Volume 228 Proceedings of NeuReps 2023
Volume 245 Proceedings of MLIC 2024
Volume 253 AABI 2024 Proceedings
Volume 248 Proceedings of CHIL 2024
Volume 235 Proceedings of ICML 2024
Volume 247 Proceedings of COLT 2024
Volume 242 Proceedings of L4DC 2024
Volume 241 Proceedings of LIDTA 2023
Volume 243 Proceedings of UniReps
Volume 226 Proceedings of the NeurIPS 2023 Gaze Meets ML Workshop
Volume 239 "I Can't Believe It's Not Better" NeurIPS Workshop 2023
Volume 238 Proceedings of AISTATS 2024
Volume 231 Proceedings of Learning on Graphs 2023
Volume 236 Proceedings of CLeaR 2024
Volume 240 Proceedings of MLCB 2024
Volume 237 Proceedings of ALT 2024
Volume 222 Proceedings of ACML 2023
Volume 227 Proceedings of Medical Imaging with Deep Learning 2023
Volume 233 Proceedings of the Northern Lights Deep Learning Workshop 2024
Volume 234 Proceedings of Conference on Parsimony and Learning 2024
Volume 219 Proceedings of Machine Learning for Healthcare 2023
Volume 225 Proceedings of the 3rd Machine Learning for Health Symposium