Technische Universität München

Department of Mathematics

Master's Thesis

Model uncertainty in statistical inference

Tobias Freidling

Supervisor: Prof. Mathias Drton

Advisors: Prof. Mathias Drton, Prof. Makoto Yamada

Submission Date: 30.09.2020

I assure the single handed composition of this Master's thesis only supported by declared resources.

Cambridge, UK,

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit zwei Themen aus der Statistik, die sich der Frage widmen, wie valide Inferenz-Ergebnisse angesichts von Unsicherheit bezüglich des gewählten statistischen Modells erreicht werden können.

Das erste Teilprojekt ist im Gebiet Kausale Inferenz angesiedelt und hat zum Ziel, Konfidenzintervalle für den kausalen Effekt zwischen den Größen in einem linearen System bestehend aus zwei Variablen zu finden. Zunächst werden die benötigten Begrifflichkeiten eingeführt und grundlegende Annahmen, wie Fehlerterme mit gleicher Varianz, anhand derer die Identifizierbarkeit des Modells gewährleistet wird, vorgestellt. Daraufhin wird veranschaulicht, dass der naive Ansatz, Modellwahl und Schätzung von Konfidenzintervallen in zwei separate Schritte zu zerlegen, oftmals zu falschen Ergebnissen führt. Außerdem zeigt sich anhand zweier Beispiele, dass Resampling-Methoden, wie Bootstrapping oder Subsampling, ebenfalls nicht angewendet werden können.
Aus diesem Grund entwickeln wir einen neuen Ansatz, der auf der Dualität von Hypothesentests und Konfidenzintervallen aufbaut. Es wird eine Familie von Tests mit Hilfe von Constrained Statistical Inference Theory entwickelt, von der wir Konfidenzintervalle ableiten. Abschließend werden die Ergebnisse der vorgestellten Methoden sowohl an künstlich erzeugten Datensätzen als auch an Benchmarks aus realen Anwendungen untersucht.

Das zweite Teilprojekt befasst sich mit Inferenz in einem Modell, das unter Berücksichtigung der vorliegenden Daten durch HSIC-Lasso ausgewählt wurde. Es wird zunächst das Polyhedral Lemma eingeführt, auf dessen Grundlage ein Hauptstrang der Forschung auf dem Gebiet der Selektiven Inferenz basiert und das auch in dieser Arbeit Anwendung findet. Als zweiten theoretischen Eckpfeiler stellen wir nachfolgend das Hilbert-Schmid Unabhängigkeitskriterium vor, das erlaubt die Abhängigkeit zweier Zufallsvariablen ohne Verwendung weitergehender Annahmen zu quantifizieren.
Auf dieser Grundlage wird eine Methode für korrekte Inferenz nach Modellwahl durch das HSIC-Lasso Verfahren entwickelt. Darüber hinaus werden verschiedene potentielle Inferenz-Größen vorgestellt und es wird ebenfalls auf Probleme bei der Anwendung auf hochdimensionale Datensätze eingegangen. Daraufhin wird mittels verschiedener künstlich generierter Daten analysiert, wie sich die Wahl von bestimmten Parametern auf die erzielten Ergebnisse auswirken. Zuletzt wird mit zwei Benchmark-Datensätzen die praktische Anwendbarkeit des vorgestellten Ansatzes unter Beweis gestellt.

Contents

1 Introduction

Model uncertainty is a theme which pervades statistics as a whole. In a broad sense, it addresses the central question of working out a reasonable model for observed data and is thus present in exploratory data analysis, hypothesis testing for nested models and the development of non-parametric methods among many others. This work, however, understands model uncertainty as a data-driven choice of model preceding an inference procedure. In such a situation we have to account for the decision made in the selection stage to ensure that the results of hypothesis testing and confidence interval calculation are valid. In particular, the fields of causal and selective inference, which have attracted a lot of research activity in the last years, are concerned with the ramifications of model choice and subsequent inference.

Causal inference typically addresses low-dimensional settings where both the direction and the strength of causal relationships among the involved variables are unknown. It goes beyond the bread and butter concept of correlation inasmuch not only coordinated behaviour between two variables is captured, but also the underlying dependence structure, in other words the data generating process, is investigated in more depth. For instance, the statement that coronary heart disease is caused by dietary fat intake provides considerably more insight than the claim that coronary heart disease and dietary fat intake are just associated, c.f. (Maathuis et al. 2019).
In order to examine causal dependencies, randomised control trials have been established as the gold standard. They are used in an experimental setting where interventions on variables can be performed and provide researchers with a powerful framework to investigate their hypotheses. Despite the power of randomised control trials, it is not always possible to conduct such experiments due to inter alia the involved costs, ethical principals or physical constraints. Moreover, it might be interesting to ask questions concerning causality on datasets that were collected without regard to the examination of causal relationships. For this reason, we follow the less assuming principal of observational data which assumes a given dataset whose collection was not influenced by the researcher, see for example (Spirtes et al. 2001).
Causality is an inherently multi-disciplinary topic and it is thus not surprising that various mathematical frameworks were developed which emphasise different aspects of causal relationships. The earliest formalisations go back to (Wright 1921) and (Spława-Neyman 1923), albeit their pioneering work was not followed up until the 1970s and 1980s. Their initial ideas were elaborated to become three different terminologies for causality that are equivalent but suited for different purposes. We briefly outline them according to (Spirtes et al. 2001), (Pearl 2009) and (Maathuis et al. 2019). Modelling causal relationships with directed edges between the nodes of a graph is a very intuitive and easy to visualise way of describing causality. However, only the existence but not the strength of causal relationships can be described well in this framework. This issue does not occur in the terminology of structural equation models which represent causality as a collection of assignments. It is important to note that these assignments describe the data generating process and can therefore not be treated as mere equations. In practice, there is often the risk of confusing the language of structural equation models with the notation of classical regression. The third framework that evolved in the context of causality are potential

outcomes or rather counterfactuals. They can become particularly complex for large systems but provide an easy language to describe interventions where some variables in the data generating process are set to a certain value.

Since we are interested in inference on causal effects but do not consider interventions, we use the framework of structural equation models in this work. More specifically, we consider a two-variable system and assume a linear relationship, a situation which is covered in much detail in (J. Peters et al. 2017). In this setting, we incorporate the uncertainty regarding the direction of the causal relationship into the causal effect. For instance, if X_2 has influence on X_1 of strength 0.5, the causal effect of X_2 on X_1 is 0.5 but the causal effect of X_1 on X_2 is 0 because the relationship is directed opposite.

In such a situation, the identifiability of the direction of the causal relationship is key. It was shown that unique identification is possible under following assumptions: non-linearity with additive errors, linearity with non-Gaussian errors, and linearity with errors of equal variance, cf. (Heinze-Deml et al. 2018) or (Maathuis et al. 2019). We concentrate on the latter setting which was analysed by (Jonas Peters and Bühlmann 2014) and (Loh and Bühlmann 2014).

In this framework, we undertake the construction of confidence intervals for the causal effect which, to the best of our knowledge, was not previously done. We show that a naive approach as well as resampling methods, such as bootstrapping or subsampling, surprisingly fail. For this reason, we use the duality of hypothesis tests and confidence intervals, and the framework of constrained statistical inference to construct suitable tests.

Unlike causal inference, selective inference treats classification and regression settings and thus implicitly assumes that the direction of causality is clear. Nevertheless, model uncertainty is a common theme, especially when the number of covariates exceeds the sample size; this is called a 'small n, large p problem'. In such a situation, we need to select a subset of variables to enter the model in order to ensure identifiability. Moreover, model selection serves the endeavour of a parsimonious and interpretable model.

Statistical research produced an extensive literature on different selection methods, of which we mention only a few. (Hocking 1976) proposes the algorithmic forward stepwise selection procedure and (Akaike 1974) puts forward the Akaike information criterion, which relies on the log-likelihood and degrees of freedom of a model. (R. Tibshirani 1996) pioneered the class of regularisation methods with his least absolute shrinkage and selection operator (Lasso). In the spirit of his proposal, a huge variety of different regularisation procedures were developed, e.g. ridge regression, elastic net, Group Lasso or SCAD. For a review of the latter see (Hastie et al. 2015).

In this work we treat the HSIC-Lasso selection procedure which was introduced in (Makoto Yamada, Jitkrittum, et al. 2014) and, as the name suggests, is based on the Hilbert-Schmidt independence criterion (HSIC) (Gretton, Bousquet, et al. 2005). This is a kernel based method that allows to detect dependence between any two random variables and marks the culminating point of a line of research concerning the use of reproducing kernel Hilbert space in machine learning. It found many applications due to its universality and lack of necessary assumptions. The HSIC-Lasso selection procedure uses the Hilbert-Schmidt independence criterion as a measure for the degree of dependence. It selects covariates, which influence the response and exhibit little dependence on other covariates, in a Lasso-type optimisation. Thus, it provides a way to harness the computational and theoretical advantages of quadratic optimisation and the universality of model-free dependence estimation.

Traditional statistical theory assumes that a model is given a priori and only its parameters are object of inference. Yet, if we conduct a data-driven model choice which precedes the inference stage, we have to account for the selection as the covariates entering the model are overly significant. Otherwise, they would not have been selected. (Leeb and Pötscher 2005) and (Leeb and Pötscher 2006) among many other papers investigate how dramatic the effects of neglecting the selection process are and how valid post-selection inference can be carried out. For instance, they show that the implications of model choice carry over into the limit $n \to \infty$. Roughly speaking, it became clear that model selection can greatly distort inference results and that this problem poses considerable difficulties for statistical research.

Two major paradigms regarding the treatment of model choice evolved. (Berk et al. 2013) developed a post-selection inference method which is protected against distorting effects of any selection procedure. By design, this method is very robust but, in practice, often leads to overly conservative confidence intervals. In contrast, (Jason D. Lee et al. 2016) and (Ryan J. Tibshirani, J. Taylor, et al. 2016) only account for the actual selection outcome by conditioning on it in the inference step. This requires control over or at least insight into the model choice which, however, is usually the case. Their seminal insight is the formulation of the selection event as a restriction on the distribution of the inference target. Under the assumption of Gaussianity, it is even possible to derive a pivotal quantity for finite sample sizes. In the following, we refer to this framework as truncated Gaussian setting. Due to its low computational costs and relative generality in terms of the inference target, several adaptations and generalisations were proposed.

This work shows how the theory based on truncated Gaussians can be applied to the HSIC-Lasso selection procedure for different inference targets. Moreover, issues, such as hyperparameter choice or computational costs, which are crucial in real-world applications are addressed.

2 Causal effect inference

This section is organised as follows. In the first Subsection 2.1 the framework of linear structural equation models is introduced and the causal effect of one variable on another is defined. In the subsequent Subsections 2.2 and 2.3, we show that a naive approach as well as resampling methods in general fail to construct valid confidence intervals. For this reason, we develop a method which uses the duality of confidence intervals and hypothesis testing and the framework of constrained statistical inference to get valid confidence intervals in Subsection 2.4. We conclude with experimental evaluations both on artificial and benchmark data in 2.5.

2.1 Two-variable linear structural equation models

2.1.1 Structural equation models

Although structural equation models do not require the notion of graphs, it is instructive to introduce them in correspondence to directed graphs as they provide an intuitive understanding of causality.

Definition 2.1. Let $\mathcal{V} = \{1, \dots, d\}$ be a set of *vertices* and $\mathcal{E} \subset \{(j,k) : j, k \in \{1, \dots, n\}, j \neq k\}$ be a set of directed *edges* joining the vertices. Then the tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is called *directed graph.*
Considering an edge $j \to k$, j is referred to as *parent* and k is called *child.* The sets of parents and children of a node j are denoted by $\mathrm{pa}(j)$ and $\mathrm{ch}(j)$ respectively.
Considering a directed path $j \to \cdots \to k$, j is said to be an *ancestor* of k and k is a *descendant* of j. The corresponding sets for a node j are $\mathrm{an}(j)$ and $\mathrm{de}(j)$. If $j = k$, the path is referred to as *cycle* and a directed graph that does not have cycles is said to be *acyclic.*

We want to model the causal dependence structure among the components of a random vector $X = (X_1, \dots, X_d)$ by linking it to a directed acyclic graph $\mathcal{G}$, abbreviated DAG. The nodes of $\mathcal{G}$ are represented by the components of X and the edges depict the causal relationships among them. A priori, both the structure of the graph and the degree of influence among the variables are unknown and subject to statistical inference.
Additional to the direction of causal relationships, the framework of structural equation models (SEMs), as described in (Maathuis et al. 2019) with earlier work by (Pearl 1995), allows to quantify their strength.

Definition 2.2. Let $X = (X_1, \dots, X_d)$ be a random vector with a causal structure that is linked to a DAG $\mathcal{G}$. A *structural equation model* for X on $\mathcal{G}$ assumes for each node $j \in \mathcal{V}$ that X_j is a function of its graphical parents and possibly a random variable ε_j,

$$X_j := f_j(X_{\mathrm{pa}(j)}, \varepsilon_j), \qquad \forall j \in \{1, \dots, d\}. \tag{2.1}$$

Note that in the definition above the operator := was used instead of the equals sign in order to point out that we want to express an asymmetric assignment of the value of f_j to X_j. This reflects the idea of a sequential data generating process. Consequently, a node has influence on its descendants but not on its ancestors. Making additional assumptions, we define a smaller, more manageable class of models.

Definition 2.3. Let $X = (X_1, \dots, X_d)$ be a centred random vector with a causal structure that is linked to a DAG $\mathcal{G}$. A *linear structural equation model* (LSEM) is given by

$$X_j := \sum_{k=1}^{d} \beta_{jk} X_k + \varepsilon_j, \qquad \forall j \in \{1, \dots, d\},$$

where $\{\varepsilon_j\}_{j \in \{1,\dots,d\}}$ are independent random variables with mean zero and $\{\beta_{jk}\}_{j,k \in \{1,\dots,d\}}$ are unknown parameters. For all of them we require

$$\beta_{jj} = 0 \quad \text{and} \quad \beta_{jk} = 0 \Leftrightarrow k \notin \mathrm{pa}(j).$$

In the definition above, we could equally use the formula

$$X_j := \sum_{k \,\in\, \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j, \qquad \forall j \in \{1, \dots, d\},$$

which is more in line with the general definition. However, using the given framework is more convenient as it describes both the causal structure and the degree of influence among the variables in terms of $\{\beta_{jk}\}_{j,k \in \{1,\dots,d\}}$. Hence, $\beta_{jk} = 0$ encodes the absence of a causal relationship $j \leftarrow k$ whereas $\beta_{jk} \neq 0$ establishes its existence and simultaneously quantifies its strength. Therefore, the dependence structure is described by a $d \times d$ matrix $B = (\beta_{jk})_{j,k \in \{1,\dots,d\}}$. Dealing with asymmetric assignments rather than a symmetric equality relation gives rise to the definition of conditional expectation in the context of causality.

Definition 2.4. In the SEM-framework the *causal conditional expectation* of X_j given $X_k = x_k$ is defined by

$$\mathrm{E}\left[X_j \| X_k = x_k\right] = \mathrm{E}\left[X_j | X_k = x_k\right] \mathbf{1}_{\{X_j \in \mathrm{de}(k)\}} + \mathrm{E}\left[X_j\right] \mathbf{1}_{\{X_j \notin \mathrm{de}(k)\}}$$

where $\mathrm{E}\left[\,\cdot\,|\,\cdot\,\right]$ denotes the usual conditional expectation.

With this definition at hand, one can investigate how X_j changes in mean when x_k is modified. This is precisely the causal influence of X_k on X_j we want to quantify. For linear models it can be expressed by a single number, cf. (Pearl 2009).

Definition 2.5. In a LSEM the *causal effect* of X_k on X_j is given by

$$\frac{d}{dx_k} \mathrm{E}\left[X_j \| X_k = x_k\right]. \tag{2.2}$$

Remark 2.1. We can directly deduce from the definition of the causal conditional expectation that the causal effect of X_k on X_j is zero if X_j is not a descendant of X_k.

Remark 2.2. In the more general setting (2.1) the causal effect of X_k on X_j is contained in f_j and its dependence on X_k. This makes it necessary to consider more than the single quantity $\frac{d}{dx_k} \mathrm{E}\left[X_j \| X_k = x_k\right]$ to adequately describe the causal relationship.

As we have seen, the direction of the causal relationship between two random variables X_j and X_k enters into the causal effect of X_k on X_j. Therefore, we have to assure that the graph is identifiable. In accordance with (W. Chen et al. 2019) and (Jonas Peters and Bühlmann 2014), we assume that $\{\varepsilon_j\}_{j \in \{1,\dots,d\}}$ have a common variance $\sigma^2 > 0$ which allows us to express X and its covariance matrix Σ in terms of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$ and B,

$$\begin{aligned} X &= (\mathrm{Id} - B)^{-1} \varepsilon, \\ \Sigma &= \mathrm{E}\left[XX^T\right] = \sigma^2 (\mathrm{Id} - B)^{-1} (\mathrm{Id} - B)^{-T}. \end{aligned} \tag{2.3}$$

2.1.2 Two-variable system

In the following, we consider a two-variable LSEM further assuming that ε_1 and ε_2 have the same variance $\sigma^2 > 0$ and treat, without loss of generality, only the causal effect of X_2 on X_1. In this framework there are only three possible models

$$\begin{array}{lll} \text{(M1)} & X_1 = \beta_{12}X_2 + \varepsilon_1, & X_2 = \varepsilon_2, \\ \text{(M2)} & X_1 = \varepsilon_1, & X_2 = \beta_{21}X_1 + \varepsilon_2, \\ \text{(M3)} & X_1 = \varepsilon_1, & X_2 = \varepsilon_2, \end{array}$$

where the third can be seen as a special case of either the first or the second with $\beta_{12} = 0$ or $\beta_{21} = 0$ respectively.
Under the equal variance assumption, (W. Chen et al. 2019) shows how the model can be identified: The child variable has a higher variance than the parent variable.

Example 2.1. If the true model is

$$\begin{aligned} X_1 &= \beta_{12}X_2 + \varepsilon_1, \\ X_2 &= \varepsilon_2, \end{aligned}$$

computing the variance of X_1 and X_2 directly yields

$$\begin{aligned} \mathrm{var}\,(X_1) &= \mathrm{var}\,(\beta_{12}X_2 + \varepsilon_1) = \beta_{12}^2\sigma^2 + \sigma^2 = \sigma^2(\beta_{12}^2 + 1), \\ \mathrm{var}\,(X_2) &= \mathrm{var}\,(\varepsilon_2) = \sigma^2 < \sigma^2(\beta_{12}^2 + 1), \end{aligned}$$

which shows that the child X_1 has larger variance.

Using this insight and (2.2), we can compute the causal effect of X_2 on X_1 as follows.

$$\begin{aligned} \frac{d}{dx_2}\mathrm{E}\left[X_1 \| X_2 = x_2\right] &= \frac{d}{dx_2}\Big(\mathrm{E}\left[X_1 | X_2 = x_2\right]\mathbf{1}_{\{X_1 \in \mathrm{de}(2)\}} + \mathrm{E}\left[X_1\right]\mathbf{1}_{\{X_1 \notin \mathrm{de}(2)\}}\Big) \\ &= \frac{d}{dx_2}\Big(\mathrm{E}\left[\beta_{12}X_2 + \varepsilon_1 | X_2 = x_2\right]\mathbf{1}_{\{X_1 \in \mathrm{de}(2)\}}\Big) \\ &= \frac{d}{dx_2}\Big(\beta_{12}x_2\mathbf{1}_{\{X_1 \in \mathrm{de}(2)\}}\Big) \\ &= \beta_{12}\mathbf{1}_{\{\mathrm{var}(X_1) > \mathrm{var}(X_2)\}} \end{aligned} \tag{2.4}$$

Assuming that one of the models (M1), (M2) and (M3) holds constrains the covariance of (X_1, X_2). We identify the general space of 2 × 2 covariance matrices with the cone

$$\mathrm{C} = \left\{(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathbb{R}^3 : \sigma_{11}, \sigma_{22} > 0,\, \sigma_{12}^2 < \sigma_{11}\sigma_{12}\right\}, \tag{2.5}$$

where we choose the strict inequalities as we want to exclude almost surely constant and perfectly correlated random variables. We find that only subspaces of C are coherent with the corresponding models: Assuming (M1), we calculate the covariance matrix of (X_1, X_2) denoted by Σ according to (2.3). To abbreviate the notation, we use $\mathrm{var}\,(X_1) = \sigma_{11}$, $\mathrm{var}\,(X_2) = \sigma_{22}$ and $\mathrm{cov}\,(X_1, X_2) = \sigma_{12}$, and get

$$\Sigma = \mathrm{E}\left[XX^T\right] = \sigma^2\begin{pmatrix} 1 + \beta_{12}^2 & \beta_{12} \\ \beta_{12} & 1 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

This establishes the identities

$$\beta_{12} = \frac{\sigma_{12}}{\sigma_{22}}, \qquad 1 + \beta_{12}^2 = \frac{\sigma_{11}}{\sigma_{22}}, \tag{2.6}$$

from which an additional constraint on the structure of Σ ensues,

$$\sigma_{12}^2 + \sigma_{22}^2 = \sigma_{11}\sigma_{22}. \tag{2.7}$$

Undertaking similar computations for (M2) and (M3) yields the covariance spaces

$$\mathrm{C}_1 := \{(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}\colon \sigma_{12}^2 + \sigma_{22}^2 = \sigma_{11}\sigma_{22}\}, \tag{2.8}$$
$$\mathrm{C}_2 := \{(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}\colon \sigma_{12}^2 + \sigma_{11}^2 = \sigma_{11}\sigma_{22}\}, \tag{2.9}$$
$$\mathrm{C}_3 := \{(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}\colon \sigma_{12} = 0,\ \sigma_{11} = \sigma_{22}\}, \tag{2.10}$$

for the respective models. Closely checking the constraints imposed in C_3, we notice that $\mathrm{C}_3 \subset \mathrm{C}_1$ and $\mathrm{C}_3 \subset \mathrm{C}_2$. Using (2.2) and (2.6), we can characterise the causal effect of X_2 on X_1 as a function of the covariance matrix entries.

Definition 2.6. In a two-variable LSEM satisfying the equal variance assumption the causal effect of X_2 on X_1 is given by

$$T\colon \mathrm{C}_1 \cup \mathrm{C}_2 \to \mathbb{R}, \quad (\sigma_{11}, \sigma_{12}, \sigma_{22}) \mapsto \frac{\sigma_{12}}{\sigma_{22}} \mathbf{1}_{\{\sigma_{11} > \sigma_{22}\}}, \tag{2.11}$$

where C_1 and C_2 are defined as in (2.8) and (2.9) respectively.

Thus, the causal effect becomes a parameter that only relies on second-order information but does not require the assumption of a particular distribution or family of distributions. Moreover, it is straightforward to use a plug-in estimator to find a point estimate for the causal effect of a data sample. In the following, we use the notations $T(\sigma_{11}, \sigma_{12}, \sigma_{22})$ and $T(\Sigma)$ interchangeably.

Lemma 2.7. *T is continuous but cannot be extended to a continuously differentiable function on an open set containing* $\mathrm{C}_1 \cup \mathrm{C}_2$.

Proof. As a composition of continuous functions, T is obviously continuous for $\sigma_{11} \neq \sigma_{22}$ as $\sigma_{22} > 0$. For the case $\sigma_{11} = \sigma_{22}$, we use the identities (2.6). In the limit $\sigma_{11} \to \sigma_{22}$, $1 + \beta_{12}^2 = \sigma_{11}/\sigma_{22}$ becomes $1 + \beta_{12}^2 = 1$. Consequently, $\beta_{12} = \sigma_{12}/\sigma_{22} \to 0$ which proves that T is continuous on the whole domain.
In order to show that T cannot be continuously differentiably extended, we consider an arbitrary open set $O \supset \mathrm{C}_1 \cup \mathrm{C}_2$. Therefore, partial derivatives for points within O are well-defined. Hence, we can consider the derivative

$$\frac{\partial T}{\partial \sigma_{12}}(\sigma_{11}, \sigma_{12}, \sigma_{22}) = \frac{1}{\sigma_{22}} \mathbf{1}_{\{\sigma_{11} > \sigma_{22}\}}$$

and the sequence $((1 + \frac{1}{n}, \frac{1}{\sqrt{n}}, 1))_{n \in \mathbb{N}} \subset \mathrm{C}_1 \subset O$. We easily find that $\frac{\partial T}{\partial \sigma_{12}}(1 + \frac{1}{n}, \frac{1}{\sqrt{n}}, 1) = 1$ for all $n \in \mathbb{N}$ whereas $\frac{\partial T}{\partial \sigma_{12}}(1, 0, 1) = 0$. □

In the ensuing sections, we follow different approaches to construct confidence intervals for the causal effect of X_2 on X_1 in a two-variable LSEM under the equal variance assumption. Throughout, we assume that the data $\mathbf{X}^n = ((X_1^{(1)}, X_2^{(1)}), \ldots, (X_1^{(n)}, X_2^{(n)}))$ created by the model satisfies the usual assumptions of independence and identical distribution.

2.2 Naive approach

Following a naive approach, we split the task of building a confidence interval under model uncertainty into two steps.
First, we determine the direction of the causal relationship and hence the model. As seen in the previous section, this can be easily done by comparing the empirical variances, denoted by $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$, as the dependent variable has larger variance. $\hat{\sigma}_{11} \leq \hat{\sigma}_{22}$ implies the causal relationship $X_1 \rightarrow X_2$, whereas $\hat{\sigma}_{11} > \hat{\sigma}_{22}$ indicates $X_1 \leftarrow X_2$.
Second, we condition on the model choice made and find that for the former case we can only use $\{0\}$ as confidence interval for the causal effect of X_2 on X_1. In the latter case we obtain the linear regression equation

$$X_1 = \beta_{12} X_2 + \varepsilon_1,$$

where the causal effect is β_{12}. According to (Fahrmeir et al. 2013), we may use a standard confidence interval for regression coefficients

$$\left(\hat{\beta}_{12} \pm \widehat{sd} \cdot z_{1-\frac{\alpha}{2}}\right),$$

which is based on the ordinary least square estimates $\hat{\beta}_{12}$ and $\hat{\sigma}$. Here, $\widehat{sd} = \hat{\sigma}/\|\mathbf{X}_2^n\|_2$ is the estimated standard deviation of $\hat{\beta}_{12}$ and z_γ denotes the γ-quantile of a standard normal distribution. This confidence interval is asymptotically valid; however, under the assumption of normal errors, we can use the exact confidence interval

$$\left(\hat{\beta}_{12} \pm \widehat{sd} \cdot t_{1-\frac{\alpha}{2};n-1}\right).$$

We denote the number of samples as n and $t_{\gamma;m}$ is the γ-quantile of a t-distribution with m degrees of freedom.
In summary, we constructed the $(1-\alpha)$-confidence interval

$$I = \begin{cases} \{0\}, & \text{if } \hat{\sigma}_{11} \leq \hat{\sigma}_{22} \\ \left(\hat{\beta}_{12} \pm \widehat{sd} \cdot z_{1-\frac{\alpha}{2}}\right), & \text{if } \hat{\sigma}_{11} > \hat{\sigma}_{22}. \end{cases}$$

Splitting the construction of confidence intervals into two steps, we tacitly condition on the model choice. Hence, the obtained interval is only valid under a correctly specified model. Suppose, for instance, that the causal relationship $X_1 \leftarrow X_2$ holds and $|\beta_{12}| \ll 1$ meanwhile having a small to moderate number of samples. Consequently, the true variances of X_1 and X_2, σ_{11} and σ_{22}, are very close, cf. Example 2.1. Therefore, their empirical estimates are similar and with probability close to 0.5 the confidence interval constructed from the data becomes $I = \{0\}$. Yet, this obviously does not contain the true parameter. This case demonstrates that the naive approach of estimating the model and the strength of the causal relationship in a stepwise procedure is impracticable. Not including the uncertainty regarding the underlying model into the length of the confidence interval is particularly detrimental in situations with feeble causality. Therefore, we need to investigate approaches which simultaneously assess the direction and the strength of the causal relationship.

2.3 Resampling techniques

As stated in Definition 2.6, the causal effect of X_2 on X_1 is a function of the parameters of the model space. This gives rise to the approach of using resampling techniques, such as bootstrapping and subsampling, in order to approximate the distribution of a root quantity depending on $T(\Sigma)$ and ultimately deduce confidence intervals. First, we briefly outline bootstrapping and subsampling, and subsequently apply these methods to the causal effect.

2.3.1 Bootstrapping and subsampling

Following (Politis et al. 1999), we consider a statistical model $\{\mathbb{P}\colon \mathbb{P} \in \mathbf{P}\}$ and assume that independent and identically distributed observations $\mathbf{X}^n = (X^{(1)}, \dots, X^{(n)})$ are given. $\mathbf{P}$ can represent a parametric and non-parametric set of distributions alike. We want to study the real-valued quantity $g(\mathbb{P})$, where g is a known function depending on $\mathbb{P}$, and denote its estimate $g(\hat{\mathbb{P}}_n)$. In non-parametric models $\hat{\mathbb{P}}_n$ may be the empirical distribution, whereas parametric models can leverage ordinary least square or maximum likelihood estimation among others.

In order to construct confidence intervals, one frequently considers a pivotal quantity $R_n(X_n, g(\mathbb{P}))$, also referred to as root. Its distribution for a sample of size n is denoted by $J_n(\mathbb{P})$ and its cumulative distribution function by $J_n(\cdot, \mathbb{P})$. The commonly chosen quantity

$$R_n(\mathbf{X}^n, g(\mathbb{P})) = \tau_n\Big(g(\hat{\mathbb{P}}_n) - g(\mathbb{P})\Big) \tag{2.12}$$

with normalising constant τ_n allows to easily derive confidence intervals if its distribution is known. Resampling methods provide an approximation to $J_n(\mathbb{P})$, that does not require R to be a (asymptotically) pivotal quantity, which makes them applicable in a wide range of problems.

Definition 2.8. Let R_n be the root as stated in (2.12) and $J_n(\mathbb{P})$ its distribution. The *bootstrap approximation* of $J_n(\mathbb{P})$, denoted by $J_n(\hat{\mathbb{P}}_n)$, is the empirical distribution of the B values

$$R_n(\mathbf{X}^{n,i}, g(\hat{\mathbb{P}}_n)) = \tau_n\Big(g(\hat{\mathbb{P}}^i_n) - g(\hat{\mathbb{P}}_n)\Big), \quad i \in \{1, \dots, B\}.$$

Here $\hat{\mathbb{P}}^i_n$ denotes the i-th estimate of $\mathbb{P}_n$ calculated from the dataset $\mathbf{X}^{n,i}$ which is obtained by resampling n values of $\mathbf{X}^n$ with replacement.

A bootstrap confidence region for $g(\mathbb{P})$ of nominal level $1-\alpha$ can be deduced by

$$B_n(1-\alpha, \mathbf{X}^n) = \left\{g(\mathbb{P})\colon J_n^{-1}\left(\frac{\alpha}{2}, \hat{\mathbb{P}}_n\right) \leq R_n(\mathbf{X}^n, g(\mathbb{P})) \leq J_n^{-1}\left(1-\frac{\alpha}{2}, \hat{\mathbb{P}}_n\right)\right\}.$$

Remark 2.3. From the viewpoint of computational statistics one can equally regard bootstrapping as a Monte Carlo approximation of $J_n(\mathbb{P})$.

While abstract, sufficient conditions for the equivalence of the bootstrapped and true distribution in the limit $n \to \infty$ are known, in many applications a differentiability assumption on g is vital. In the following, F and $\hat{F}_n$ denote the cumulative distribution functions, abbreviated cdf, of $\mathbb{P}$ and $\hat{\mathbb{P}}_n$ respectively and $\mathbf{F}$ is the set of cdfs corresponding to $\mathbf{P}$.

Definition 2.9. The functional g defined on $\mathbf{F}$ is *Fréchet differentiable* at F with respect to $\|\cdot\|$ if there exists a function h_F such that $\int h_F dF = 0$ and for any $G \in \mathbf{F}$,

$$g(G) = g(F) + \int h_F \, d(G - F) + o(\|G - F\|)$$

holds as $\|G - F\| \to 0$. The function h_F is called *Fréchet derivative* of g at F.

Theorem 2.10. *Assume* $\mathbf{F}$ *contains distributions with finite support and* g *has Fréchet derivative* h_F *at* F *with respect to* $\|\cdot\|_\infty$*. Also, assume that* $0 < \int h_F^2 dF < \infty$ *and choose* $\tau_n = n^{1/2}$ *in (2.12). Then with probability one*

$$\rho_L\big(J_n(F), J_n(\hat{F}_n)\big) \to 0, \qquad \text{as } n \to \infty,$$

for the Lévy metric ρ_L *and*

$$\mathbb{P}\big(g(\mathbb{P}) \in B_n(1 - \alpha, \mathbf{X}^n)\big) \to 1 - \alpha, \qquad \text{as } n \to \infty.$$

In a two-variable LSEM assuming equal variance, $\mathbf{F}$ contains all distributions whose covariance matrices lie within $\mathrm{C}_1 \cup \mathrm{C}_2$. Hence, in an attempt to leverage Theorem 2.10, we investigate the Fréchet differentiability of $T(\Sigma(F))$. Applying the chain rule, we quickly see that T has to be totally differentiable with respect to its arguments. However, according to Lemma (2.7) T cannot be extended to a larger domain such that it is continuously differentiable and consequently has no total derivative. Therefore, Theorem 2.10 cannot be applied to assert the validity of bootstrap confidence intervals.

Nevertheless, further following this approach is sensible as in many cases, where the standard bootstrap fails, modifications which ensure a correct approximation are known. (Andrews 2000) illustrates some of them with a short example and gives references to other works concerned with the same or similar issue.

Definition 2.11. Let R_n be the root as stated in (2.12) and $J_n(\mathbb{P})$ its distribution. The *subsampling approximation* of $J_n(\cdot, \mathbb{P})$, denoted by $L_{n,b}$, is the empirical distribution of the B values

$$R_n(\mathbf{X}^{n,b,i}, g(\hat{\mathbb{P}}_n)) = \tau_n \left(g(\hat{\mathbb{P}}^i_{n,b}) - g(\hat{\mathbb{P}}_n) \right), \quad i \in \{1, \ldots, B\}.$$

Here $\hat{\mathbb{P}}^i_{n,b}$ denotes the i-th estimate of $\mathbb{P}_n$ calculated from the dataset $\mathbf{X}^{n,b,i}$ which is obtained by resampling b values of $\mathbf{X}^n$ without replacement.

A subsampling confidence region for $g(\mathbb{P})$ of nominal level $1 - \alpha$ can be deduced by

$$B_{n,b}(1 - \alpha, \mathbf{X}^n) = \left\{ g(\mathbb{P}) \colon L^{-1}_{n,b}\left(\frac{\alpha}{2}\right) \leq R_n(\mathbf{X}^n, g(\mathbb{P})) \leq L^{-1}_{n,b}\left(1 - \frac{\alpha}{2}\right) \right\}.$$

As for bootstrapping, results on the validity of the confidence intervals in the limit $n \to \infty$ are known.

Theorem 2.12. *Assume that* $J(F_n)$ *converges weakly to a non-degenerate limit law* $J(F)$ *as* $n \to \infty$*. Also assume that* $\tau_b/\tau_n \to 0$*,* $b \to \infty$*,* $b/n \to 0$ *and* $B \to \infty$ *as* $n \to \infty$*. If* x *is a continuity point of* $J(\cdot, \mathbb{P})$*, then*

$$L_{n,b}(x) \to J(x, \mathbb{P}), \qquad \text{as } n \to \infty.$$

Moreover, if $J(\cdot, \mathbb{P})$ *is continuous at the respective quantiles,*

$$\mathbb{P}\big(g(\mathbb{P}) \in B_{n,b}(1 - \alpha, X_n)\big) \to 1 - \alpha, \qquad \text{as } n \to \infty.$$

Note that Theorem 2.12 does not pose requirements on g and can, in this sense, handle a larger class of problems than bootstrapping. However, in scenarios in which both procedures yield asymptotically correct confidence intervals the bootstrap approximation often converges faster than subsampling.

Both methods rely on the basic assumption that $J(F_n)$ converges weakly to a non-degenerate limit law. In the case of bootstrapping, Theorem 2.10 does not state this underlying assumption explicitly but rather ensures it with stronger conditions. Therefore, we will aim to define an appropriate root for the causal effect of X_2 on X_1 and investigate its asymptotic distribution in the following sections.

2.3.2 Continuous extension

A straightforward choice for the root is

$$\tau_n\Big(T(\widehat{\Sigma}_n) - T(\Sigma)\Big), \tag{2.13}$$

where $\widehat{\Sigma}_n$ denotes the empirical covariance matrix and which is defined by

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^T = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{pmatrix}.$$

However, the root (2.13) is not-well defined as T only takes values in $\mathrm{C}_1 \cup \mathrm{C}_2$ and $\widehat{\Sigma}_n$ is only restricted to the space of general 2×2 covariance matrices C. Two solutions for this issue come to mind.

First, instead of using the empirical covariance matrix one can attempt to construct an estimator $\tilde{\Sigma}_n$ that leverages the underlying structure of the model and takes values only in $\mathrm{C}_1 \cup \mathrm{C}_2$. Since we have to verify the convergence of the root distribution to a limiting law, this approach, though, is less feasible as it is unclear how the asymptotic distribution of

$$\tau_n\left(T(\tilde{\Sigma}_n) - T(\Sigma)\right)$$

can be derived.

Second, one can extend T to $\tilde{T}$ which is defined on the whole domain C and use the root

$$R(\mathbf{X}^n, \Sigma) = \tau_n\left(\tilde{T}(\widehat{\Sigma}_n) - \tilde{T}(\Sigma)\right). \tag{2.14}$$

This approach does not harness the structure of the model and consequently looses power. Yet, it is possible to derive the asymptotic distribution of $R(\mathbf{X}^n, \Sigma)$ using results on asymptotic normality of the empirical covariance matrix. Therefore, we follow this approach and focus on continuous extensions of T as we want to guarantee that for estimates $\widehat{\Sigma}_n$ close to $\mathrm{C}_1 \cup \mathrm{C}_2$ the assigned causal effect is also close to the real causal effect. In the following, we present two examples for such continuous extensions.

Example 2.2. Choosing $\tilde{T}_1$ of the form $\tilde{T}_1(\sigma_{11}, \sigma_{12}, \sigma_{22}) = f(\sigma_{11}, \sigma_{12}, \sigma_{22})\mathbf{1}_{\{\sigma_{11} > \sigma_{22}\}}$ for some continuous function f ensures that the causal effect on C_2 is 0. In order to fulfil the continuity requirement and the correct representation of the causal effect on C_1, f has to satisfy

$$f(\sigma_{11}, \sigma_{12}, \sigma_{22}) = 0 \qquad \text{for } \sigma_{11} = \sigma_{22}, \tag{2.15}$$

$$f(\sigma_{11}, \sigma_{12}, \sigma_{22}) = \sigma_{12}/\sigma_{22} \qquad \text{for } \sigma_{12}^2 + \sigma_{22}^2 = \sigma_{11}\sigma_{22}. \tag{2.16}$$

We propose

$$f(\sigma_{11},\sigma_{12},\sigma_{22}) = \operatorname{sign}(\sigma_{12})\sqrt[4]{\sigma_{12}^2(\sigma_{11}-\sigma_{22})/\sigma_{22}^3}$$

and directly see that (2.15) is fulfilled. Using (2.7), we can also verify (2.16) as

$$\begin{aligned} f(\sigma_{11},\sigma_{12},\sigma_{22}) &= \operatorname{sign}(\sigma_{12})\sqrt[4]{\sigma_{12}^2\left(\frac{\sigma_{12}^2+\sigma_{22}^2}{\sigma_{22}}-\sigma_{22}\right)/\sigma_{22}^3} \\ &= \operatorname{sign}(\sigma_{12})\sqrt[4]{\sigma_{12}^4/\sigma_{22}^4} = \sigma_{12}/\sigma_{22}. \end{aligned}$$

Moreover, we note that $\operatorname{sign}(\sigma_{12})\sqrt{|\sigma_{12}|}$ is a continuous function of σ_{12} and σ_{22} only takes positive values. Hence, f is continuous and satisfies all required conditions which demonstrates that

$$\tilde{T}_1(\sigma_{11},\sigma_{12},\sigma_{22}) = \operatorname{sign}(\sigma_{12})\sqrt[4]{\sigma_{12}^2(\sigma_{11}-\sigma_{22})/\sigma_{22}^3}\,\mathbf{1}_{\{\sigma_{11}>\sigma_{22}\}}$$

is a possible continuous extension on C.

Example 2.3. Since $\sigma_{22} \leq \sigma_{11}$ on C_1 and $\sigma_{22} \geq \sigma_{11}$ on C_2, it is reasonable to look for a continuous extension reflecting the two different models by considering $\min\{\sigma_{11},\sigma_{22}\}$. Using the definitions (2.8) and (2.9), we see

$$\min\{\sigma_{11},\sigma_{22}\}(\sigma_{11}-\sigma_{22})+\sigma_{12}^2 = \begin{cases} 2\sigma_{12}^2, & \text{if } (\sigma_{11},\sigma_{12},\sigma_{22}) \in \mathrm{C}_1, \\ 0, & \text{if } (\sigma_{11},\sigma_{12},\sigma_{22}) \in \mathrm{C}_2. \end{cases} \tag{2.17}$$

Harnessing this finding, we propose

$$\tilde{T}_2(\sigma_{11},\sigma_{12},\sigma_{22}) = \frac{\operatorname{sign}(\sigma_{12})}{\sigma_{22}}\sqrt[4]{\frac{\sigma_{12}^2}{2}\left|\min\{\sigma_{11},\sigma_{22}\}(\sigma_{11}-\sigma_{22})+\sigma_{12}^2\right|}$$

as extension of the causal effect on C. $\tilde{T}_2$ is well-defined as only non-negative values occur in the root and positive values in the denominator. In addition, $\tilde{T}_2$ is continuous as a composition of continuous functions and we can use (2.17) to verify that $\tilde{T}_2$ takes the value σ_{12}/σ_{22} on C_1 and 0 on C_2. For this reason, $\tilde{T}_2$ is a valid continuous extension of the causal effect on C.

2.3.3 Asymptotic distribution

The asymptotic validity of both bootstrapping and subsampling relies on the assumption that the distribution of the root (2.14) converges to a non-degenerate limit law. In order to establish such an asymptotic behaviour, we first examine the convergence of the estimator $\widehat{\Sigma}$. Assuming finite fourth moments, which is equivalent to centred fourth moments in our scenario, we define

$$\sigma_{jklh} = \mathrm{E}\left[X_j X_k X_l X_h\right]$$

and denote its estimator

$$\hat{\sigma}_{jklh} = \frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} X_l^{(i)} X_h^{(i)}.$$

(Steiger and Hakstian 1982) examine the asymptotic normality of the empirical covariance estimator in great generality. For our case their work yields

$$\sqrt{n}\left((\hat{\sigma}_{11},\hat{\sigma}_{12},\hat{\sigma}_{22})^T-(\sigma_{11},\sigma_{12},\sigma_{22})^T\right)\xrightarrow{\mathrm{D}}\mathcal{N}(0,\Gamma), \quad \text{as } n\to\infty, \tag{2.18}$$

with

$$\Gamma=\begin{pmatrix}\sigma_{1111}-\sigma_{11}^2 & \sigma_{1112}-\sigma_{11}\sigma_{12} & \sigma_{1122}-\sigma_{11}\sigma_{22}\\ \sigma_{1112}-\sigma_{11}\sigma_{12} & \sigma_{1122}-\sigma_{12}^2 & \sigma_{1222}-\sigma_{12}\sigma_{22}\\ \sigma_{1122}-\sigma_{11}\sigma_{22} & \sigma_{1222}-\sigma_{12}\sigma_{22} & \sigma_{2222}-\sigma_{22}^2\end{pmatrix}.$$

In the following, we investigate the asymptotic distributions of $\tilde{T}_1$ and $\tilde{T}_2$ in order to find sequences $(\tau_n)_{n\in\mathbb{N}}$ such that the root defined in (2.14) converges to a non-degenerate random variable.

Example 2.2 (Continued). We distinguish three cases for the true $(\sigma_{11},\sigma_{12},\sigma_{22})$ according to (2.8) – (2.10).

First case: $(\sigma_{11},\sigma_{12},\sigma_{22})\in \mathrm{C}_1\setminus\mathrm{C}_3$. Since $\sigma_{11}>\sigma_{22}$ and f is continuously differentiable, so is $\tilde{T}_1$. Therefore, (2.18) allows us to apply the delta method, Theorem 2.23, to $\tilde{T}_1$ which yields

$$\sqrt{n}\left(\tilde{T}_1(\widehat{\Sigma}_n)-\tilde{T}_1(\Sigma)\right)\xrightarrow{\mathrm{D}}\mathcal{N}\left(0,\sigma_\infty^2\right), \quad \text{as } n\to\infty,$$

where $\sigma_\infty^2=(\nabla f(\Sigma))^T\Gamma(\nabla f(\Sigma))$. Detailed calculations can be found in the Appendix 2.A.2. Consequently, we find that $\tau_n=\mathcal{O}(n^{1/2})$.

Second case: $(\sigma_{11},\sigma_{12},\sigma_{22})\in \mathrm{C}_2\setminus\mathrm{C}_3$. In this case, $\sigma_{11}<\sigma_{22}$ holds which induces $\tilde{T}_1(\Sigma)=0$. We prove that $\tau_n(\tilde{T}_1(\widehat{\Sigma}_n)-\tilde{T}_1(\Sigma))\to 0$ in distribution as $n\to\infty$.
To this end, let g be a continuous, bounded, $\mathbb{R}$-valued function on the real numbers and $(\tau_n)_{n\in\mathbb{N}}$ be an arbitrary real-valued sequence. According to the strong law of large numbers, $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ converge a.s. to σ_{11} and σ_{22} respectively implying $\mathbb{P}(\hat{\sigma}_{11}\le\hat{\sigma}_{22})\to 1$. Therefore, we can estimate

$$\left|\mathrm{E}\left[g\left(\tau_n f(\hat{\sigma}_{11},\hat{\sigma}_{12},\hat{\sigma}_{22})\right)\mathbf{1}_{\{\hat{\sigma}_{11}>\hat{\sigma}_{22}\}}\right]\right|\le\mathbb{P}(\hat{\sigma}_{11}>\hat{\sigma}_{22})\mathrm{E}\left[\|g\|_\infty\right]\to 0, \quad \text{as } n\to\infty,$$

as g is bounded. Using this interim result, we find

$$\begin{aligned}&\mathrm{E}\left[g\left(\tau_n(\tilde{T}_1(\widehat{\Sigma})-\tilde{T}_1(\Sigma))\right)\right]=\mathrm{E}\left[g\left(\tau_n\tilde{T}_1(\widehat{\Sigma})\right)\right]=\mathrm{E}\left[g\left(\tau_n f(\hat{\sigma}_{11},\hat{\sigma}_{12},\hat{\sigma}_{22})\right)\mathbf{1}_{\{\hat{\sigma}_{11}>\hat{\sigma}_{22}\}}\right)\right]\\ &=g(0)\cdot\mathbb{P}(\hat{\sigma}_{11}\le\hat{\sigma}_{22})+\mathrm{E}\left[g\big(\tau_n f(\hat{\sigma}_{11},\hat{\sigma}_{12},\hat{\sigma}_{22})\big)\mathbf{1}_{\{\hat{\sigma}_{11}>\hat{\sigma}_{22}\}}\right]\\ &\to g(0)\cdot 1+0=g(0), \quad \text{as } n\to\infty.\end{aligned}$$

Hence, the root converges to 0 in distribution, regardless of the chosen $(\tau_n)_{n\in\mathbb{N}}$, and is consequently degenerate in the limit.

Third case: $(\sigma_{11},\sigma_{12},\sigma_{22})\in\mathrm{C}_3$. In order to investigate the asymptotic distribution of the root (2.14), we define the helper function h as

$$h(\hat{\sigma}_{11},\hat{\sigma}_{12},\hat{\sigma}_{22}):=\operatorname{sign}(\hat{\sigma}_{12})\sqrt[4]{\hat{\sigma}_{12}^2|\hat{\sigma}_{11}-\hat{\sigma}_{22}|}\mathbf{1}_{\{\hat{\sigma}_{11}>\hat{\sigma}_{22}\}}.$$

Clearly, it is continuous in all arguments and, as $\sigma_{12} = 0$ and $\sigma_{11} = \sigma_{22}$, we can perform the manipulations

$$\begin{aligned} n^{3/8}\, h(\hat\sigma_{11}, \hat\sigma_{12}, \hat\sigma_{22}) &= \operatorname{sign}(\hat\sigma_{12}) \sqrt[4]{\hat\sigma_{12}^2 \left(\sqrt{n}\right)^3 |\hat\sigma_{11} - \hat\sigma_{22}|}\, \mathbf{1}_{\{\hat\sigma_{11} > \hat\sigma_{22}\}} \\ &= \operatorname{sign}\left(\sqrt{n}(\hat\sigma_{12} - \sigma_{12})\right) \sqrt[4]{\left(\sqrt{n}(\hat\sigma_{12} - \sigma_{12}\right)^2 \left|\left(\sqrt{n}(\hat\sigma_{11} - \sigma_{11})\right) - \sqrt{n}\,(\hat\sigma_{22} - \sigma_{22}))\right|} \\ &\quad \mathbf{1}_{\{\sqrt{n}(\hat\sigma_{11} - \sigma_{11}) > \sqrt{n}(\hat\sigma_{22} - \sigma_{22})\}} \\ &= h\left(\sqrt{n}(\hat\sigma_{11} - \sigma_{11}), \sqrt{n}(\hat\sigma_{12} - \sigma_{12}), \sqrt{n}(\hat\sigma_{22} - \sigma_{22})\right). \end{aligned}$$

Therefore, we can apply the continuous mapping theorem 2.24, in the following abbreviated as CMT, which yields

$$n^{3/8}\, h(\widehat{\Sigma}_n) = h\big(\sqrt{n}\,(\widehat{\Sigma}_n - \Sigma)\big) \xrightarrow{\text{D}} h\big(\mathcal{N}\,(0, \Gamma)\,\big), \quad \text{as } n \to \infty.$$

Moreover, $\hat\sigma_{22} \xrightarrow{\text{P}} \sigma_{22}$ holds due to the weak law of large numbers and the CMT 2.24 proves that $\hat\sigma_{22}^{-3/4} \xrightarrow{\text{P}} \sigma_{22}^{-3/4} \neq 0$ as well.
Combining the results for h and $\hat\sigma_{22}$ with Slutsky's Theorem 2.25, we see

$$n^{3/8}\, \tilde T_1(\widehat{\Sigma}_n) = n^{3/8} \hat\sigma_{22}^{-3/4} h(\widehat{\Sigma}_n) \xrightarrow{\text{D}} \sigma_{22}^{-3/4} h\big(\mathcal{N}\,(0, \Gamma)\,\big), \quad \text{as } n \to \infty.$$

As continuous, non-constant function of a Gaussian random variable the limit law is indeed non-degenerate. Since $\tilde T_1(\Sigma) = 0$, we have derived that (2.14) can only have a valid limit distribution if $\tau_n = \mathcal{O}(n^{3/8})$.

Summarising the upper findings, we have shown that $(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}_1 \setminus \mathrm{C}_3$ demands $\tau_n = \mathcal{O}(n^{1/2})$ and C_3 requires $\tau_n = \mathcal{O}(n^{3/8})$, whereas on $\mathrm{C}_2 \setminus \mathrm{C}_3$ no sequence can fulfil the condition of a non-degenerate limit law. Since the true value of Σ is unknown, we cannot choose a valid sequence $(\tau_n)_{n \in \mathbb{N}}$ which renders a resampling approach based on $\tilde T_1$ and the root (2.14) theoretically unjustified.

Example 2.3 (Continued). Commensurate with the first example, we distinguish three cases for $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ according to (2.8) – (2.10).

First case: $(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}_1 \setminus \mathrm{C}_3$. $\tilde T_2$ is continuously differentiable on $\mathrm{C}_1 \setminus \mathrm{C}_3$ because (2.17) implies that only positive values enter the absolute value expression of $\tilde T_2$ which is the only potential source for non-differentiability. Therefore, we can apply the delta method 2.23 using (2.18) and obtain

$$\sqrt{n}\left(\tilde T_2(\widehat{\Sigma}_n) - \tilde T_2(\Sigma)\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, \sigma_\infty^2\right), \quad \text{as } n \to \infty,$$

where $\sigma_\infty^2 = (\nabla \tilde T_2(\Sigma))^T\, \Gamma(\nabla \tilde T_2(\Sigma))$. Detailed calculations can be found in the Appendix 2.A.2. Consequently, we find that $\tau_n = \mathcal{O}(n^{1/2})$.

Second case: $(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}_2 \setminus \mathrm{C}_3$. Investigating the limit distribution, we restrict our attention to the set $\{\hat\sigma_{11} \leq \hat\sigma_{22}\}$ which implies $\min\{\hat\sigma_{11}, \hat\sigma_{22}\} = \hat\sigma_{11}$. This is justified by the almost sure convergences $\hat\sigma_{11} \to \sigma_{11}$ and $\hat\sigma_{22} \to \sigma_{22}$, according to the strong law of large numbers, and $\sigma_{11} < \sigma_{22}$ which leads to $\mathbb{P}(\hat{\sigma}_{11} \leq \hat{\sigma}_{22}) \to 1$ as $n \to \infty$. As in the third case of the first example, we define a continuous helper function by

$$h(x_1, x_2, x_3, x_4) := \sqrt[4]{\big|x_1 - x_2 - x_3 + x_4\big|}.$$

Making use of the relationship $\sigma_{11}\sigma_{22} - \sigma_{11}^2 - \sigma_{12}^2 = 0$, we rewrite parts of $\tilde{T}_2$ as follows:

$$\begin{aligned}
n^{1/8} \sqrt[4]{\big|\hat{\sigma}_{11}(\hat{\sigma}_{11} - \hat{\sigma}_{22}) + \hat{\sigma}_{12}^2\big|} &= n^{1/8} \Big[\big|(\hat{\sigma}_{11}^2 - \sigma_{11}^2) - (\hat{\sigma}_{11}\hat{\sigma}_{22} - \sigma_{11}\sigma_{22}) + (\hat{\sigma}_{12}^2 - \sigma_{12}^2)\big|\Big]^{1/4} \\
&= n^{1/8} \Big[\big|(\hat{\sigma}_{11} - \sigma_{11})(\hat{\sigma}_{11} + \sigma_{11}) - \hat{\sigma}_{22}(\hat{\sigma}_{11} - \sigma_{11}) \\
&\qquad - \sigma_{11}(\hat{\sigma}_{22} - \sigma_{22}) + (\hat{\sigma}_{12} - \sigma_{12})(\hat{\sigma}_{12} + \sigma_{12})\big|\Big]^{1/4} \\
&= h\Big((\hat{\sigma}_{11} + \sigma_{11})\sqrt{n}(\hat{\sigma}_{11} - \sigma_{11}), \hat{\sigma}_{22}\sqrt{n}(\hat{\sigma}_{11} - \sigma_{11}), \\
&\qquad \sigma_{11}\sqrt{n}(\hat{\sigma}_{22} - \sigma_{22}), (\hat{\sigma}_{12} + \sigma_{12})\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12})\Big).
\end{aligned}$$

The weak law of large numbers implies $\hat{\sigma}_{11} + \sigma_{11} \to 2\sigma_{11}$, $\hat{\sigma}_{22} \to \sigma_{22}$ and $\hat{\sigma}_{12} + \sigma_{12} \to 2\sigma_{12}$ in probability. Further, (2.18) and Slutsky's theorem 2.25 prove that the arguments of h converge to a Gaussian random variable $\mathcal{N}(0, \bar{\Gamma})$ in distribution.
Besides, the weak law of large numbers and the CMT 2.24 show that

$$2^{-1/4} \operatorname{sign}(\hat{\sigma}_{12})\sqrt{|\hat{\sigma}_{12}|}\hat{\sigma}_{22}^{-1} \xrightarrow{\text{P}} 2^{-1/4} \operatorname{sign}(\sigma_{12})\sqrt{|\sigma_{12}|}\sigma_{22}^{-1}, \quad \text{as } n \to \infty.$$

Finally, according to Slutsky's theorem 2.25, we find

$$\begin{aligned}
n^{1/8}\, \tilde{T}_2(\widehat{\Sigma}_n) &= 2^{-1/4} \operatorname{sign}(\hat{\sigma}_{12})\sqrt{|\hat{\sigma}_{12}|}\ \hat{\sigma}_{22}^{-1}\, n^{1/8} \sqrt[4]{\big|\hat{\sigma}_{11}(\hat{\sigma}_{11} - \hat{\sigma}_{22}) + \hat{\sigma}_{12}^2\big|} \\
&\xrightarrow{\text{D}} 2^{-1/4} \operatorname{sign}(\sigma_{12})\sqrt{|\sigma_{12}|}\ \sigma_{22}^{-1}\, h(\mathcal{N}(0, \bar{\Gamma})), \quad \text{as } n \to \infty.
\end{aligned}$$

Since $\tilde{T}_2(\Sigma) = 0$ on $\mathrm{C}_2 \setminus \mathrm{C}_3$, we have shown that τ_n in (2.14) has to be of order $\mathcal{O}(n^{1/8})$ to obtain a non-degenerate limit law.

Third case: $(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}_3$. Again, we define a continuous helper function

$$h(x_1, x_2, x_3, x_4) := \operatorname{sign}(x_1) \sqrt[4]{\frac{1}{2}x_1^2 \,\big|x_2 - x_3 + x_4\big|}.$$

In the following, we heavily use $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} = 0$ to derive

$$\begin{aligned}
&n^{3/8}\, \tilde{T}_2(\hat{\sigma}_{11}, \hat{\sigma}_{12}, \hat{\sigma}_{22}) \\
&\quad = \hat{\sigma}_{22}^{-1} \operatorname{sign}\left(\frac{\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12})}{\hat{\sigma}_{22}^2}\right) \Bigg[\frac{1}{2}\big(\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12})\big)^2 \\
&\qquad\quad \times \Big|\min\{\hat{\sigma}_{11}, \hat{\sigma}_{22}\}\sqrt{n}\big((\hat{\sigma}_{11} - \sigma_{11}) - (\hat{\sigma}_{22} - \sigma_{22})\big) + \hat{\sigma}_{12}\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12})\Big|\Bigg]^{1/4} \\
&\quad = h\Big(\hat{\sigma}_{22}^{-2}\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12}), \min\{\hat{\sigma}_{11}, \hat{\sigma}_{22}\}\sqrt{n}(\hat{\sigma}_{11} - \sigma_{11}), \\
&\qquad\quad \min\{\hat{\sigma}_{11}, \hat{\sigma}_{22}\}\sqrt{n}(\hat{\sigma}_{22} - \sigma_{22}), \hat{\sigma}_{12}\sqrt{n}(\hat{\sigma}_{12} - \sigma_{12})\Big).
\end{aligned}$$

The weak law of large numbers and the CMT 2.24 establish $\hat{\sigma}_{22}^{-2} \to \sigma_{22}^{-2}$, $\min\{\hat{\sigma}_{11}, \hat{\sigma}_{22}\} \to \min\{\sigma_{11}, \sigma_{22}\}$ and $\hat{\sigma}_{12} \to \sigma_{12}$ in probability. Combining these findings with (2.18), Slutsky's theorem 2.25 proves that the arguments of h converge in distribution to a Gaussian random variable $\mathcal{N}(0, \tilde{\Gamma})$. Finally, we apply the CMT 2.24 and obtain

$$n^{3/8}\, \tilde{T}_2(\widehat{\Sigma}_n) \xrightarrow{\mathrm{D}} h\big(\mathcal{N}(0, \tilde{\Gamma})\big), \quad \text{as } n \to \infty.$$

Analogously to the first example, since $\tilde{T}_2(\Sigma) = 0$ holds, the root 2.14) can only have a valid limit distribution if $\tau_n = \mathcal{O}(n^{3/8})$.

Summarising the upper findings, we have shown that $(\sigma_{11}, \sigma_{12}, \sigma_{22}) \in \mathrm{C}_1 \setminus \mathrm{C}_3$ demands $\tau_n = \mathcal{O}(n^{1/2})$, $\mathrm{C}_2 \setminus \mathrm{C}_3$ requires $\tau_n = \mathcal{O}(n^{3/8})$ and C_3 imposes $\tau_n = \mathcal{O}(n^{1/8})$. As in the first example, the choice of $(\tau_n)_{n\in\mathbb{N}}$ depends on the unknown value $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ which renders a resampling method based on $\tilde{T}_2$ and (2.14) theoretically unjustified.

We have encountered that the seemingly obvious idea of using bootstrapping or subsampling to construct confidence intervals for the causal effect of X_1 on X_2 emerges as rather intricate. Using a common choice of root, we have shown that for both continuous extensions of the causal effect from $\mathrm{C}_1 \cup \mathrm{C}_2$ to C that we proposed the convergence rates depend on the unknown value of Σ. For this reason, we cannot choose an appropriate normalising sequence $(\tau_n)_{n\in\mathbb{N}}$ for the root which deprives the presented approach of any theoretical foundation.
While we have only investigated two examples, our findings can be the starting point of a more in-depth analysis of the interplay between causal effect estimation and convergence rates.

2.4 Inverting tests

Since the approach of using resampling techniques is not successful, we develop a distinct method which relies on the dual relationship of hypothesis tests and confidence intervals. In order to construct such tests, we harness the theory of constrained statistical inference, otherwise called order restricted inference.

2.4.1 Duality of testing and confidence regions

According to (Lehmann and Romano 2005), we consider a generic statistical model $\{\mathbb{P}_{\theta,\lambda} \colon (\theta, \lambda) \in \Theta \times \Lambda\}$ and a quantity $s(\theta)$ that depends on θ. For each attainable s_0, the acceptance region of a level-α test H_0: $s(\theta) = s_0$ against H_1: $s(\theta) \neq s_0$ is denoted by $A(s_0)$. If we define a confidence region by

$$C(x) = \{s_0 : x \in A(s_0)\},$$

then

$$s_0 \in C(x) \Leftrightarrow x \in A(s_0),$$

holds and hence

$$\mathbb{P}_{\theta,\lambda}\left(s(\theta) \in C(X)\right) \geq 1 - \alpha \quad \forall\, (\theta, \lambda) \in \Theta \times \Lambda.$$

Thus, any family of level-α acceptance regions leads to a family of confidence sets of confidence level $1-\alpha$. Note that the there are no further requirements on the test procedure and it can vary for different values of s_0.

We want to make use of this flexibility and aim to develop statistical tests for all possible values of the causal effect. In this respect, the model parametrisation $(\beta, \sigma^2, m) \in \mathbb{R} \times (0,\infty) \times \{0,1\}$, where m decides on the direction of the causal relationship, seems natural. However, constrained statistical inference, as introduced in the next section, relies on the likelihood ratio statistic and assumes that all its parameters are continuous. This clearly conflicts with the discrete parameter m.

Enlarging m to a continuous parameter with values in $[0,1]$, which could be interpreted as interpolating between the two directions of the causality, is conceivable, albeit its ramifications on the model and its interpretation are not straightforward. For this reason, we use the parameter space C and express the causal effect of X_2 on X_1 as

$$\frac{\sigma_{12}}{\sigma_{22}} \mathbf{1}_{\{\sigma_{11}>\sigma_{22}\}}. \tag{2.19}$$

In this setting we neglect some of the given structure and thus loose power, but likelihood ratio methods to construct test statistics become applicable.

2.4.2 Constrained statistical inference

The theory of constrained statistical inference is briefly explained here relying on (Silvapulle and Sen 2005). It makes use of likelihood ratio statistics and consequently requires further assumptions compared to the initially distribution-free setting of the considered two-variable LSEM. We introduce relevant results in this subsection and apply the theory in the following two.

Definition 2.13. Let $X^{(1)}, \dots, X^{(n)}$ be independently and identically distributed random variables with common probability density function $f(x;\theta), \theta \in \Theta \subseteq \mathbb{R}^p$, where x can be uni- or multivariate. The *log-likelihood* and the entries of the *Fisher information matrix* for one observation are defined by

$$\begin{aligned} \ell_n(\theta) &= \sum_{i=1}^{n} \log f(X^{(i)};\theta), \\ [\mathcal{I}(\theta)]_{i,j} &= \mathrm{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X;\theta) \right) \middle| \theta \right], \quad i,j \in \{1,\dots,p\}, \end{aligned}$$

respectively. Let $\Theta^* \subset \Theta$. The *maximum likelihood estimator* $\hat{\theta}_{\Theta^*}$ is given by

$$\hat{\theta}_{\Theta^*} = \operatorname*{argmax}_{\theta \in \Theta^*} \ell_n(\theta).$$

We require the following regularity conditions.

Assumptions 2.14. *In the setting of Definition 2.13, the following is assumed.*

1. *$\hat{\theta}_{\Theta^*}$ is $\sqrt{n}$-consistent whenever the true parameter θ_0 is contained in Θ^*.*

2. *Distinct values of θ correspond to distinct distributions.*

3. *The first three partial derivatives of* $\log f(x;\theta)$ *with respect to* θ *exist almost everywhere.*

4. *There exists a* $G(y)$ *such that* $\int G(y)dy < \infty$ *and the absolute values of the first three partial derivatives of* $\log f(x;\theta)$ *with respect to* θ *are bounded by* $G(y)$ *in a neighbourhood of* θ_0.

5. *The Fisher information matrix* $\mathcal{I}(\theta)$ *is finite and positive definite.*

These conditions are not minimal but facilitate developing the theory without concerning oneself with technical details.

Definition 2.15. Assume the framework stated above and let $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ be nested models. The *likelihood ratio statistic* λ_n for a sample of size n is defined as

$$\lambda_n = 2\left(\sup_{\theta\in\Theta_1} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n(\theta)\right) = 2\left(\ell_n(\hat{\theta}_{\Theta_1}) - \ell_n(\hat{\theta}_{\Theta_0})\right).$$

In the well-studied case where Θ_0 is a linear subspace and $\Theta_1 = \Theta$, that is the general alternative, the limiting distribution of the likelihood ratio statistic is known.

Theorem 2.16. *Let* $R \in \mathbb{R}^{m\times p}$ *and* $r = \operatorname{rank}(R)$. *Testing* $H_0 : R\theta = 0$ *against* $H_1 : \theta \in \Theta$ *implies* $\lambda_n \xrightarrow{D} \chi^2_r$ *as* $n \to \infty$ *under the null hypothesis, where* χ^2_d *denotes the Chi-squared distribution with* d *degrees of freedom.*

If the hypothesis to be tested cannot be represented as a linear space, the limiting distribution of λ_n is more intricate. It involves the local geometry at θ_0, which is expressed in the tangent cone and its regularity.

Definition 2.17. Let $\Theta^* \subseteq \mathbb{R}^p$ and $\theta_0 \in \Theta^*$. The *tangent cone* to Θ^* at θ_0, denoted by $\mathcal{T}(\Theta^*;\theta_0)$, is the set of all vectors w for which a sequence of positive numbers $(t_n)_{n\in\mathbb{N}}$ converging to zero and $(\theta_n)_{n\in\mathbb{N}} \subset \Theta^*$ converging to θ_0 exist such that

$$t_n^{-1}(\theta_n - \theta_0) \to w, \quad \text{as } n \to \infty. \tag{2.20}$$

$\mathcal{T}(\Theta^*;\theta_0)$ is *Chernoff-regular* if for all its elements and all such $(t_n)_{n\in\mathbb{N}}$ a corresponding $(\theta_n)_{n\in\mathbb{N}}$ can be found such that (2.20) holds.

Remark 2.4. The tangent cone $\mathcal{T}(\Theta^*;\theta_0)$ is closed and a cone in the sense that, if $w \in \mathcal{T}(\Theta^*;\theta_0)$ holds, then $\lambda w \in \mathcal{T}(\Theta^*;\theta_0)$ for all $\lambda > 0$.

The importance of the concept of Chernoff-regularity was first discovered by (Chernoff 1954) and was later interpreted in the context of different definitions of approximating cones, see e.g. (Geyer 1994). While it is not immediately obvious how the upper definition can be verified for a given hypothesis Θ^*, (Drton 2009) establishes Chernoff-regularity for a wide class of spaces.

Lemma 2.18. *If* $\Theta^* \subseteq \mathbb{R}^p$ *is a semi-algebraic set, i.e. a finite union of sets defined by polynomial equations and inequalities, then* Θ^* *is Chernoff-regular everywhere.*

Furthermore, under the Mangasarian-Fromowitz constraint qualification (MF-CQ) and continuous differentiability we can directly compute $\mathcal{T}(\Theta^*;\theta_0)$ with the following proposition.

Proposition 2.19. *Suppose that $\Theta \subseteq \mathbb{R}^p$ is open and let Θ^* be given by*

$$\Theta^* = \{\theta \in \Theta \colon h_1(\theta) = \ldots = h_l(\theta) = 0, h_{l+1}(\theta) \geq 0, \ldots, h_k(\theta) \geq 0\},$$

where $h_1, \ldots, h_k$ are continuously differentiable. Let $\theta_0 \in \Theta^*$ and let $a_i = (\partial/\partial\theta)h_i(\theta_0)$ for $i = 1, \ldots, k$ and $J(\theta_0) = \{i \colon h_i(\theta_0) = 0, l+1 \leq i \leq k\}$. Assume that the MF-CQ is satisfied at θ_0, i.e. there exists a non-zero $b \in \mathbb{R}^p$ such that $a_1^T b = \ldots = a_l^T b = 0$, $a_1, \ldots, a_l$ are linearly independent and $a_i^T b > 0$ for $i \in J(\theta_0)$. Then $\mathcal{T}(\Theta^*;\theta_0)$ is equal to

$$\left\{\theta \in \mathbb{R}^p \colon a_i^T\theta = 0 \ \ \forall\, i = 1, \ldots, l;\ a_i^T\theta \geq 0 \ \ \forall\, i \in J(\theta_0)\right\}.$$

Remark 2.5. Loosely speaking, a condition given by $h_j(\theta) \geq 0, j \in \{l+1, \ldots, k\}$, only effects the tangent cone if θ_0 fulfils it with equality.

We consider the general testing problem

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \text{against} \quad \mathrm{H}_1 : \theta \in \Theta_1$$

for nested models $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$. Let θ_0 be the true parameter and define the norm $\|\cdot\|$ on $\mathbb{R}^p$ as

$$\|x\| = \sqrt{x^T\mathcal{I}(\theta_0)x}, \quad \text{for } x \in \mathbb{R}^p.$$

For $\Theta^* \subseteq \mathbb{R}^p$ and $x \in \mathbb{R}^p$, we use the abbreviation

$$\|x - \Theta^*\| = \inf_{\theta\in\Theta^*} \|x - \theta\|.$$

Theorem 2.20. *Let $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ be nested models and assume that Θ is open. If the null hypothesis holds and Θ_0 is Chernoff-regular at θ_0, the distribution of the likelihood ratio statistic λ_n in the limit $n \to \infty$ is equal to*

$$\|Z - \mathcal{T}(\Theta_0;\theta_0)\|^2 - \|Z - \mathcal{T}(\Theta_1;\theta_0)\|^2, \tag{2.21}$$

where $Z \sim \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$.

Remark 2.6. Results on the asymptotic distribution of λ_n are also available when Θ is not open but require additional assumptions.

Although Theorem 2.20 provides an elegant characterisation of the limiting distribution, the testing problem remains complicated as (2.21) is dependent on the true parameter θ_0 and, in general, the distribution cannot be explicitly characterised.

Under further assumptions on Θ_0 and Θ_1, however, a closed form asymptotic distribution can be derived. We turn our attention to testing problems that use the general alternative, i.e. $\Theta_1 = \Theta$, which causes the second term in (2.21) to vanish. If Θ_0 fulfils additional conditions, (Wolak 1989) and (Silvapulle and Sen 2005) state the following result.

Theorem 2.21. *Let $h^{(1)}(\theta)$ and $h^{(2)}(\theta)$ be continuously differentiable, vector-valued functions which characterise the null hypothesis in the testing problem*

$$H_0 : \theta \in \Theta_0 = \{\theta \colon h^{(1)}(\theta) \geq 0,\, h^{(2)}(\theta) = 0\} \quad \textit{against} \quad H_1 : \theta \in \Theta \subseteq \mathbb{R}^p.$$

Assume that θ_0 lies on the boundary of Θ_0 and that the MF-CQ is fulfilled. Let a_i denote $(\partial/\partial\theta)h_i^{(1)}(\theta_0)$ and let $\{j_1, \dots, j_m\}$ denote $\{i \colon h_i^{(1)}(\theta_0) = 0\}$. Set $H^{(1)}(\theta_0) = (a_{j_1}, \dots, a_{j_m})^T$ and $m = m(\theta_0) = \operatorname{rank}(H^{(1)}(\theta_0))$. If $h^{(2)}$ is not specified, set $r = 0$; otherwise, $H^{(2)}(\theta_0) = \nabla h^{(2)}(\theta_0)$ and assume that the matrix has full row-rank $r = \operatorname{rank}(H^{(2)}(\theta_0))$. If $r + m \leq p$, then

$$\mathbb{P}(\lambda_n \geq \cdot \,|\, \theta = \theta_0) \to \sum_{i=0}^{m} w_{m-i}(m, V(\theta_0))\, \mathbb{P}(\chi^2_{r+i} \geq \cdot), \quad \textit{as } n \to \infty, \tag{2.22}$$

where

$$\begin{aligned} V(\theta_0) &= H^{(1)}(\theta_0)\mathcal{I}(\theta_0)^{-1}H^{(1)}(\theta_0)^T \\ &\quad - H^{(1)}(\theta_0)\mathcal{I}(\theta_0)^{-1}H^{(2)}(\theta_0)^T(H^{(2)}(\theta_0)\mathcal{I}(\theta_0)^{-1}H^{(2)}(\theta_0)^T)^{-1}H^{(2)}(\theta_0)\mathcal{I}(\theta_0)^{-1}H^{(1)}(\theta_0)^T. \end{aligned}$$

If $h^{(2)}$ is not specified, the second summand above is zero. $\{w_k\}_{k\in\{0,\dots,m\}}$ are positive weights such that $\sum_{k=0}^m w_k = 1$ and χ^2_d denotes a Chi-squared distribution with d degrees of freedom.

Remark 2.7. If θ_0 lies in the interior of Θ_0, the tangent cone at θ_0 is $\mathbb{R}^p$ which induces the limiting statistic (2.21) to be zero almost surely.

Remark 2.8. The distribution (2.22) is a mixture of χ^2-random variables and is also referred to as Chi-bar-squared distribution and denoted by $\bar{\chi}^2(\Theta_0, V)$. In general, its weights $w_i(q, V)$ depend on the positive definite matrix V and thus in (2.22) also on θ_0. For small q, closed form representations are known:

1. Let $q = 1$. Then
$$w_0(1, V) = w_1(1, V) = 0.5.$$

2. Let $q = 2$. Then
$$\begin{aligned} w_0(2, V) &= 0.5\,\pi^{-1}\cos^{-1}(\rho_{12}), \quad w_1(2, V) = 0.5, \\ w_2(2, V) &= 0.5 - 0.5\,\pi^{-1}\cos^{-1}(\rho_{12}), \end{aligned}$$
where ρ_{12} is the correlation coefficient $v_{12}(v_{11}v_{22})^{-1/2}$.

Explicit results for $q \in \{3, 4\}$ are known; for higher dimensions, however, the weights can only be obtained by simulation.

The limiting distribution stated in Theorem 2.21 still depends on θ_0 through $m(\theta_0)$ and $V(\theta_0)$. When conducting a test at significance level α, we consider the asymptotic distribution

$$\sup_{\theta\in\Theta_0} \sum_{i=0}^{m} w_{m-i}(m, V(\theta))\, \mathbb{P}(\chi^2_{r+i} \geq \cdot),$$

in order to secure that, regardless of θ_0, the type-I error is at maximum α. The value of θ_0 for which the supremum is attained is called least favourable null value.

2.4.3 Standard approach

Building on the two previous sections, we construct testing procedures for different values c_0 of the causal effect of X_2 on X_1. Conducting these tests for a range of values allows us to derive confidence regions covering the true causal effect at a given level $1-\alpha$.
To this end, we assume that the log-likelihood ℓ_n satisfies the regularity conditions stated in the previous section and use the cone C as defined in (2.5), which is clearly an open set, as parameter space Θ. Moreover, we only consider testing subspaces given by semi-algebraic sets against the general alternative. Therefore, we do not have to worry about Chernoff-regularity, cf. Lemma 2.18, can calculate tangent cones using Proposition 2.19 and derive asymptotic distributions for different values of the null hypothesis according to Theorem 2.21.
We use the likelihood ratio statistic

$$\lambda_n^{c_0} = 2\left(\ell_n(\widehat{\Sigma}_n) - \sup_{\Sigma \in \mathrm{H}_0^{c_0}} \ell_n(\Sigma)\right), \tag{2.23}$$

where $\mathrm{H}_0^{c_0}$ denotes the null hypothesis for testing c_0 and $\widehat{\Sigma}_n$ is the unrestricted maximum likelihood estimate. Depending on c_0 the subspace of C that is associated with the null hypothesis differs which makes it necessary to distinguish three scenarios.

First case: $|c_0| \geq 1$. A value different from 0 indicates that X_1 is the dependent variable which corresponds to $\sigma_{11} \geq \sigma_{22}$ and $\sigma_{12}/\sigma_{22} = c_0$, according to (2.6). Yet, if $|c_0| \geq 1$ holds, applying the strict Cauchy-Schwarz inequality, which is enforced by the definition of C, and inserting the equation $\sigma_{12} = c_0\,\sigma_{22}$ yields

$$\sigma_{11} > \frac{\sigma_{12}^2}{\sigma_{22}} = c_0^2\,\sigma_{22} \geq \sigma_{22}.$$

Consequently, the inequality condition is automatically fulfilled and thus not part of the null hypothesis. Hence, we test

$$\mathrm{H}_0^{c_0} : \sigma_{12} = c_0\,\sigma_{22} \quad \text{against} \quad \mathrm{H}_1 : \Sigma \in C. \tag{2.24}$$

Since the null hypothesis is a linear subspace of C, the classical theory of likelihood ratio tests applies. We use Theorem 2.16 to obtain the asymptotic distribution of the likelihood ratio statistic

$$\mathbb{P}\left(\lambda_n^{c_0} \leq \cdot\,\right) \to \mathbb{P}\left(\chi_1^2 \leq \cdot\,\right), \quad \text{as } n \to \infty.$$

Second case: $|c_0| \in (0,1)$. Similar to the first case, the constraints $\sigma_{11} \geq \sigma_{22}$ and $\sigma_{12} = c_0\,\sigma_{22}$ have to be satisfied, however, the first does not automatically hold. Hence, it is part of the null hypothesis as well. We conduct the test

$$\mathrm{H}_0^{c_0} : \sigma_{12} = c_0\,\sigma_{22},\, \sigma_{11} \geq \sigma_{22} \quad \text{against} \quad \mathrm{H}_1 : \Sigma \in C. \tag{2.25}$$

The limiting distribution of $\lambda_n^{c_0}$ depends on the tangent cone at the unknown, true $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ leading to two possible situations.
If $\sigma_{11} > \sigma_{22}$ holds, the tangent cone is $\{(x,y,z) \in \mathbb{R}^3 : y = c_0 z\}$ according to Proposition 2.19. In this case, the limiting distribution is, similar to the first case, χ_1^2.

If $\sigma_{11} = \sigma_{22}$, the tangent cone is $\{(x, y, z) \in \mathbb{R}^3 : y = c_0 z,\, x \geq z\}$. Since this is not a linear subspace but a half-space, we resort to Theorem 2.21 stating that the asymptotic distribution is a mixture of χ_1^2 and χ_2^2. The weights are 0.5 each according to Remark 2.8. The last favourable null value satisfies $\sigma_{11} = \sigma_{22}$ as the Chi-bar-squared distribution of the second case is larger than χ_1^2; hence, we obtain

$$\sup_{\Sigma \in \mathrm{H}_0^{c_0}} \mathbb{P}\left(\lambda_n^{c_0} \leq \cdot\,\right) \to \frac{1}{2}\mathbb{P}\left(\chi_1^2 \leq \cdot\,\right) + \frac{1}{2}\mathbb{P}\left(\chi_2^2 \leq \cdot\,\right), \quad \text{as } n \to \infty.$$

Third case: $c_0 = 0$. In the representation of the causal effect of X_2 on X_1 given by (2.19), a value of 0 indicates that either $\sigma_{11} \leq \sigma_{22}$ or $\sigma_{12} = 0$ holds. Considering (2.6), we see that the assumptions of a two-variable LSEM establish $\sigma_{11} = \sigma_{22} \Leftrightarrow \sigma_{12} = 0$ which implies that $\sigma_{12} = 0$ is a special case of the first inequality. Therefore, it suffices to test

$$\mathrm{H}_0^0 : \sigma_{11} \leq \sigma_{22} \quad \text{against} \quad \mathrm{H}_1 : \Sigma \in C. \tag{2.26}$$

As in the case of $|c_0| \in (0, 1)$, the limiting distribution of λ_n^0 depends on the true, unknown value of $(\sigma_{11}, \sigma_{12}, \sigma_{12})$.

If $\sigma_{11} < \sigma_{22}$, the tangent cone is just $\mathbb{R}^3$ according to Proposition 2.19 which implies that λ_n^0 is equal to zero almost surely.

If $\sigma_{11} = \sigma_{22}$ however, the tangent cone is given by $\{(x, y, z) \in \mathbb{R}^3 : x \leq z\}$. Following Theorem 2.21, the asymptotic distribution is a mixture of χ_0^2 and χ_1^2 with the weights 0.5 each, where $\chi_0^2 \equiv 0$.

Consequently, the least favourable null value clearly has to satisfy $\sigma_{11} = \sigma_{22}$ which is equivalent to

$$\sup_{\Sigma \in \mathrm{H}_0^0} \mathbb{P}\left(\lambda_n^{c_0} \leq \cdot\,\right) \to \frac{1}{2}\mathbb{P}\left(\chi_0^2 \leq \cdot\,\right) + \frac{1}{2}\mathbb{P}\left(\chi_1^2 \leq \cdot\,\right), \quad \text{as } n \to \infty.$$

Considering the intricate general asymptotic distribution of Theorem 2.21, it is remarkable that we could derive rather simple limit distributions. This is mainly due to the fact that the weights are independent of the true but unknown $(\sigma_{11}, \sigma_{22}, \sigma_{12})$ as we only deal with linear and half-spaces. This greatly facilitates our analysis and lets us avoid overly conservative bounds for (2.22).

In order to compute a confidence interval for the causal effect of X_2 on X_1 we carry out tests for different c_0-values which are equally spaced with distance h and lie in the interval $[b_{low}, b_{high}]$. For each c_0, we calculate the maximum likelihood statistic $\lambda_n^{c_0}$ and compare it to the $(1-\alpha)$-quantile of the respective limit distribution. For each of the three scenarios we collect the c_0-values that pass the test and construct intervals from them, heuristically adding the length $h/2$ at both sides. In the case $|c_0| \geq 1 \Leftrightarrow c_0 \in (-\infty, -1] \cup [1, \infty)$, we have to build an interval for positive and negative c_0-values, for $|c_0| \in (0, 1)$, only a single interval suffices and, for $c_0 = 0$, the contribution to the confidence region is either $\emptyset$ or $\{0\}$. The detailed procedure is laid out in Algorithm 1.

Since the limit distributions differ among the three scenarios and the obtained confidence region is a union of intervals, it is not guaranteed that it is in fact an interval. Later simulations show that, depending on the sample size and causal effect, holes and other discontinuities may occur frequently.

Algorithm 1: Standard approach for log-likelihood ℓ

```
Input: observations $\mathbf{X}^n$, level $\alpha$, boundaries $b_{low}, b_{high}$, step length $h$
Output: $(1-\alpha)$-confidence region
$c_0 \leftarrow b_{low}$
$S_1, S_2, S_3 \leftarrow \emptyset$
$q_1 \leftarrow F^{-1}_{\chi^2_1}(1-\alpha)$
$q_2 \leftarrow (0.5F_{\chi^2_1} + 0.5F_{\chi^2_2})^{-1}(1-\alpha)$
$q_3 \leftarrow (0.5F_{\chi^2_0} + 0.5F_{\chi^2_1})^{-1}(1-\alpha)$
while $c_0 \leq b_{high}$ do
    if $|c_0| \geq 1$ then
        $\lambda_n^{c_0} \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{y=c_0 z} \ell_n(x,y,z))$
        if $\lambda_n^{c_0} \leq q_1$ then
            $S_1 \leftarrow S_1 \cup \{c_0\}$
    else if $|c_0| \in (0,1)$ then
        $\lambda_n^{c_0} \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{y=c_0 z,\, x \geq z} \ell_n(x,y,z))$
        if $\lambda_n^{c_0} \leq q_2$ then
            $S_2 \leftarrow S_2 \cup \{c_0\}$
    else
        $\lambda_n^{c_0} \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{x \leq z} \ell_n(x,y,z))$
        if $\lambda_n^{c_0} \leq q_3$ then
            $S_3 \leftarrow S_3 \cup \{c_0\}$
    $c_0 \leftarrow c_0 + h$
$I_1 \leftarrow [\min(S_1 \cap [b_{low}, -1]) - h/2, \max(S_1 \cap [b_{low}, -1]) + h/2]$
  $\cup[\min(S_1 \cap [1, b_{high}]) - h/2, \max(S_1 \cap [1, b_{high}]) + h/2]$
$I_2 \leftarrow [\min(S_2) - h/2, \max(S_2) + h/2]$
return $I_1 \cup I_2 \cup S_3$
```

2.4.4 Two-step approach

In the previous section, we used the largest possible asymptotic distribution for the cases $|c_0| \in (0,1)$ and $c_0 = 0$ to ensure that the type-I error is controlled. However, if $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ is not the least favourable null value, the real limit distribution of $\lambda_n^{c_0}$ is smaller which leads to overly conservative confidence regions. Figure 1 illustrates that, even for small sample sizes and common values of β_{12} and σ^2, the limit law deviates from the least favourable null distribution considerably. For both $|c_0| \in (0,1)$ and $c_0 = 0$, the limit distribution depends on the value of $\sigma_{11} - \sigma_{22}$ as explained above.

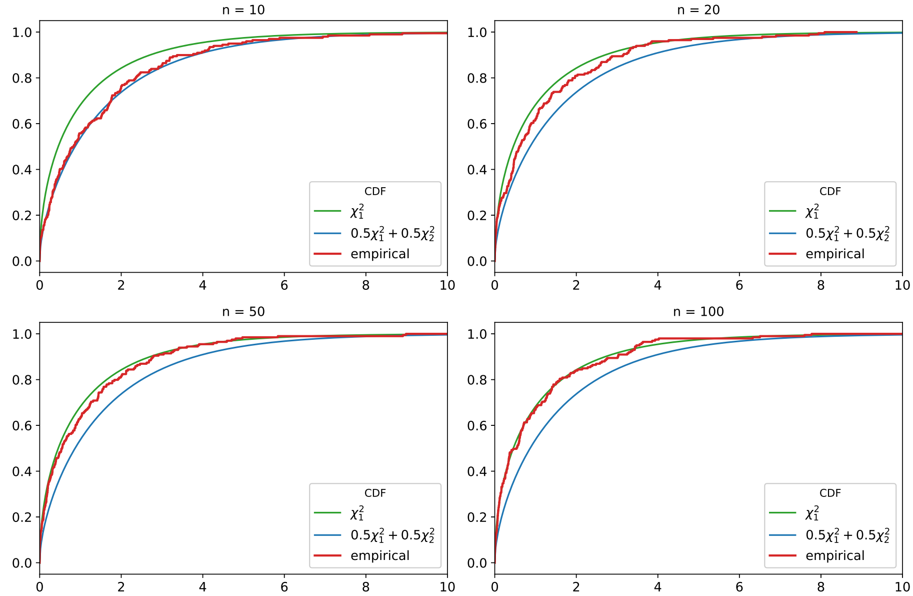

Figure 1: Empirical cdf of $\lambda_n^{c_0}$ for different sample sizes $n \in \{10, 20, 50, 100\}$. The model parameters are set to $\beta_{12} = 0.5, \sigma^2 = 1$ and for each n the empirical cdf is based on 200 simulated datasets. As n grows, the law of $\lambda_n^{c_0}$ deviates from the least favourable null distribution, $0.5\chi_1^2 + 0.5\chi_2^2$, and approaches χ_1^2.

In the following, we summarise a two-step procedure, first presented by (Silvapulle 1996), which tackles the general problem of the influence of nuisance parameters on hypothesis testing and subsequently apply the result to our testing problem.
We consider a statistical model for a random variable X parametrised by $(\theta, \lambda) \in \Theta \times \Lambda$ and the hypothesis test

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{against} \quad \mathrm{H}_1 : \theta \in \Theta^* \subseteq \Theta.$$

Let T_λ be a test statistic for a given λ, where large values of T_λ are evidence against H_0, let $t_\lambda(X)$ be the observed value and denote the true value of the nuisance parameter λ_0. Consequently, the p-value is given by

$$p_0(X) = \mathbb{P}_{\theta_0,\lambda_0}(T_{\lambda_0} \geq t_{\lambda_0}(X)|X).$$

Since λ_0, however, is unknown, we resort to

$$p(X) = \sup_{\lambda \in \Lambda} \mathbb{P}_{\theta_0,\lambda}(T_\lambda \geq t_\lambda(X)|X).$$

While rejecting H_0, if p does not exceed a certain significance level α, clearly controls the type-I error, this testing procedure might often be too conservative as the whole range of possible nuisance parameters is accounted for, regardless of the data.
(Silvapulle 1996) first estimates a $(1-\alpha_1)$-confidence region $\mathcal{C}(\alpha_1)$ for the nuisance parameter, where $0<\alpha_1<\alpha$. Second, the p-value is computed as

$$p^*(X) = \alpha_1 + \sup_{\lambda \in \mathcal{C}(\alpha_1)} \mathbb{P}_{\theta_0,\lambda}(T_\lambda \geq t_\lambda(X)|X).$$

Theorem 2.22. *For a given $\alpha \in (0,1)$ and $0<\alpha_1<\alpha$, rejecting H_0 if $p^*(X) \leq \alpha$ ensures that the type-I error does not exceed α.*

Proof. We estimate the type-I error as follows

$$\begin{aligned}\mathbb{P}_{\theta_0,\lambda_0}(p^*(X)\leq\alpha) &= \mathbb{P}_{\theta_0,\lambda_0}\left(p^*(X)\leq\alpha, \lambda_0\in\mathcal{C}(\alpha_1)\right) + \mathbb{P}_{\theta_0,\lambda_0}(p^*(X)\leq\alpha, \lambda_0\notin\mathcal{C}(\alpha_1))\\ &\leq \mathbb{P}_{\theta_0,\lambda_0}\Big(\sup_{\lambda\in\mathcal{C}(\alpha_1)} \mathbb{P}_{\theta_0,\lambda}(T_\lambda\geq t_\lambda(X)|X)\leq\alpha-\alpha_1, \lambda_0\in\mathcal{C}(\alpha_1)\Big)+\alpha_1\\ &\leq \mathbb{P}_{\theta_0,\lambda_0}\left(\mathbb{P}_{\theta_0,\lambda_0}(T_{\lambda_0}\geq t_{\lambda_0}(X)|X)\leq\alpha-\alpha_1\right)+\alpha_1\\ &= \mathbb{P}_{\theta_0,\lambda_0}(p_0(X)\leq\alpha-\alpha_1)+\alpha_1\\ &= (\alpha-\alpha_1)+\alpha_1=\alpha.\end{aligned}$$

The two inequalities are a direct consequence of the monotonicity of the probability measure and the last equality ensues from the uniform distribution of p-values, as stated in (Lehmann and Romano 2005). □

The outlined procedure improves the straightforward approach of using p as p-value because the supremum only considers values of the nuisance parameter that are realistic in view of the data. The envisaged type-I error rate α is divided up into α_1, the error probability for the confidence region of λ_0, and $\alpha-\alpha_1$, the maximal value of $\sup_{\lambda\in\mathcal{C}(\lambda_0,\alpha_1)}\mathbb{P}_\lambda(T_\lambda\geq t)$ which H_0 is rejected for.

The asymptotic distribution of $\lambda_n^{c_0}$ in the cases $|c_0|\in(0,1)$ and $c_0=0$ depends on the nuisance parameter $\mathbf{1}_{\{\sigma_{11}=\sigma_{22}\}}$ as the tangent cones at the true value $(\sigma_{11},\sigma_{12},\sigma_{22})$ are different for $\sigma_{11}=\sigma_{22}$ and $\sigma_{11}\neq\sigma_{22}$ as explained in the previous section. In order to apply the outlined two-step procedure, we first have to construct $(1-\alpha_1)$-confidence regions for the nuisance parameter. Since $\mathbf{1}_{\{\sigma_{11}=\sigma_{22}\}}$ only takes two values and the supremum is consequently only taken over two limit distributions at most, it is enough to build a confidence interval for $\sigma_{11}-\sigma_{22}$ and verify if 0 is contained.
To this end, we make use of the asymptotic normality of $\widehat{\Sigma}$, stated in (2.18), as well as the linear transformation

$$\sigma_{11}-\sigma_{22}=(1,0,-1)\,(\sigma_{11},\sigma_{12},\sigma_{22})^T$$

and obtain

$$\sqrt{n}\big((\hat{\sigma}_{11}-\hat{\sigma}_{22})-(\sigma_{11}-\sigma_{22})\big)\xrightarrow{\mathrm{D}}\mathcal{N}\left(0,\sigma_\infty^2\right),\quad \text{as } n\to\infty, \tag{2.27}$$

where

$$\sigma_\infty^2 = \sigma_{1111} + \sigma_{2222} - 2\sigma_{1122} + 2\sigma_{11}\sigma_{22} - \sigma_{11}^2 - \sigma_{22}^2. \tag{2.28}$$

Replacing the exact fourth and second moments in (2.28) by their estimates yields an estimator $\hat{\sigma}_\infty^2$ which, according to the law of large numbers, converges in probability to σ_∞^2 as $n \to \infty$. Applying this finding on (2.27), we can deduce with Slutsky's theorem 2.25 that

$$\sqrt{n}\,\frac{(\hat{\sigma}_{11} - \hat{\sigma}_{22}) - (\sigma_{11} - \sigma_{22})}{\hat{\sigma}_\infty} \xrightarrow{\text{D}} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

This allows us to construct the approximate $(1-\alpha_1)$-confidence interval

$$\left((\hat{\sigma}_{11} - \hat{\sigma}_{22}) \pm \sqrt{\frac{\hat{\sigma}_\infty^2}{n}}\, z_{1-\alpha_1/2}\right), \tag{2.29}$$

where z_γ denotes the γ-th quantile of a standard normal distribution.

If 0 is contained in the confidence interval (2.29), we have to consider both possible limit distributions. For $|c_0| \in (0,1)$ they are χ_1^2 and $0.5\chi_1^2 + 0.5\chi_2^2$, and for $c_0 = 0$ they are given by χ_0^2 and $0.5\chi_0^2 + 0.5\chi_1^2$ respectively, as explained in the previous section. Since for both $|c_0| \in (0,1)$ and $c_0 = 0$ the latter distribution is larger, we take this law as the limit distribution.

If 0 is not contained in the confidence interval (2.29), we only have to consider the limit distributions for $\sigma_{11} \neq \sigma_{22}$ which are χ_1^2 and $\chi_0^2 \equiv 0$ respectively.

Having completed the first step, we test different values of c_0 in the same way the standard approach does, yet, we use the $(1-(\alpha-\alpha_1))$-quantile of the limit distribution chosen by the first step. Algorithm 2 lays out the details of the two-step procedure.

The parameter α_1 gives an additional degree of freedom that can be used to tune the confidence intervals. However, for the case $|c_0| \in (0,1)$, the two-step procedure can only be superior to the standard approach if the $(1-\alpha)$-quantile of $0.5\chi_1^2 + 0.5\chi_2^2$ is larger than the $(1-(\alpha-\alpha_1))$-quantile of χ_1^2.

2.5 Experiments

We investigate the behaviour of the standard and two-step approach, presented in Subsection 2.4, for simulated data. Moreover, we apply Algorithm 1 and 2 to benchmark datasets, for which the direction of the causal relationship is known, in order to examine their performance on real world data.

2.5.1 Toy problem

We use the toy model

$$X_1 = \beta X_2 + \varepsilon_1, \quad X_2 = \varepsilon_2, \quad (\varepsilon_1, \varepsilon_2)^T \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}), \tag{2.30}$$

whose log-likelihood clearly fulfils the Assumptions 2.14. Sampling data for different values of σ^2 and β, we use (Kraft 1988)'s sequential least squares programming algorithm as implemented in Python's `SciPy` package as optimisation routine. Details on the parametrisation of the log-likelihood and the optimisation constraints can be found in Appendix 2.B and the source code is also available as a Jupyter Notebook on Github.

Algorithm 2: Two-step approach for log-likelihood ℓ

```
Input: observations $\mathbf{X}^n$, level $\alpha$, level $\alpha_1$ for nuisance parameter,
       boundaries $b_{low}, b_{high}$, step length $h$
Output: $(1-\alpha)$-confidence region
$c_0 \leftarrow b_{low}$
$S_1, S_2, S_3 \leftarrow \emptyset$
                                                        // 1st step
$q_1 \leftarrow F^{-1}_{\chi^2_1}(1-\alpha)$
$\hat{\sigma}^2_\infty \leftarrow \hat{\sigma}_{1111} + \hat{\sigma}_{2222} - 2\hat{\sigma}_{1122} + 2\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}^2_{11} - \hat{\sigma}^2_{22}$
if $0 \in (\hat{\sigma}_{11} - \hat{\sigma}_{22} \pm \sqrt{\hat{\sigma}^2_\infty/n}\, z_{1-\alpha_1/2})$ then
    $q_2 \leftarrow (0.5F_{\chi^2_1} + 0.5F_{\chi^2_2})^{-1}(1-(\alpha-\alpha_1))$
    $q_3 \leftarrow (0.5F_{\chi^2_0} + 0.5F_{\chi^2_1})^{-1}(1-(\alpha-\alpha_1))$
else
    $q_2 \leftarrow F^{-1}_{\chi^2_1}(1-(\alpha-\alpha_1))$
    $q_3 \leftarrow 0$
                                                        // 2nd step
while $c_0 \leq b_{high}$ do
    if $|c_0| \geq 1$ then
        $\lambda^{c_0}_n \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{y=c_0 z} \ell_n(x,y,z))$
        if $\lambda^{c_0}_n \leq q_1$ then
            $S_1 \leftarrow S_1 \cup \{c_0\}$
    else if $|c_0| \in (0,1)$ then
        $\lambda^{c_0}_n \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{y=c_0 z,\, x \geq z} \ell_n(x,y,z))$
        if $\lambda^{c_0}_n \leq q_2$ then
            $S_2 \leftarrow S_2 \cup \{c_0\}$
    else
        $\lambda^{c_0}_n \leftarrow 2(\ell_n(\widehat{\Sigma}_n) - \sup_{x \leq z} \ell_n(x,y,z))$
        if $\lambda^{c_0}_n \leq q_3$ then
            $S_3 \leftarrow S_3 \cup \{c_0\}$
    $c_0 \leftarrow c_0 + h$
$I_1 \leftarrow [\min(S_1 \cap [b_{low}, -1]) - h/2, \max(S_1 \cap [b_{low}, -1]) + h/2]$
  $\cup[\min(S_1 \cap [1, b_{high}]) - h/2, \max(S_1 \cap [1, b_{high}]) + h/2]$
$I_2 \leftarrow [\min(S_2) - h/2, \max(S_2) + h/2]$
return $I_1 \cup I_2 \cup S_3$
```

Since σ^2 appears as scaling factor, it does not qualitatively affect the variety of possible scenarios. On this account, we set $\sigma^2 = 1$ throughout all simulations and further fix $\alpha = 0.05$. We consider the standard and two-step approach, for the latter of which we use $\alpha_1 \in \{0.01, 0.02\}$. Moreover, we only need to examine non-negative values of β due to symmetry.

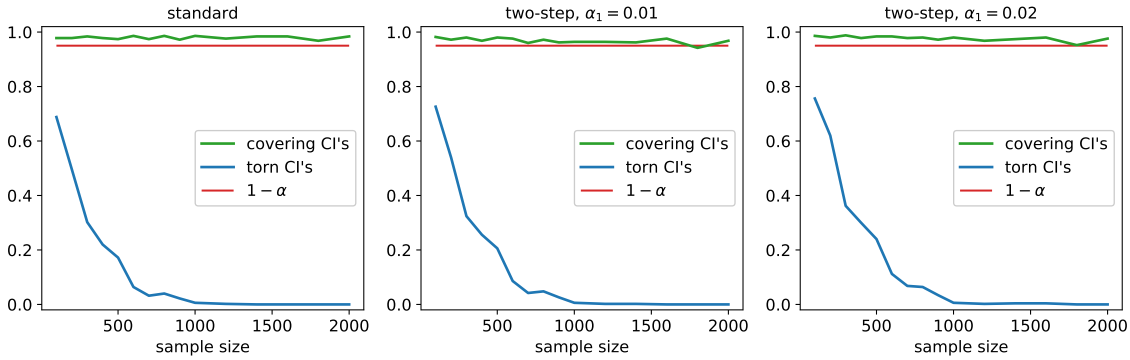

Figure 2: Share of covering and torn confidence intervals for $\beta = 0.5$ and $X_1 \leftarrow X_2$. The percentage of instances where the confidence interval covers the causal effect and where it is discontinuous (torn) are calculated on 500 simulated datasets. Potential values c_0 were tested with step length 0.01 in the range $[-1, 2]$.

First, we study the behaviour of the proposed approaches for a value of the causal effect that is away from the critical points $-1, 0$ and 1 where the least favourable null distribution changes. Therefore, we set the direction of causality as $X_1 \leftarrow X_2$ and run a simulation for $\beta = 0.5$ whose results are depicted in Figure 2. We see that, even for small sample sizes, all approaches yield slightly conservative confidence intervals; the two-step approach with $\alpha_1 = 0.01$, however, comes closest to the envisaged coverage of $1 - \alpha$. We further observe that, while discontinuous confidence intervals are common for small sample sizes, their occurrence rapidly decreases as n grows.

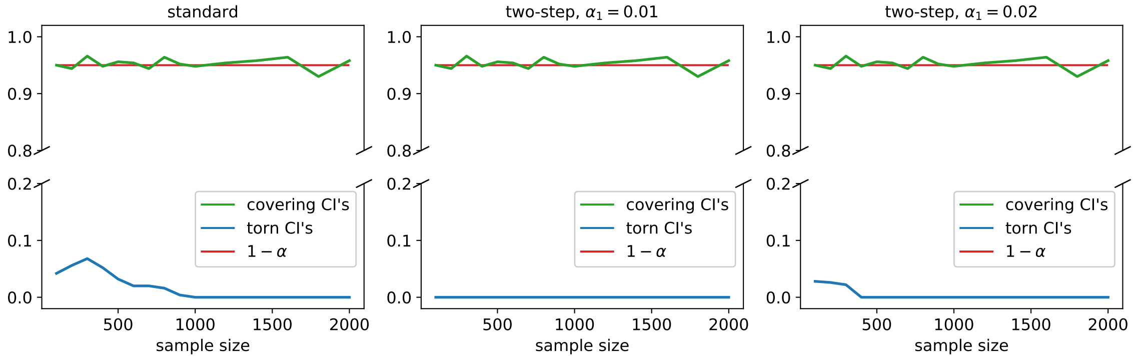

Figure 3: Share of covering and torn confidence intervals for $\beta = 1.1$ and $X_1 \leftarrow X_2$. 500 datasets were simulated to calculate the respective shares and potential values c_0 were tested in the range $[-0.4, 2.6]$ with step length 0.01.

Next, we consider values of the causal effect that are close to 0 and 1 where we expect discontinuous, or rather torn, confidence intervals to be found more frequently. Figure 3 exhibits that all approaches achieve the desired coverage and torn confidence intervals are

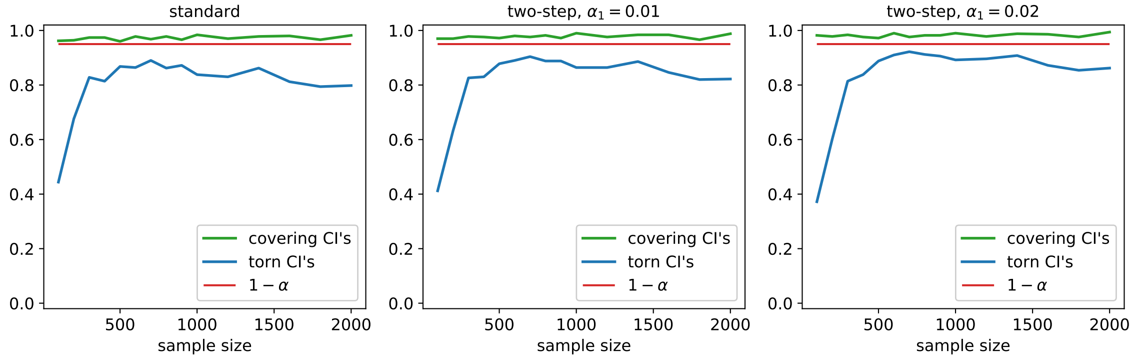

Figure 4: Share of covering and torn confidence intervals for $\beta = 0.2$ and $X_1 \leftarrow X_2$. 500 datasets were simulated to calculate the respective shares and potential values c_0 were tested in the range $[-1.3, 1.7]$ with step length 0.01.

rare. This finding is corroborated by simulations for $\beta = 1$ and $\beta = 0.9$. Contrary to this, Figure 4 shows that the percentage of discontinuous confidence intervals for small $|\beta|$ sharply rises and stays at a high level, even for larger sample sizes. Additionally, we undertook simulations for $\beta = 0.1$, which exhibits a slower growth to the peak level, and for $\beta = 0.3$, where a decline of the share of torn confidence intervals can be observed for smaller sample sizes than for $\beta = 0.2$. These phenomena can be traced back to the likelihood of inclusion of $\{0\}$ into the confidence interval which causes a potential discontinuity. Therefore, the share of torn confidence intervals eventually decreases, albeit this may only occur for very large sample sizes.

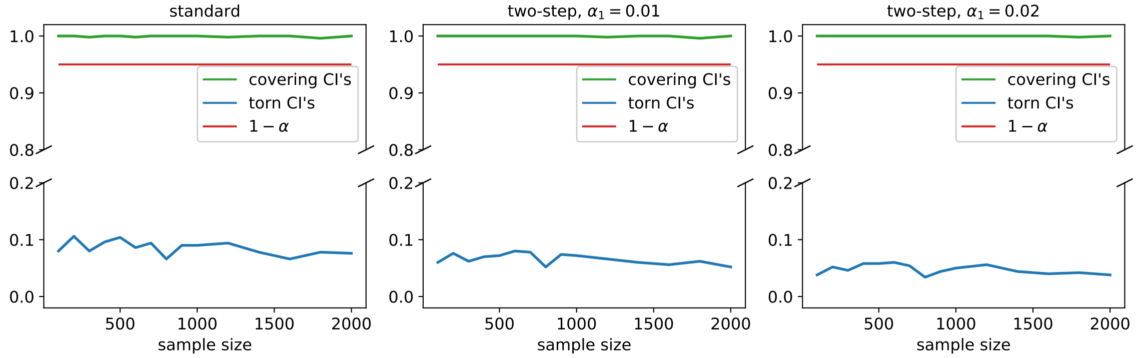

Figure 5: Share of covering and torn confidence intervals for $\beta = 0$. 500 datasets were simulated to calculate the respective shares and potential values c_0 were tested in the range $[-1.5, 1.5]$ with step length 0.01.

Moreover, we conduct an experiment for independent random variables X_1 and X_2 which is equivalent to $\beta = 0$. As illustrated by Figure 5, the occurrence of torn confidence intervals stays below 10% throughout all tested approaches and sample sizes while all calculated confidence intervals are clearly conservative with an almost constant coverage of 100%.

Finally, we examine the performance of the confidence intervals for the causal effect of X_2 and X_1 when the true causality is directed opposite. To this end, we set $\beta = 0.5$ and switch the roles of X_1 and X_2 in (2.30). Figure 6 illustrates that the share of torn

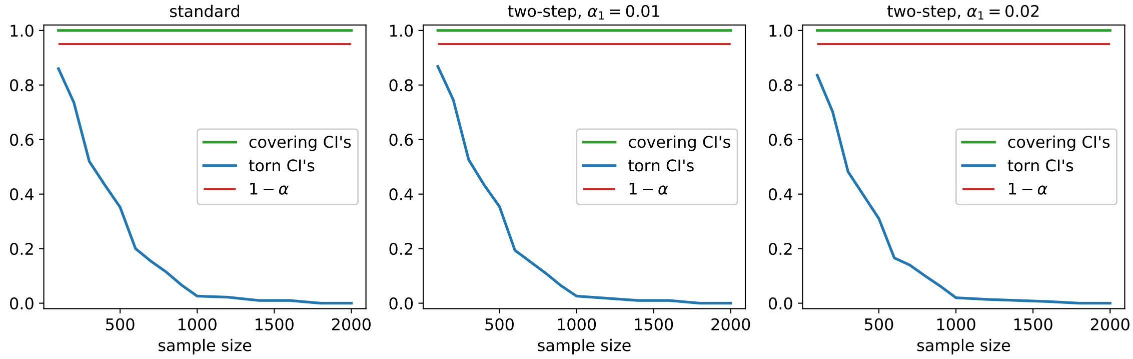

Figure 6: Share of covering and torn confidence intervals for $\beta = 0.5$ and $X_1 \to X_2$. The percentage of instances where the confidence interval covers the causal effect and where it is discontinuous (torn) are calculated on 500 simulated datasets. Potential values c_0 were tested with step length 0.01 in the range $[-1, 2]$.

confidence intervals decreases analogously to Figure 2 while the coverage stays constantly at 100% as in Figure 5. Simulations for $\beta = 0.2$ and $\beta = 0.1$ corroborate that the share of torn confidence intervals behaves similar to the corresponding experiments with $X_1 \leftarrow X_2$ and that the obtained confidence intervals are conservative.

In summary, we find that both the standard and two-step approach show a similar performance yielding valid, often conservative confidence intervals which are achieved even for small sample sizes. Moreover, the likelihood of obtaining a discontinuous confidence interval highly depends on the value of β and can be considerable even for large sample sizes.

2.5.2 Benchmarks

In order to evaluate the performance of the standard approach and the two-step approach for different values of α_1 on real-world data, we use the CauseEffectPairs dataset presented in (Mooij et al. 2016). It consists of data for 100 different cause-effect pairs from various fields for which the "ground-truth" causal directions were determined by domain knowledge. The source code for the following inference is available on Github.

We specifically consider pairs that render the assumptions of a linear relationship as well as the equal variance assumption plausible. In particular, we use the pairs `0066` and `0067`, alias `stock1` and `stock2`, which contain the daily stock returns of companies and indices. In the former pair, one enterprise holds a large portion of another company; in the latter pair, one stock is a typical representative in a subindex. Moreover, we consider the pairs `0089` and `0090`, alias `root1` and `root2`, which describe the degree of root decomposition in forests and grasslands respectively six months and one year after the start of observation. Furthermore, we use pair `0076`, alias `food`, containing the average annual rate of change of population and total dietary consumption. All of these pairs account for the size of the observed quantities by considering shares instead of absolute values. Hence, assuming equal variance of the error terms is justified.

Having identified valid pairs, we centre the data and scale it such that the causal effect among the different pairs is of comparable size. For each pair, we scale both X_1 and X_2 by the inverse of the standard deviation of X_1. It is crucial to use the same scaling factor

as we could not identify the direction of the causal relationship otherwise.
Since the approach of inverting tests requires the knowledge of the underlying distribution, we need to find a reasonable suggestion. The probability plot in Figure 7 shows that the ordered data points match the theoretical quantiles of a Gaussian distribution for all pairs in question. Hence, assuming an underlying normal distribution is plausible.

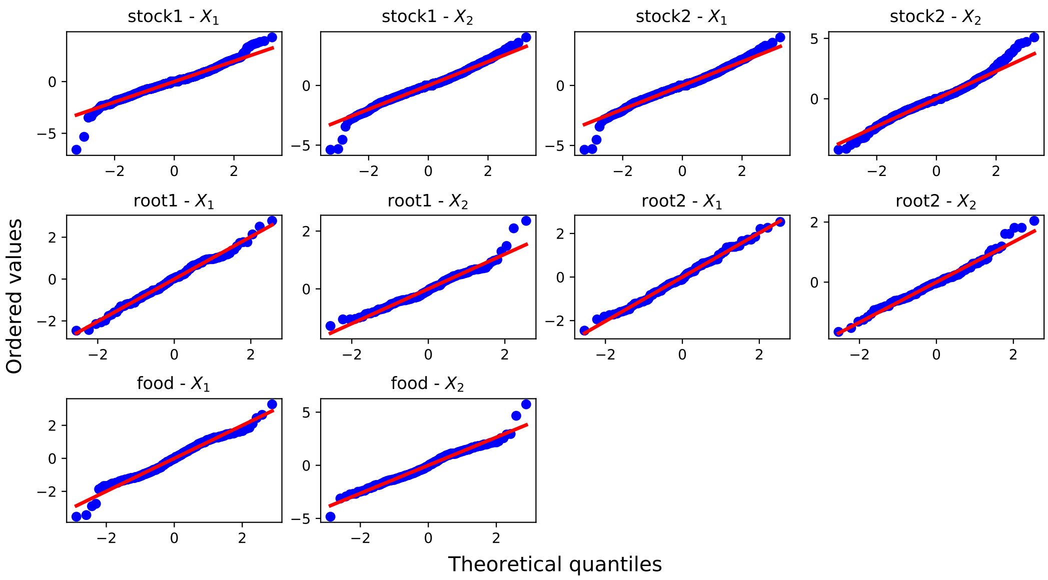

Figure 7: Gaussian probability plot of the five pairs. The ordered values for each pair and each X_1 and X_2 are plotted against an approximation of the medians of the respective quantiles of a fitted normal distribution.

Having verified all requirements and identified a reasonable log-likelihood, we proceed to calculate confidence intervals for the causal effects of X_2 on X_1. We set $\alpha = 0.05$ and use both the standard approach and the two-step approach with $\alpha_1 \in \{0.01, 0.02, 0.03\}$. Drawing from Figure 8, we notice that the estimate of the causal effect could actually

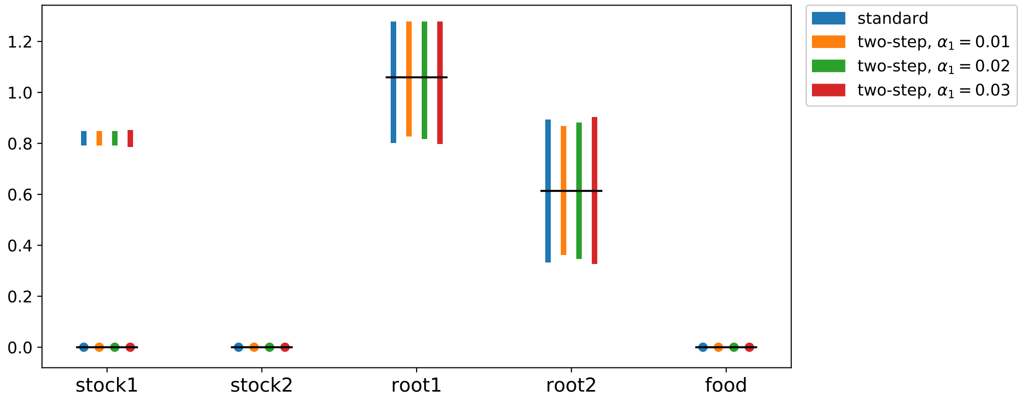

Figure 8: 95% confidence intervals for the causal effect of X_2 on X_1. The black lines are the point estimates of the respective causal effects. Potential values c_0 were tested with step length 0.005 in the range $[-3, 3]$.

recover the true direction of the causal relationship between X_1 and X_2 as stated in Table 1. Moreover, the two-step approach with $\alpha_1 = 0.01$ exhibits the best performance, albeit only for some data pairs a substantial difference could be observed.

pair	direction of causality
`stock1`	$X_1 \to X_2$
`stock2`	$X_1 \to X_2$
`root1`	$X_1 \leftarrow X_2$
`root2`	$X_1 \leftarrow X_2$
`food`	$X_1 \to X_2$

Table 1: Data pairs and true causal direction

Finally, we calculate the confidence intervals for the causal effect in the opposite direction $X_1 \to X_2$. Considering Figure 9, we find that the tested approaches yield overall similar confidence intervals. Commensurate with Figure 8, we observe that the confidence intervals for `stock1` are discontinuous by a large margin. This phenomenon can be explained by taking the small difference between the empirical variances of X_1 and X_2 into account. As the (scaled) values 1.00 and 1.01 are very close, the ambiguity in the decision on the causal direction is considerable. While the point estimate chooses correctly, the confidence interval estimate, designed to reflect uncertainty, reflects the difficulty of the task as a discontinuous interval.

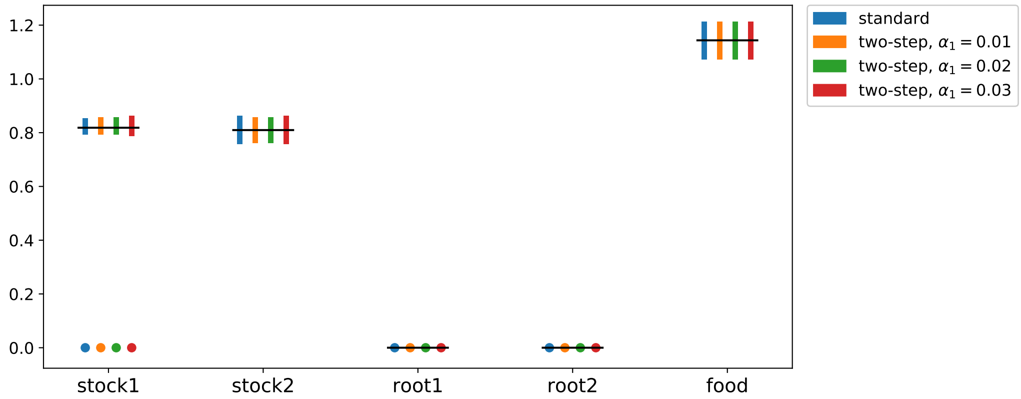

Figure 9: 95% confidence intervals for the causal effect of X_1 on X_2. The black lines are the point estimates of the respective causal effects. Potential values c_0 were tested with step length 0.005 in the range $[-3, 3]$.

Appendix 2.A Asymptotic distribution of continuous extensions

2.A.1 Convergence theorems

The delta method, the continuous mapping theorem and Slutsky's theorem are essential tools to characterise asymptotic distributions. We state them as presented in (Lehmann and Romano 2005).

Theorem 2.23 (Delta method). *Suppose $(X_n)_{n\in\mathbb{N}}$ are random vectors in $\mathbb{R}^k$ and assume $\tau_n(X_n - \mu) \xrightarrow{D} \mathcal{N}(0, \Sigma)$ where μ is a constant vector, Σ is a positive definite matrix and $(\tau_n)_{n\in\mathbb{N}}$ is a sequence of constants $\tau_n \to \infty$. Suppose g is a function from $\mathbb{R}^k$ to $\mathbb{R}$ which is differentiable at μ, then*

$$\tau_n\big(g(X_n) - g(\mu)\big) \xrightarrow{D} \mathcal{N}\left(0, (\nabla g(\mu))^T \Sigma(\nabla g(\mu))\right), \quad \text{as } n \to \infty.$$

Theorem 2.24 (Continuous mapping theorem). *Suppose $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{P} Y$ and let g be a continuous map from $\mathbb{R}^k$ to $\mathbb{R}^s$. Then $g(X_n) \xrightarrow{D} g(X)$ and $g(Y_n) \xrightarrow{P} g(Y)$.*

Theorem 2.25 (Slutsky's theorem). *Suppose $(X_n)_{n\in\mathbb{N}}$ is a sequence of real-valued random variables such that $X_n \xrightarrow{D} X$. Further, suppose $(A_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ satisfy $A_n \xrightarrow{P} a$ and $B_n \xrightarrow{P} b$, where a and b are constants. Then $A_nX_n + B_n \xrightarrow{D} aX + b$.*

2.A.2 Calculations of limit variance

We use Mathematica to facilitate the computation of the gradients of $\tilde{T}_1$ and $\tilde{T}_2$, and the asymptotic variance σ^2_∞. In the following we state the notebooks used for the $\tilde{T}_1$ and $\tilde{T}_2$ respectively. The code used is also available on Github.

Calculations for Example 2.2:

```
GammaM = {{σ1111 − σ11^2, σ1112 − σ11 ∗ σ12, σ1122 − σ11 ∗ σ22},
{σ1112 − σ11 ∗ σ12, σ1122 − σ12^2, σ1222 − σ12 ∗ σ22},
{σ1122 − σ11 ∗ σ22, σ1222 − σ12 ∗ σ22, σ2222 − σ22^2}};

TraditionalForm[t = Sign[σ12] ∗ (σ12^2 ∗ (σ11 − σ22) /σ22^3) ^(1/4)]
```

$$\sqrt[4]{\frac{\sigma_{12}^2(\sigma_{11}-\sigma_{22})}{\sigma_{22}^3}}\operatorname{sgn}(\sigma_{12})$$

```
grad = Grad[t, {σ11, σ12, σ22}];

grad = Replace[grad, Sign′[σ12] → 0, All];

TraditionalForm[

Simplify[grad, σ11 > 0 && σ22 > 0 && σ12^2 < σ11 ∗ σ22 && σ12 ≠ 0 && σ11 > σ22]]
```

$$\left\{\frac{\sigma_{12}^2\operatorname{sgn}(\sigma_{12})}{4((\sigma_{11}-\sigma_{22})\sigma_{22})^{3/4}|\sigma_{12}|^{3/2}}, \frac{\sigma_{12}\sqrt[4]{\frac{\sigma_{11}-\sigma_{22}}{\sigma_{22}^3}}\operatorname{sgn}(\sigma_{12})}{2|\sigma_{12}|^{3/2}}, -\frac{\sigma_{12}^2(3\sigma_{11}-2\sigma_{22})\operatorname{sgn}(\sigma_{12})}{4(\sigma_{11}-\sigma_{22})^{3/4}\sigma_{22}^{7/4}|\sigma_{12}|^{3/2}}\right\}$$

sigmaInf = (GammaM.grad).grad;

TraditionalForm[

Simplify [sigmaInf, $\boldsymbol{\sigma_{11} > 0 \ \&\& \ \sigma_{22} > 0 \ \&\& \ \sigma_{12}{}^\wedge 2 < \sigma_{11} * \sigma_{22} \ \&\& \ \sigma_{12} \neq 0 \ \&\&}$

$\boldsymbol{\sigma_{11} > \sigma_{22} \ \&\& \ \sigma_{1111} > 0 \ \&\& \ \sigma_{1122} > 0 \ \&\& \ \sigma_{2222} > 0 \ \&\& \ \sigma_{1122}{}^\wedge 2 < \sigma_{1111} * \sigma_{2222}]]}$

$$(\sigma_{12}^2\left(\left(9\sigma_{2222}\sigma_{11}^2 - 6\sigma_{22}\left(\sigma_{1122}+2\sigma_{2222}\right)\sigma_{11} + \sigma_{22}^2\left(\sigma_{1111}+4\left(\sigma_{1122}+\sigma_{2222}\right)\right)\right)\sigma_{12}^2\right.$$
$$-4\left(\sigma_{11}-\sigma_{22}\right)\sigma_{22}\left(3\sigma_{11}\sigma_{1222}-\sigma_{22}\left(\sigma_{1112}+2\sigma_{1222}\right)\right)\sigma_{12}+4\left(\sigma_{11}-\sigma_{22}\right)^2\sigma_{22}^2\sigma_{1122}))$$
$$/\left(16\left(\sigma_{11}-\sigma_{22}\right)^{3/2}\sigma_{22}^{7/2}\left|\sigma_{12}\right|^3\right)$$

Calculations for Example 2.3:

GammaM = $\boldsymbol{\{\{\sigma_{1111} - \sigma_{11}{}^\wedge 2, \sigma_{1112} - \sigma_{11} * \sigma_{12}, \sigma_{1122} - \sigma_{11} * \sigma_{22}\},}$

$\boldsymbol{\{\sigma_{1112} - \sigma_{11} * \sigma_{12}, \sigma_{1122} - \sigma_{12}{}^\wedge 2, \sigma_{1222} - \sigma_{12} * \sigma_{22}\},}$

$\boldsymbol{\{\sigma_{1122} - \sigma_{11} * \sigma_{22}, \sigma_{1222} - \sigma_{12} * \sigma_{22}, \sigma_{2222} - \sigma_{22}{}^\wedge 2\}\};}$

TraditionalForm[$\boldsymbol{t}$ **= Sign** $\boldsymbol{[\sigma_{12}] / \sigma_{22} * (\sigma_{12}{}^\wedge 2/2 * }$ **Abs** $\boldsymbol{[\sigma_{22} * (\sigma_{11} - \sigma_{22}) + \sigma_{12}{}^\wedge 2])\,{}^\wedge(1/4)]}$

$$\frac{\sqrt[4]{\sigma_{12}^2}\sqrt[4]{\left|\sigma_{12}^2+(\sigma_{11}-\sigma_{22})\sigma_{22}\right|}\operatorname{sgn}(\sigma_{12})}{\sqrt[4]{2}\sigma_{22}}$$

grad = Grad $\boldsymbol{[t, \{\sigma_{11}, \sigma_{12}, \sigma_{22}\}]\,;}$

grad = Replace [grad, Sign′ $\boldsymbol{[\sigma_{12}] \to 0,}$ **All] ;**

grad = Replace [grad, Abs′ $\boldsymbol{[\sigma_{12}{}^\wedge 2 + (\sigma_{11} - \sigma_{22}) * \sigma_{22}] \to 1,}$ **All] ;**

TraditionalForm[

Simplify [grad, $\boldsymbol{\sigma_{11} > 0 \ \&\& \ \sigma_{22} > 0 \ \&\& \ \sigma_{12}{}^\wedge 2 < \sigma_{11} * \sigma_{22} \ \&\& \ \sigma_{12} \neq 0 \ \&\& \ \sigma_{11} > \sigma_{22}]]}$

$$\left\{\frac{\sqrt{|\sigma_{12}|}\operatorname{sgn}(\sigma_{12})}{4\sqrt[4]{2}\left(\sigma_{12}^2+(\sigma_{11}-\sigma_{22})\sigma_{22}\right)^{3/4}}, \frac{\sigma_{12}\left(2\sigma_{12}^2+(\sigma_{11}-\sigma_{22})\sigma_{22}\right)\operatorname{sgn}(\sigma_{12})}{2\sqrt[4]{2}\sigma_{22}\left(\sigma_{12}^4+(\sigma_{11}-\sigma_{22})\sigma_{22}\sigma_{12}^2\right)^{3/4}}, -\frac{\left(4\sigma_{12}^2+(3\sigma_{11}-2\sigma_{22})\sigma_{22}\right)\sqrt{|\sigma_{12}|}\operatorname{sgn}(\sigma_{12})}{4\sqrt[4]{2}\sigma_{22}^2\left(\sigma_{12}^2+(\sigma_{11}-\sigma_{22})\sigma_{22}\right)^{3/4}}\right\}$$

sigmaInf = (GammaM.grad).grad;

TraditionalForm[

Simplify [sigmaInf, $\boldsymbol{\sigma_{11} > 0 \ \&\& \ \sigma_{22} > 0 \ \&\& \ \sigma_{12}{}^\wedge 2 < \sigma_{11} * \sigma_{22} \ \&\& \ \sigma_{12} \neq 0 \ \&\&}$

$\boldsymbol{\sigma_{11} > \sigma_{22} \ \&\& \ \sigma_{1111} > 0 \ \&\& \ \sigma_{1122} > 0 \ \&\& \ \sigma_{2222} > 0 \ \&\& \ \sigma_{1122}{}^\wedge 2 < \sigma_{1111} * \sigma_{2222}]]}$

$$(16\sigma_{2222}\sigma_{12}^6 - 32\sigma_{22}\sigma_{1222}\sigma_{12}^5 + 8\sigma_{22}\left(\sigma_{22}\left(\sigma_{1122}-2\sigma_{2222}\right)+3\sigma_{11}\sigma_{2222}\right)\sigma_{12}^4$$

$$+ 8\sigma_{22}^2\left(\sigma_{22}\left(\sigma_{1112} + 4\sigma_{1222}\right) - 5\sigma_{11}\sigma_{1222}\right)\sigma_{12}^3 + \sigma_{22}^2(9\sigma_{2222}\sigma_{11}^2 + 2\sigma_{22}\left(5\sigma_{1122} - 6\sigma_{2222}\right)\sigma_{11}$$
$$+ \sigma_{22}^2\left(\sigma_{1111} - 12\sigma_{1122} + 4\sigma_{2222}\right))\sigma_{12}^2 - 4\left(\sigma_{11} - \sigma_{22}\right)\sigma_{22}^3\left(3\sigma_{11}\sigma_{1222} - \sigma_{22}\left(\sigma_{1112} + 2\sigma_{1222}\right)\right)\sigma_{12}$$
$$+ 4(\sigma_{11} - \sigma_{22})^2\sigma_{22}^4\sigma_{1122})/(16\sqrt{2}\sigma_{22}^4(\sigma_{12}^2 + (\sigma_{11} - \sigma_{22})\sigma_{22})^{3/2}|\sigma_{12}|)$$

Appendix 2.B Simulation details

Using the model (2.30), we assume normally distributed data and thus consider the negative log-likelihood

$$-\ell(\Sigma) = \frac{1}{2}\Big(2n\ln(2\pi) + n\ln(\det(\Sigma)) + \sum_{j=1}^{n} x_j^T\Sigma^{-1}x_j\Big)$$

for a data sample $(x_1, \ldots, x_n)$, $x_j \in \mathbb{R}^2$ for all $j \in \{1, \ldots, n\}$. The Algorithms 1 and 2 rely on the minimisation of $-\ell$ for the different sets of constraints given by (2.24) – (2.26). To this end, we use (Kraft 1988)'s sequential least squares programming algorithm that requires the derivatives of $-\ell$ with respect to Σ. Since the inverse of Σ appears in the sum which compounds the derivation of $-\partial\ell/\partial\Sigma$, we choose the concentration matrix $K = \Sigma^{-1}$ as parametrisation of the negative log-likelihood.

$$K = \begin{pmatrix} k_1 & k_2 \\ k_2 & k_3 \end{pmatrix} = \Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix} \tag{2.31}$$

Employing the following identities, cf. (Minka 1997),

$$\frac{\partial(x^TKx)}{\partial K} = xx^T, \qquad \frac{\partial\ln(\det(K))}{\partial K} = (K^T)^{-1},$$

we compute the derivative

$$\frac{\partial(-\ell(K))}{\partial K} = -\frac{n}{2}K^{-1} + \frac{1}{2}\sum_{j=1}^{n} x_jx_j^T.$$

As last step before applying the optimisation routine, we have to express (2.24) – (2.26) in terms of the new parametrisation. Using (2.31), we find

$$\begin{aligned} \sigma_{12} &= c_0\,\sigma_{22} &&\Leftrightarrow & -k_2 &= c_0\,k_1, \\ \sigma_{11} &\geq \sigma_{22} &&\Leftrightarrow & k_3 &\geq k_1, \\ \sigma_{11} &\leq \sigma_{22} &&\Leftrightarrow & k_3 &\leq k_1. \end{aligned}$$

References

Andrews, Donald W. K. (2000). "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space." In: *Econometrica* 68.2, pp. 399–405.

Chen, Wenyu, Mathias Drton, and Y. Samuel Wang (2019). "On causal discovery with an equal-variance assumption." In: *Biometrika* 106.4, pp. 973–980.

Chernoff, Herman (1954). "On the distribution of the likelihood ratio." In: *Ann. Math. Stat.* 25, pp. 573–578.

Drton, Mathias (2009). "Likelihood ratio tests and singularities." In: *Ann. Stat.* 37.2, pp. 979–1012.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian Marx (2013). *Regression. Models, methods and applications.* Berlin: Springer, pp. xiv + 698.

Geyer, Charles J. (1994). "On the asymptotics of constrained M-estimation." In: *Ann. Stat.* 22.4, pp. 1993–2010.

Heinze-Deml, Christina, Marloes H. Maathuis, and Nicolai Meinshausen (2018). "Causal Structure Learning". In: *Annual Review of Statistics and Its Application* 5.1, pp. 371–391.

Kraft, Dieter (1988). "A software package for sequential quadratic programming". In: *Tech. Rep. DFVLR-FB 88-28.*

Lehmann, E. L. and Joseph P. Romano (2005). *Testing statistical hypotheses. 3rd ed.* 3rd ed. New York, NY: Springer, pp. xiv + 784.

Loh, Po-Ling and Peter Bühlmann (2014). "High-dimensional learning of linear causal networks via inverse covariance estimation." In: *J. Mach. Learn. Res.* 15, pp. 3065–3105.

Maathuis, Marloes, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, eds. (2019). *Handbook of graphical models.* Boca Raton, FL: CRC Press, pp. xviii + 536.

Minka, Tom (1997). "Old and New Matrix Algebra Useful for Statistics". In:

Mooij, Joris M., Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf (2016). "Distinguishing cause from effect using observational data: methods and benchmarks." In: *J. Mach. Learn. Res.* 17. Id/No 32, p. 102.

Pearl, Judea (1995). "Causal Diagrams for Empirical Research". In: *Biometrika* 82.4, pp. 669–688.

— (2009). *Causality. Models, reasoning, and inference. 2nd revised ed.* 2nd revised ed. Cambridge: Cambridge University Press, pp. xviii + 464.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* Cambridge, MA, USA: MIT Press.

Peters, Jonas and Peter Bühlmann (2014). "Identifiability of Gaussian structural equation models with equal error variances." In: *Biometrika* 101.1, pp. 219–228.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf (1999). *Subsampling.* New York, NY: Springer, pp. xv + 347.

Silvapulle, Mervyn J. (1996). "A test in the presence of nuisance parameters." In: *J. Am. Stat. Assoc.* 91.436, pp. 1690–1693.

Silvapulle, Mervyn J. and Pranab K. Sen (2005). *Constrained statistical inference. Inequality, order, and shape restrictions.* Hoboken, NJ: John Wiley & Sons, pp. xvii + 532.

Spirtes, Peter, Clark Glymour, and Richard Scheines (2001). *Causation, prediction, and search. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson. 2nd ed.* 2nd ed. Cambridge, MA: MIT Press, pp. xxii + 496.

Spława-Neyman, Jerzy (1923). *Essai d'application de la statistique mathématique à la résolution de quelques problèmes agricoles.* Revue mensuelle de statistique publiée par l'Office central de statistique de la République polonaise 6, 29 S. (1923).

Steiger, James H. and A. Ralph Hakstian (1982). "The asymptotic distribution of elements of a correlation matrix: Theory and application." In: *Br. J. Math. Stat. Psychol.* 35, pp. 208–215.

Wolak, Frank A. (1989). "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models". In: *Econometric Theory* 5.1, pp. 1–35.

Wright, Sewall (1921). "Correlation and causation". In: *Journal of agricultural research* 20, pp. 557–580.

3 Post-selection inference with HSIC-Lasso

This section is organised as follows. The Subsections 3.1 and 3.2 introduce post-selection inference and the Hilbert-Schmidt independence criterion (HSIC) as the two main cornerstones of this work. The following Subsection 3.3 combines both these fields of research to obtain an asymptotically valid approach for inference after selection with HSIC-Lasso. Subsequently, the performance of this procedure is evaluated both on artificial and benchmark data in the Subsections 3.4 and 3.5 respectively.

3.1 Post-selection inference

Taking the effects of a preceding selection procedure into account when doing inference is an idea that can be traced back as early as (Fisher 1956), (Cox 1975) and (Pötscher 1991). However, it was not until 2005 that this issue gained traction and established post-selection inference (PSI) as a field in its own right with several branches and lines of research.

3.1.1 Different approaches

Throughout this subsection, we use the model

$$Y \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{Id})$$

with iid. data, known σ^2 and (R. Tibshirani 1996)'s Lasso-selection procedure with tuning parameter λ as a running example to illustrate different approaches for PSI. The sample size is denoted by n and the number of covariates, or rather the dimension of β, by p respectively.

(Cox 1975) proposed to randomly split the dataset into two folds, one on which the selection is carried out and one which is used for inference. Hence, under the common iid. assumption both procedures are independent and do not need to be adapted in order to guarantee correct inference results. Although this approach is strikingly simple and guaranteed to be valid for any kind of selection procedure, it outright disregards the selection-data on the inference stage which leaves space for improvement.

Two standard approaches to deduce p-values and confidence intervals for selected variables are Bayesian methods and bootstrapping. As laid out by (Park and Casella 2008), a Bayesian Lasso can be realised by considering β a random quantity with exponential prior

$$\pi(\beta \,|\, \sigma, \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_j|}.$$

Hence, we can infer the posterior distribution of β, e.g. with Gibbs sampling, which allows for hypothesis testing and confidence interval estimation.

Following a non-parametric bootstrap approach, cf. (Efron 1979) and (Efron and Robert J. Tibshirani 1993), we sample datasets of size n with replacement from the original data, carry out the selection procedure and use the empirical distribution of β for inference.

This method can be further refined using a parametric bootstrap and is also applicable when λ is not fixed and, for example, chosen by cross-validation.
(Hastie et al. 2015) points out that the Bayesian approach is of complexity $\mathcal{O}(p^2)$, whereas bootstrapping only exhibits $\mathcal{O}(p)$. Nevertheless, both methods are not suitable for high- and ultra high-dimensional problems as they heavily rely on computationally expensive repeated sampling.

Debiased Lasso is a method that lies on the edge of what is considered post-selection inference but is included in this short compendium due to its wide reception. It was developed by several authors, for instance (C.-H. Zhang and S. S. Zhang 2014), (Bühlmann 2013), (van de Geer et al. 2014) and (Javanmard and Montanari 2014), and is tailored for high-dimensional ($p \gg n$) linear models. In the case $p > n$, fitting the full linear model $Y = X\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$, and using the traditional confidence intervals for the regression coefficients is not feasible. Instead, using a debiased version $\hat{\beta}^d$ of the Lasso estimate $\hat{\beta}_\lambda$ is suggested and defined by

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\Theta X^T(y - X\hat{\beta}_\lambda) = \beta + \frac{1}{n}\Theta X^T\varepsilon + \hat{\Delta}.$$

The matrix Θ is an approximate inverse of $\hat{\Sigma} = \frac{1}{n}X^T X$ that is chosen such that $\hat{\Delta} = (\mathrm{Id} - \frac{1}{n}\Theta X^T X)(\hat{\beta}_\lambda - \beta)$ vanishes as $n \to \infty$. Hence, the approximation

$$\hat{\beta}^d \sim \mathcal{N}\Big(\beta, \frac{\sigma^2}{n}\Theta\,\hat{\Sigma}\,\Theta^T\Big)$$

can be used to derive confidence intervals for β. Note that this approach aims at inference on the model that includes all covariates, the so called full model, putting its main focus on enlarging the traditional linear regression approach. Hence, the results will differ from techniques that consider models with a subset of selected covariates, hereinafter partial models.

As described in (Jason D. Lee et al. 2016), the inherent difficulty of PSI is the dependence of the inference target on the selection procedure. Denoting a model estimator by $\hat{M} \subset \{1, \dots, p\}$, the regression coefficients in the selected (partial) model are given by

$$\hat{\beta}^{\hat{M}} = (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{\hat{M}}^T Y,$$

where $X_{\hat{M}}$ only contains the columns of X with index in $\hat{M}$. Commensurate with classical linear regression, we might require $\mathbb{P}(\hat{\beta}_j^{\hat{M}} \in C_j^{\hat{M}}) \geq 1 - \alpha$ for a confidence region $C_j^{\hat{M}}$. However, the event inside the probability measure is not well-defined as, for a fixed model M, β_j^M is undefined if $j \notin M$. The PoSI approach, put forward by (Berk et al. 2013), solves this issue by demanding

$$\mathbb{P}\left(\hat{\beta}_j^{\hat{M}} \in C_j^{\hat{M}}\, \forall j \in \hat{M}\right) \geq 1 - \alpha \quad \forall \hat{M}. \tag{3.1}$$

Here, $\hat{M}$ is not a deterministic subset of $\{1, \dots, p\}$ but refers to any model selection procedure. The requirement (3.1) exhibits two major characteristics: First, the coverage property of the confidence interval is family-wise for all selected variables $j \in \hat{M}$; second, confidence intervals are universally valid, i.e. correct regardless of the selection procedure. In analogy with the classical theory,

$$\left(\hat{\beta}_j^{\hat{M}} \pm K\sqrt{(X_{\hat{M}}^T X_{\hat{M}})_{jj}^{-1}}\,\hat{\sigma}\right)$$

is suggested as confidence interval where $\hat{\sigma}$ is an estimate of the standard deviation in the full model and K is chosen depending on X, the level α, the degrees of freedom in $\hat{\sigma}$ and the space of considered models such that universal validity is achieved. For more details, we refer to the original paper.

PoSI is especially useful in situations where the selection procedure is unknown or account of it is not trustworthy because valid inference is still guaranteed. However, accounting for all possible selection procedures often yields very conservative confidence intervals. In fact, it is shown that K can grow as quickly as $\mathcal{O}(p^{1/2})$. Moreover, PoSI is computationally very expensive which renders it infeasible for high-dimensional situations.

3.1.2 Truncated Gaussians and polyhedral lemma

Contrary to PoSI, (Jason D. Lee et al. 2016) drop the universal validity requirement and, instead of family-wise coverage, propose conditional coverage which amounts to

$$\mathbb{P}\big(\hat{\beta}_j^M \in C_j^M | \hat{M} = M\big) \geq 1 - \alpha \quad \forall j \in M.$$

Conditioning on $\{\hat{M} = M\}$ is sensible as a confidence interval C_j^M is calculated if and only if M is selected. This also lifts the need for comparing regression coefficients across different models, as demanded by '$\forall j \in \hat{M}$' in (3.1).

Following the principal of conditional coverage, the distribution of a large class of statistics given a selection procedure can be captured. We assume that $Y \sim \mathcal{N}(\mu, \Sigma)$, where μ is unknown and Σ is known, and the quantity of interest for inference is given by $\eta_M^T \mu$. The vector η_M is fixed and can depend on the selected model. Taking linear regression as an example, μ is modelled by $X\beta$ and we set $\eta_M = \mathrm{e}_j(X_M^T X_M)^{-1} X_M^T$ for inference on β_j^M. In order to deduce valid confidence intervals, characterising the distribution of

$$\eta_M^T Y | \{\hat{M} = M\}$$

is essential. The selection procedure $\hat{M}(Y)$ is inherently dependent on the response and we assume that the selection event can be expressed in terms of Y in an affine linear fashion, i.e.

$$\{\hat{M} = M\} = \{Y \in \mathbb{R}^n \colon A(M)Y \leq b(M)\} =: \{AY \leq b\},$$

for $A(M) \in \mathbb{R}^{k \times n}$ and $b(M) \in \mathbb{R}^k$. Geometrically, $\{AY \leq b\}$ describes a polyhedron which leads to the name polyhedral lemma. (Jason D. Lee et al. 2016) show that the selection event of Lasso can indeed be characterised this way if one additionally conditions on the signs of the selected variables. The seminal insight of the authors is that $\{AY \leq b\}$ can be disentangled rewriting Y in terms of $\eta_M^T Y$ and a component Z that is independent of $\eta_M^T Y$.

Lemma 3.1 (Polyhedral lemma)**.** *Let $Y \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$, $\eta \in \mathbb{R}^n$, $A \in \mathbb{R}^{k\times n}$ and $b \in \mathbb{R}^k$. Then Z, defined by*

$$Z := \left(\mathrm{Id} - C\eta^T\right) Y, \qquad C := (\eta^T\Sigma\eta)^{-1}\Sigma\eta,$$

and $\eta^T Y$ are independent. Furthermore,

$$\{AY \leq b\} = \left\{\mathcal{V}^-(Z) \leq \eta^T Y \leq \mathcal{V}^+(Z), \mathcal{V}^0(Z) \geq 0\right\},$$

where

$$\mathcal{V}^-(Z) := \max_{j:(AC)_j<0} \frac{b_j - (AZ)_j}{(AC)_j}, \tag{3.2}$$

$$\mathcal{V}^+(Z) := \max_{j:(AC)_j>0} \frac{b_j - (AZ)_j}{(AC)_j}, \tag{3.3}$$

$$\mathcal{V}^0(Z) := \max_{j:(AC)_j=0} b_j - (AZ)_j.$$

Drawing from this understanding, we see that the selection procedure has an effect on the inference inasmuch as it restricts the values the quantity of interest can assume. Against this backdrop, we define a truncated Gaussian distribution as follows.

Definition 3.2. Let $\mu \in \mathbb{R}$, $\sigma^2 > 0$ and $a, b \in \mathbb{R}$ such that $a < b$. Then the cumulative distribution function of a *Gaussian distribution* $\mathcal{N}(\mu, \sigma^2)$ *truncated to the interval* $[a, b]$ is given by

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi(\frac{x-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},$$

where Φ denotes the cdf of $\mathcal{N}(0, 1)$.

Concluding the presented line of thought, we are now able to state a pivotal quantity that can be used for inference.

Theorem 3.3. *Under the assumptions of Lemma 3.1 it holds that*

$$F^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}_{\eta^T\mu,\eta^T\Sigma\eta}(\eta^T Y)|\{AY \leq b\} \sim \mathrm{Unif}\,(0, 1), \tag{3.4}$$

where $\mathcal{V}^-$ and $\mathcal{V}^+$ are given by (3.2) and (3.3) respectively.

(Ryan J. Tibshirani, J. Taylor, et al. 2016) describes how one- and two-sided hypothesis testing and confidence interval calculation can be done. Suppose we want to test

$$\mathrm{H}_0 : \eta_M^T\mu = 0 \quad \text{against} \quad \mathrm{H}_1 : \eta_M^T\mu > 0. \tag{3.5}$$

Then the statistic

$$T = 1 - F^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}_{0,\eta^T\Sigma\eta}(\eta^T Y)$$

is a valid p-value for H_0 conditional on $\{AY \leq b\}$. Further, defining δ_α for $0 \leq \alpha \leq 1$ such that

$$1 - F^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}_{\delta_\alpha,\eta^T\Sigma\eta}(\eta^T Y) = \alpha$$

yields a valid one-sided confidence interval $[\delta_\alpha, \infty)$ conditional on $\{AY \leq b\}$.
Likewise, we consider the two-sided hypothesis testing problem

$$\mathrm{H}_0 : \eta_M^T \mu = 0 \quad \text{against} \quad \mathrm{H}_1 : \eta_M^T \mu \neq 0$$

and use the statistic

$$T = 2 \min \left\{ F_{0,\eta^T \Sigma \eta}^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}(\eta^T Y), 1 - F_{0,\eta^T \Sigma \eta}^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}(\eta^T Y) \right\}.$$

Again, T is a valid conditional p-value and defining $\delta_{\alpha/2}$ and $\delta_{1-\alpha/2}$ such that

$$\begin{aligned} 1 - F_{\delta_{\alpha/2},\eta^T \Sigma \eta}^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}(\eta^T Y) &= \alpha/2, \\ 1 - F_{\delta_{1-\alpha/2},\eta^T \Sigma \eta}^{[\mathcal{V}^-(Z),\mathcal{V}^+(Z)]}(\eta^T Y) &= 1 - \alpha/2 \end{aligned}$$

yields a valid confidence interval $[\delta_{\alpha/2}, \delta_{1-\alpha/2}]$ conditional on $\{AY \leq b\}$.
The polyhedral lemma 3.1 in combination with Theorem 3.3 sparked an entire branch of research within PSI as it has many favourable traits. Since it does not rely on computationally costly operations, it is suitable for high-dimensional settings. In addition, it is not tailored to a specific target or selection procedure but can be applied as long as the quantity of interest can be written as $\eta_M^T \mu$ and the selection event can be represented as $\{AY \leq b\}$. Furthermore, it is remarkable that Theorem 3.3 is non-asymptotic in that it exactly holds for finite sample sizes. Even if the assumption of normality is violated, generalisations for asymptotically Gaussian quantities are often possible. This is harnessed in section 3.3.
Concurrent to the development of the polyhedral lemma, (Ryan J. Tibshirani, J. Taylor, et al. 2016) introduced the spacing test, tailored to PSI for the last angle regression algorithm which calculates the Lasso regularisation path, cf. (Efron, Hastie, et al. 2004). It also builds on the core elements of polyhedral selection events and truncated Gaussian distributions. Since then, numerous generalisations and adaptations of this framework have been proposed. For instance, (Ryan J. Tibshirani, Rinaldo, et al. 2018) and (Tian and J. Taylor 2017) drop the assumption of normality and investigate large sample properties for $n > p$ and $n < p$ respectively. (Tian, Loftus, et al. 2018) tackle the issue of incorporating the estimation of Σ, (Hyun et al. 2018) use the generalised Lasso as selection procedure and (J. Taylor and R. Tibshirani 2018) present applications to logistic regression and the graphical Lasso.

3.2 Hilbert-Schmidt independence criterion (HSIC)

Detecting dependence between random variables via kernel-based approaches and reproducing kernel Hilbert spaces (RKHS) was a very active field of research at the beginning of the new millennium producing several proposals, such as (Bach and Jordan 2003), (Achard et al. 2003) or (Fukumizu et al. 2004). This effort culminated in the introduction of the Hilbert-Schmidt independence criterion (HSIC) by (Gretton, Bousquet, et al. 2005) and its subsequent examination and application.

3.2.1 Theoretical derivation

The main incentive to develop advanced techniques to describe dependence relations between two random variables X and Y arises from the fact that the covariance

$$\operatorname{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

is designed for linear relationships only. If the dependence structure, however, is of nonlinear nature, the covariance can only partly capture the relationship between X and Y or completely fails to do so. Nevertheless, general, or rather model-free, independence can be expressed in terms of the covariance as follows, cf. (Gretton, Alexander Smola, et al. 2005).

Proposition 3.4. *The random variables X and Y are independent if and only if* $\operatorname{cov}(f(X), g(Y)) = 0$ *for each pair (f, g) of bounded, continuous functions.*

There are two lines of thought leading to the Hilbert-Schmidt independence criterion, one presented in (Gretton, Bousquet, et al. 2005) regarding HSIC as the Hilbert-Schmidt norm of a cross-covariance operator and one thinking of HSIC as mean maximum discrepancy on product spaces according to (Q. Zhang et al. 2018). In this work, we follow the latter derivation and link it with the first approach and Proposition 3.4 at the end.
First, we introduce the concept of reproducing kernel Hilbert spaces.

Definition 3.5. Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on D with scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. A function $k\colon D\times D\to\mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

1. $k(\cdot,x)\in\mathcal{H} \quad \forall x\in D$,
2. $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x) \quad \forall x\in D\ \forall f\in\mathcal{H}$.

If $\mathcal{H}$ has a reproducing kernel, it is called a *reproducing kernel Hilbert space (RKHS).*

Remark 3.1. As an immediate consequence of the upper definition, we get

$$k(x,y) = \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}} \quad \forall x,y\in D.$$

The following theorem, proved by (Aronszajn 1950), provides sufficient conditions for a function k to be a reproducing kernel.

Theorem 3.6 (Moore-Aronszajn)**.** *Let $k\colon D\times D\to\mathbb{R}$, be symmetric and positive definite, that is*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i,x_j) \geq 0, \quad \forall n\geq 1\ \forall a\in\mathbb{R}^n\ \forall x\in D^n.$$

Then there is a unique RKHS $\mathcal{H}_k$ with reproducing kernel k.

Against this backdrop, we may ask how properties of the kernel k translate into characteristics of $\mathcal{H}_k$. The notion of a universal kernel, introduced by (Steinwart 2002), helps to shed light on this issue.

Definition 3.7. A continuous kernel k on a compact metric space (D,d) is called *universal* if $\mathcal{H}_k$ is dense in $C(D)$, the space of continuous functions on D, with respect to $\|\cdot\|_\infty$.

It is shown that both the Gaussian and exponential kernel, defined by

$$\begin{aligned} k(x,y) &= \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right), \quad \sigma^2 > 0, \\ k(x,y) &= \exp\left(-\frac{\|x-y\|_2}{2\sigma}\right), \quad \sigma > 0 \end{aligned}$$

respectively, are universal.
Second, we introduce the particularly useful framework of embedding distributions into Hilbert spaces according to (Alex Smola et al. 2007).

Definition 3.8. Let k be a bounded kernel on D and $\mathbb{P}$ a probability measure on D. The *kernel embedding* of $\mathbb{P}$ into the RKHS $\mathcal{H}_k$ is $\mu_k(\mathbb{P}) \in \mathcal{H}_k$ such that

$$\mathrm{E}\left[f(X)\right] = \int_D f(x)\,\mathrm{d}\mathbb{P}(x) = \langle f, \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k}, \quad X \sim \mathbb{P}, \forall f \in \mathcal{H}_k.$$

Remark 3.2. Alternatively, $\mu_k(\mathbb{P})$ can be defined by

$$\mu_k(\mathbb{P}) = \int_D k(\cdot, x)\,\mathrm{d}\mathbb{P}(x).$$

Definition 3.8 allows us to use Hilbert space theory on distributions which gives rise to the definition of maximum mean discrepancy (MMD), see for example (Borgwardt et al. 2006) and (Gretton, Borgwardt, et al. 2012), which measures the distance between probability measures.

Definition 3.9. Let k be a bounded kernel and $\mathbb{P}$ and $\mathbb{Q}$ probability measures on D. The *maximum mean discrepancy (MMD)* between $\mathbb{P}$ and $\mathbb{Q}$ with respect to k is defined as

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2.$$

Lemma 3.10. *In the setting of definition 3.9,* MMD_k *is a metric on probability measures if* k *is a universal kernel.*

Proof. (Alex Smola et al. 2007) show that $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective for universal k. Hence, any two different measures have two distinct embeddings. The statement directly follows from the norm properties of $\|\cdot\|_{\mathcal{H}_k}$. □

The maximum mean discrepancy can be used to test whether two given data samples stem from the same distribution. Since our goal is to find a measure for the dependence between two random variables X and Y, we use MMD to compare the joint distribution $\mathbb{P}_{X,Y}$ and the product of the marginals $\mathbb{P}_X\mathbb{P}_Y$.
To this end, consider any two kernels k and l on the domains D_X and D_Y. It is easy to verify that $K = k \otimes l$ given by

$$K\big((x,y),(x',y')\big) = k(x,x')\,l(y,y'), \quad x,x' \in D_X,\ y,y' \in D_Y$$

is a valid kernel on the product space $D_X \times D_Y$. Employing Remark 3.2, we can define a dependence measure between X and Y based on RKHSs.

Definition 3.11. Let X and Y be random variables and k and l be bounded kernels on the domains D_X and D_Y respectively. The *Hilbert-Schmidt independence criterion* $\mathrm{HSIC}_{k,l}(X,Y)$ for X and Y based on the kernels k and l is given by

$$\begin{aligned}\mathrm{HSIC}_{k,l}(X,Y) &= \mathrm{MMD}_{k\otimes l}(\mathbb{P}_{X,Y}, \mathbb{P}_X\mathbb{P}_Y)\\ &= \left\|\mathrm{E}_{XY}\left[k(\cdot,X)\otimes l(\cdot,Y)\right] - \mathrm{E}_X\left[k(\cdot,X)\right]\,\mathrm{E}_Y\left[l(\cdot,Y)\right]\right\|^2_{\mathcal{H}_{k\otimes l}}. \end{aligned} \tag{3.6}$$

The name of HSIC stems from the point of view held by (Gretton, Bousquet, et al. 2005). The term within the norm in (3.6) can be identified with the cross-covariance operator $C_{XY}\colon \mathcal{H}_k \to \mathcal{H}_l$ for which

$$\langle f, C_{XY}g\rangle_{\mathcal{H}_k} = \mathrm{cov}\left(f(X), g(Y)\right) \quad \forall f \in \mathcal{H}_k \forall g \in \mathcal{H}_l \tag{3.7}$$

holds. Consequently, HSIC is the squared Hilbert-Schmidt norm $\|C_{XY}\|^2_{\mathrm{HS}}$.
Coming full circle, we see that using universal kernels k and l, which causes $k\otimes l$ to be universal as well, has two important implications. First, Lemma 3.10 states that HSIC is indeed a valid metric to measure dependence between random variables. Second, Definition 3.7 yields that $\mathcal{H}_k$ and $\mathcal{H}_l$ are dense in $C(D_X)$ and $C(D_Y)$ respectively. Hence, (3.7) directly reflects the characterisation of independence given in Proposition 3.4.
It can be shown that HSIC can be expressed in terms of kernels which is a more convenient perspective to develop estimators.

Lemma 3.12. *Assuming the setting of Definition 3.2.1, let X' and Y' be independent copies of X and Y. The Hilbert-Schmidt independence criterion has the representation*

$$\begin{aligned}\mathrm{HSIC}_{k,l}(X,Y) =& \,\mathrm{E}_{X,X',Y,Y'}\left[k(X,X')\,l(Y,Y')\right] + \mathrm{E}_{X,X'}\left[k(X,X')\right]\mathrm{E}_{Y,Y'}\left[l(Y,Y')\right]\\ &- 2\,\mathrm{E}_{X,Y}\left[\mathrm{E}_{X'}\left[k(X,X')\right]\mathrm{E}_{Y'}\left[l(Y,Y')\right]\right].\end{aligned}$$

3.2.2 Estimators

Since the introduction of the Hilbert-Schmidt independence criterion several estimators have been proposed. We assume that a data sample $\{(x_j,y_j)\}_{j=1}^n$ is given and that the kernels k and l are universal and w.l.o.g. bounded by 1. (Gretton, Bousquet, et al. 2005) propose a simple estimator, which, however, exhibits a bias of order $\mathcal{O}(n^{-1})$, whereas (Song et al. 2012) correct this unfavourable trait putting forward an unbiased estimator.

Definition 3.13. Let K and L be defined by $K_{ij} = k(x_i,x_j)$ and $L_{ij} = l(x_i,x_j)$ for $1 \leq i,j \leq n$ and set $\tilde{K} = K - \mathrm{diag}(K)$, $\tilde{L} = L - \mathrm{diag}(L)$ and $\Gamma = \mathrm{Id} - \frac{1}{n}11^T$, where $1 \in \mathbb{R}^n$ has one at every entry. The *biased* and *unbiased HSIC-estimators* $\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X,Y)$ and $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X,Y)$ are defined as

$$\begin{aligned}\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X,Y) &= \frac{1}{(n-1)^2}\,\mathrm{tr}(K\Gamma L\Gamma),\\ \widehat{\mathrm{HSIC}}_{\mathrm{u}}(X,Y) &= \frac{1}{n(n-3)}\left(\mathrm{tr}(\tilde{K}\tilde{L}) + \frac{1^T\tilde{K}1\,1^T\tilde{L}1}{(n-1)(n-2)} - \frac{2}{n-2}1^T\tilde{K}\tilde{L}1\right).\end{aligned}$$

For both estimators the respective authors state concentration inequalities.

Theorem 3.14. *For $n > 1$ and all δ with probability of at least $1 - \delta$*

$$\left|\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X,Y) - \mathrm{HSIC}(X,Y)\right| \leq \sqrt{\frac{\log(6/\delta)}{\alpha^2 n}} + \frac{C}{n},$$
$$\left|\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X,Y) - \mathrm{HSIC}(X,Y)\right| \leq 8\sqrt{\frac{\log(2/\delta)}{n}},$$

where $\alpha^2 > 0.24$ and C are constants.

Regarding the asymptotic distribution of $\widehat{\mathrm{HSIC}}_{\mathrm{b}}$ and $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$, (Q. Zhang et al. 2018) summarise that both estimators scaled by $n^{1/2}$ converge to a Gaussian distribution if X and Y are dependent. However, for $X \perp Y$, the asymptotic distribution is not normal. This is a cumbersome property as we do not know the true dependence between X and Y. For this reason, we turn our attention to estimators that are asymptotically normal in either case.

In the following, it proves advantageous to use the framework of U-statistics in order to develop and establish properties of estimators. More details on this topic can be found in Appendix 3.A. (Song et al. 2012) prove that $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$ indeed has an according representation.

Theorem 3.15. *Using the notation of Definition 3.13, $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$ is a U-statistic of degree 4 with kernel*

$$h(i,j,q,r) = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} K_{st}(L_{st} + L_{uv} - 2L_{su}).$$

The sum is taken over all 24 quadruples (s,t,u,v) that can be selected without replacement from (i,j,q,r) and the notation of h was reduced to only contain the indices.

(Q. Zhang et al. 2018) and recently (Lim et al. 2020) suggested new estimators for the Hilbert-Schmidt independence criterion.

Definition 3.16. Let $B \in \mathbb{N}$ and subdivide the data into folds of size B, $\{\{(x_i^b, y_i^b)\}_{i=1}^B\}_{b=1}^{n/B}$. The *block estimator* $\widehat{\mathrm{HSIC}}_{\mathrm{block}}$ with block size B is given by

$$\widehat{\mathrm{HSIC}}_{\mathrm{block}}(X,Y) = \frac{1}{n/B} \sum_{b=1}^{n/B} \widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^b, Y^b), \tag{3.8}$$

where $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^b, Y^b)$ denotes the unbiased estimator on the data $\{(x_i^b, y_i^b)\}_{i=1}^B$.

Let $\mathcal{S}_{n,4}$ be the set of all 4-subsets of $\{1,\ldots,n\}$ and let $\mathcal{D}$ be a multiset containing m elements of $\mathcal{S}_{n,4}$ randomly chosen with replacement. Further, suppose $m = \mathcal{O}(n)$ and define $l := \lim_{n,m\to\infty} m/n$. The *incomplete U-statistics estimator* $\widehat{\mathrm{HSIC}}_{\mathrm{inc}}$ of size l is defined by

$$\widehat{\mathrm{HSIC}}_{\mathrm{inc}}(X,Y) = \frac{1}{m} \sum_{(i,j,q,r)\in\mathcal{D}} h(i,j,q,r). \tag{3.9}$$

We show that both estimators are asymptotically normal and prove a multidimensional version of the central limit theorem.

Theorem 3.17. *Let $\{(x_j^{(1)}, \dots, x_j^{(p)}, y_j)\}_{j=1}^n$ be an iid. data sample and define $H_0 = (\mathrm{HSIC}(X^{(1)}, Y), \dots, \mathrm{HSIC}(X^{(p)}, Y))^T$, and H_{block} and H_{inc} accordingly. Assume that B and l are the same for all entries of H_{block} and H_{inc} respectively, let $n/B \to \infty$ and choose $\mathcal{D}$ for all elements of H_{inc} independently. Then*

$$\sqrt{n/B}\big(H_{\text{block}} - H_0\big) \xrightarrow{D} \mathcal{N}(0, \Sigma_{\text{block}}), \tag{3.10}$$

$$\sqrt{m}\big(H_{\text{inc}} - H_0\big) \xrightarrow{D} \mathcal{N}(0, \Sigma_{\text{inc}}), \tag{3.11}$$

with positive definite matrices Σ_{block} and Σ_{inc}.

Remark 3.3. It is possible to derive formulas for Σ_{block} and Σ_{inc}. Nevertheless, we omit these considerations as they are technical and are not used in the following work.

Proof. The statement for H_{block} is a direct consequence of the multidimensional Central Limit Theorem. The LHS of (3.10) can be written as

$$\sqrt{n/B}\left(\frac{1}{n/B}\sum_{b=1}^{n/B}\begin{pmatrix}\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^{b,(1)}, Y^b) \\ \vdots \\ \widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^{b,(p)}, Y^b)\end{pmatrix} - \begin{pmatrix}\mathrm{HSIC}(X^{(1)}, Y) \\ \vdots \\ \mathrm{HSIC}(X^{(p)}, Y)\end{pmatrix}\right).$$

The n/B random variables in the sum are independent and identically distributed due to the iid. assumption and data subdivision. Moreover, the involved estimators are unbiased and $n/B \to \infty$.
The proof for H_{inc} is deferred to Appendix 3.A. □

3.2.3 HSIC-Lasso

We consider a regression or classification setting for an independent identically distributed data sample $\{(x_j^{(1)}, \dots, x_j^{(p)}, y_j)\}_{j=1}^n$. The task of detecting covariates that are influential on the response is a common problem but particularly difficult without additional assumptions, such as linearity or $p < n$. (Makoto Yamada, Jitkrittum, et al. 2014) propose a method for this setting which is based on the Hilbert-Schmidt independence criterion, does not require any assumptions on the model and scales well for high-dimensional data.

Definition 3.18. Let $\{(x_j^{(1)}, \dots, x_j^{(p)}, y_j)\}_{j=1}^n$ be an iid. data sample. Using the notation of Definition 3.13 with the respective kernels, set $\bar{L} = \Gamma L \Gamma$ and $\bar{K}^{(k)} = \Gamma K^{(k)} \Gamma$ for $k \in \{1, \dots, p\}$. Let $\hat{\beta}$ be given by

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}_+^p} \frac{1}{2}\|\bar{L} - \sum_{k=1}^{p} \beta_k \bar{K}^{(k)}\|_{\text{Frob}}^2 + \lambda\|\beta\|_1,$$

where $\lambda > 0$ is the tuning parameter of the Lasso penalty and $\mathbb{R}_+$ denotes the non-negative real numbers. The *HSIC-Lasso selection procedure* picks covariates whose corresponding $\hat{\beta}$-entries are positive.

Remark 3.4. The good computational properties of HSIC-Lasso mainly stem from the fact that the Frobenius norm can be rewritten as L^2-norm by vectorisation of $\bar{L}$ and $\bar{K}^{(k)}, k \in \{1, \ldots, p\}$. This yields a common case of a restricted Lasso problem for which many off-the-shelf algorithms are available.

The functioning of HSIC-Lasso can be best understood considering an alternative representation.

Lemma 3.19. *In the setting of Definition 3.18, it holds that*

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p_+}{\operatorname{argmin}} - \sum_{k=1}^{p} \beta_k \, \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(k)}, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \beta_k \beta_l \, \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(k)}, X^{(l)}) + \lambda \|\beta\|_1. \tag{3.12}$$

We see that the dependence between response and covariates is captured by estimates of the Hilbert-Schmidt independence criterion which renders HSIC-Lasso model-free.
Moreover, it becomes apparent that three competing components are present in the optimisation problem. Considering the first and third term together, a highly influential covariate $X^{(k)}$ yields a large value of $\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(k)}, Y)$ which also induces $\hat{\beta}_k$ to become large in order to minimise the whole expression. Conversely, an uninfluential covariate leads to an HSIC-estimate close to zero and the respective entry in $\hat{\beta}$ tends to be forced to zero by the regularisation term. As for the second term, we note that the dependence structure among the covariates is also taken into account pushing the $\hat{\beta}$-entries of highly dependent covariates to zero.
We expect two different effects arising. Among influential covariates, highly dependent and thus redundant variables tend to be sorted out which leads to a more parsimonious model. In the case of an uninfluential covariate being highly dependent on an influential covariate, the first term is able to correctly assess the independent variable in the limit $n \to \infty$. For small or moderate sample sizes however, strong dependence on an influential variable can lead the selection procedure astray. In this situation, the second term is able to mitigate this effect by punishing dependence among covariates in the selection process.
(M. Yamada et al. 2018) and (Climente-González et al. 2019) develop algorithms for HSIC-Lasso that scale well for ultra high-dimensional data. (Takahashi et al. 2020)'s work constitutes a good reference for a case study.
The idea of HSIC-Lasso is further generalised and analysed by (Poignard and Makoto Yamada 2020) considering other regularisations terms, e.g. bridge or SCAD penalties, and establishing the oracle property.

3.3 Post-selection inference with HSIC-Lasso

Bringing together the results of Subsections 3.1 and 3.2, we develop an approach of post-selection inference for the HSIC-Lasso selection procedure.

3.3.1 HSIC-Lasso adaptation

Taking the representation (3.12) of HSIC-Lasso as a starting point, we make two changes in order to satisfy the requirements of Theorem 3.3. Since Gaussianity is demanded, we replace the biased HSIC-estimator with an asymptotically normal one, namely the block or incomplete U-statistics estimator, see Definition 3.16. Moreover, it is possible to use a weighted Lasso penalty without complicating the coming considerations.

Definition 3.20. In the setting of Definition 3.18, the *normal weighted HSIC-Lasso selection procedure* is given by

$$\begin{aligned}\hat{\beta} &= \underset{\beta\in\mathbb{R}^p_+}{\operatorname{argmin}} -\sum_{k=1}^{p} \beta_k\, H(X^{(k)},Y) + \frac{1}{2}\sum_{k,l=1}^{p} \beta_k\beta_l\, \tilde{H}(X^{(k)},X^{(l)}) + \lambda\,\beta^T w \\ &=: \underset{\beta\in\mathbb{R}^p_+}{\operatorname{argmin}} -\beta^T H + \frac{1}{2}\,\beta^T M\beta + \lambda\,\beta^T w,\end{aligned}$$

where H is an asymptotically Gaussian and $\tilde{H}$ any HSIC-estimator, $\lambda > 0$ is the tuning parameter of the regularisation term and $w \in \mathbb{R}^p_+$ is a fixed weight vector.

Assuming that M is positive definite, we can reformulate the upper representation in terms of a Lasso-problem as follows

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p_+}{\operatorname{argmin}} \frac{1}{2}\|Y - U\beta\|_2^2 + \lambda\,\beta^T w. \tag{3.13}$$

U is determined by the Cholesky decomposition $M = U^T U$ and Y is the solution to $H = U^T Y$. This formulation facilitates the computation of the estimate as there is a variety of efficient algorithms and software packages for Lasso problems available.

Having developed a selection procedure that relies on asymptotically normal random variables, there is the need to establish an asymptotic version of Theorem 3.3. Moreover, we estimate the covariance matrix Σ and do not assume it as given. Hence, a statement of the following kind is desirable in order to theoretically underpin the coming steps:

$$F^{[\mathcal{V}^-(Z_n),\mathcal{V}^+(Z_n)]}_{\eta^T_{M_n}\mu,\,\eta^T_{M_n}\widehat{\Sigma}_n\eta_{M_n}}(\eta^T_{M_n}Y_n)|\{A_{M_n}Y_n \le b_{M_n}\} \xrightarrow{\mathrm{D}} \mathrm{Unif}\,(0,1), \quad \text{as } n\to\infty,$$

where $(Y_n)_{n\in\mathbb{N}}$ converges to a normal distribution, $(\widehat{\Sigma}_n)_{n\in\mathbb{N}}$ to its covariance matrix and $(M_n)_{n\in\mathbb{N}}$ denotes the sequence of selected models. (Ryan J. Tibshirani, Rinaldo, et al. 2018) concern themselves with a similar asymptotic statement; however, their set-up differs from ours inasmuch a standard linear regression model as well as knowledge of the true covariance matrix is assumed.
While the simulations of Subsection 3.4 strongly hint that the upper statement is actually true, the proof of such a theorem is beyond the scope of this work.

3.3.2 Selection and inference

Having introduced a HSIC-Lasso version that is suitable for post-selection inference based on truncated Gaussians, we proceed to define the quantities of interest for hypothesis testing, confidence interval construction and the like. They are referred to as inference targets in the following. Moreover, we introduce a new notation for a matrix $A \in \mathbb{R}^{q\times q}$ and index sets $I, J \subset \{1,\ldots,q\}$. $I^c := \{1,\ldots,q\} \setminus I$ and $A_{IJ} \in \mathbb{R}^{|I|\times|J|}$ is given by the rows and columns of A whose indices are contained in I and J respectively. Furthermore, we abbreviate $A_J := A_{\{1,\ldots,q\},J}$.

Definition 3.21. Let $\hat{S} = \{j\colon \hat{\beta}_j > 0\}$ be the model selection random variable associated with a normal weighted HSIC-Lasso according to Definition 3.20. Suppose it assumes the model S and let $j \in S$. Inference targets are defined by

$$\begin{aligned}
&\text{HSIC-target:} && H_j := \mathrm{e}_j^T H, && \\
&\text{partial target:} && \hat{\beta}_{j,S}^{\text{par}} := \mathrm{e}_j^T M_{SS}^{-1} H_S, && \\
&\text{full target:} && \hat{\beta}_{j}^{\text{full}} := \mathrm{e}_j^T M^{-1} H, && \\
&\text{carved target:} && \hat{\beta}_{j,I}^{\text{car}} := \mathrm{e}_j^T M_{II}^{-1} H_I, && \text{if } j \in I, \\
& && \hat{\beta}_{j,I}^{\text{car}} := \mathrm{e}_j^T M_{I\cup\{j\},I\cup\{j\}}^{-1} H_{I\cup\{j\}}, && \text{if } j \notin I,
\end{aligned}$$

where e_j is the j-th unit vector and $I \subset \{1, \dots, p\}$ is a model selected for a larger regularisation parameter λ.

The inference targets defined above follow two different rationale: the HSIC-target H_j characterises the dependence of X_j and Y regardless of other covariates, whereas the β-targets adjust to the entire dependence structure of a model. The difference among the latter targets is the size of the model that is taken into account.

Revisiting the general notation $\eta_M^T Y|\{\hat{M} = M\}$, (Fithian et al. 2014) point out that conditioning on the entire set $\{\hat{M} = M\}$ is not necessary in situations where η_M does not depend on all of the information captured by this event. In fact, the more information is contained in a selection event the wider confidence intervals tend to get. For this reason, it is advisable to only include the selection information affecting the inference target. Against this backdrop, we take a closer look at the β-targets.

The partial target is closely connected to regression coefficients in linear regression which becomes apparent by rewriting it in terms of the Lasso-representation (3.13):

$$\hat{\beta}_{j,S}^{\text{par}} = \mathrm{e}_j^T \left(U_S^T U_S\right)^{-1} U_S^T Y.$$

It accounts for the dependence structure in the selected model S and is thus a natural choice. However, $\hat{\beta}_{j,S}^{\text{par}}$ depends on the entire information of selected and not selected variables, which renders the minimal conditioning event $\{\hat{S} = S\}$ quite large.

The full and carved target are based on the ideas of (Liu et al. 2018) who propose the same quantities for linear regression. They can be interpreted as approximations of the partial target originating in the observation that confidence intervals for the partial regression coefficient are often unacceptably long.

The full target tackles this issue by accounting for the dependence structure among all variables, regardless of whether they were selected. Hence, the only selection information that $\hat{\beta}_j^{\text{full}}$ relies on is that j was included into the model and it consequently suffices to consider $\hat{\beta}_j^{\text{full}}|\{j \in \hat{S}\}$ for valid PSI.

The carved target reduces the information conditioned on by selecting a model I with a larger regularisation parameter λ and consequently fewer contained variables. This approach can be interpreted as identifying the most important covariates. $\hat{\beta}_{j,I}^{\text{car}}$ only accounts for the dependence structure between the most influential variables I and j. Therefore, inference on $\hat{\beta}_{j,I}^{\text{car}}$ only needs to consider the selection information $\{\hat{I} = I, j \in \hat{S}\}$.

Based on these considerations, we derive the truncation points associated with the different selection events.

Theorem 3.22. *Let $\hat{S} = \{j \colon \hat{\beta}_j > 0\}$ be the model selection of a normal weighted HSIC-Lasso according to Definition 3.20, assume that M is positive definite and let $\eta \in \mathbb{R}^p$. Then $\{\hat{S} = S\} = \{A\,(H_S, H_{S^c})^T \leq b\}$ with*

$$A = -\frac{1}{\lambda}\begin{pmatrix} M_{SS}^{-1} & \mid & 0 \\ M_{S^cS}M_{SS}^{-1} & \mid & \mathrm{Id} \end{pmatrix}, \qquad b = \begin{pmatrix} -M_{SS}^{-1}\, w_S \\ w_{S^c} - M_{S^cS}M_{SS}^{-1}\, w_S \end{pmatrix}, \tag{3.14}$$

where 0 denotes a matrix filled with zeros. The formulae of the respective truncation points $\mathcal{V}_S^-$ and $\mathcal{V}_S^+$ are stated in the polyhedral lemma 3.1.
The truncation points for the event $\{j \in \hat{S}\}$ are given by

$$\mathcal{V}_j^-(Z) = \frac{1}{\mathrm{e}_j^T C}\left[\mathrm{e}_j^T M \hat{\beta}_{-j} - \mathrm{e}_j^T Z + \lambda w_j\right], \qquad \mathcal{V}_j^+(Z) = \infty, \tag{3.15}$$

where $C = (\eta^T \Sigma \eta)^{-1}\Sigma\eta$ and $Z = (\mathrm{Id} - C\eta^T)H$.

Proof. Statement (3.14): W.l.o.g. we assume that the first $|S|$ covariates of $\{X_1, \dots, X_p\}$ were included into the model. The KKT conditions, see Definition 3.26, provide an equivalent characterisation of the solution to the Lasso-optimisation as M is positive definite and Slater's condition is clearly fulfilled, cf. Theorem 3.27. We obtain

$$\begin{gathered} 0 = -H + M\hat{\beta} + \lambda w - u, \\ \beta_j \geq 0, \qquad u_j \geq 0, \qquad \beta_j u_j = 0, \qquad \forall j \in \{1, \dots, p\}. \end{gathered}$$

In order to characterise $\{\hat{S} = S\}$, we partition the upper inequalities along S and S^c and obtain

$$\hat{\beta}_S = M_{SS}^{-1}(H_S - \lambda\, w_S), \tag{3.16}$$

$$0 \leq H_{S^c} + (M\hat{\beta})_{S^c} - \lambda\, w_{S^c}. \tag{3.17}$$

These results translate into two set of inequalities. First, all entries of $\hat{\beta}$ must be non-negative which implies

$$0 \leq M_{SS}^{-1}(H_S - \lambda\, w_S) \quad \Leftrightarrow \quad -\lambda^{-1}\left(M_{SS}^{-1} \,|\, 0\right) H \leq -M_{SS}^{-1}\, w_S.$$

Second, $M\hat{\beta} = M_S\hat{\beta}_S$ holds by definition of $\hat{S}$. Hence, we can plug (3.16) into (3.17) and obtain

$$\begin{aligned} & 0 \leq H_{S^c} + M_{SS^c}\big(M_{SS}^{-1}(H_S - \lambda\, w_S)\big) - \lambda\, w_{S^c} \\ \Leftrightarrow \quad & -\lambda^{-1}\left(M_{SS^c}M_{SS}^{-1} \,|\, \mathrm{Id}\right) H \leq w_{S^c} - M_{SS^c}M_{SS}^{-1}\, w_S. \end{aligned}$$

Both these set of inequalities describe the selection in an affine linear fashion. In this setting, we can use the polyhedral lemma 3.1 to get the truncation points $\mathcal{V}_S^-$ and $\mathcal{V}_S^+$.

Statement (3.15): We decompose H into a component in direction of η and one perpendicular to η

$$H = (\eta^T H) \cdot C + Z$$

Again, we apply the KKT conditions and obtain

$$0 = (\eta^T H) \cdot C - Z + M\hat{\beta} + \lambda w - u,$$

with $u \in \mathbb{R}^p_+$. Since $j \notin \hat{S} \Leftrightarrow \hat{\beta}_j = 0$ holds by definition of $\hat{S}$, the inequality

$$0 \leq \mathrm{e}_j^T \left[(\eta^T H) \cdot C - Z + M\hat{\beta}_{-j} + \lambda w \right], \tag{3.18}$$

where the j-th entry of $\hat{\beta}_{-j}$ is set to zero, ensues for this case. Rearranging (3.18), we find

$$\eta^T H \leq \frac{1}{\mathrm{e}_j^T C} \left[\mathrm{e}_j^T M \hat{\beta}_{-j} - \mathrm{e}_j^T Z + \lambda w_j \right]. \tag{3.19}$$

Consequently, for the event $\{j \in \hat{S}\}$ the lower truncation point $\mathcal{V}_j^-(Z)$ is the RHS of (3.19) and $\mathcal{V}_j^+(Z) = \infty$. □

Concluding this subsection, we summarise the truncation points for the proposed targets.

Proposition 3.23. *We use the terminology of Theorem 3.22 and assume w.l.o.g. that the first $|S|$ covariates of $\{X_1, \ldots, X_p\}$ were included into the model. The truncation points and η-vectors of the different inference targets given in Definition 3.21 are*

target	$\mathcal{V}^-$	$\mathcal{V}^+$	η
HSIC	$\mathcal{V}_j^-$	∞	e_j
partial	$\mathcal{V}_S^-$	$\mathcal{V}_S^+$	$\mathrm{e}_j^T (M_{SS}^{-1} \,\vert\, 0)$
full	$\mathcal{V}_j^-$	∞	$\mathrm{e}_j^T M^{-1}$
carved	$\max\{\mathcal{V}_j^-, V_I^-\}$	$\mathcal{V}_I^+$	$\mathrm{e}_j^T (M_{II}^{-1} \,\vert\, 0),$ *if* $j \in I$, $\mathrm{e}_j^T (M_{I\cup\{j\},I\cup\{j\}}^{-1} \,\vert\, 0),$ *if* $j \notin I$

Proof. The η-vectors of the targets' representation as $\eta^T H$ are a direct result of Definition 3.21 adding zero-entries where necessary.
The HSIC- and full target only depend on the information that j was selected by the normal weighted HSIC-Lasso. Hence, it suffices to condition on $\{j \in \hat{S}\}$ and Theorem 3.22 provides the suitable truncation points.
In contrast, the partial target depends on the entire information $\{\hat{S} = S\}$. We use the affine linear representation of the selection procedure as stated in the first half of Theorem 3.22 in combination with the polyhedral lemma 3.1 to get the truncation points.
The carved target is a combination of the previous two cases having the conditioning set $\{\hat{I} = I, j \in \hat{S}\}$. Consequently, we have to take the maximum of the lower truncation points $\mathcal{V}_j^-$ and $\mathcal{V}_I^-$ and the minimum of the upper truncation points $\mathcal{V}_j^+ = \infty$ and $\mathcal{V}_I^+$. □

3.3.3 Practical application

While the previous subsection has established the theoretical framework for valid post-selection inference with a normal weighted HSIC-Lasso selection procedure, we now concentrate on developing an algorithm that handles difficulties arising in practical application.

Positive definiteness Throughout the foregoing arguments it was assumed that the matrix M defined by $M_{ij} = \tilde{H}(X^{(i)}, X^{(j)}), i, j \in \{1, \ldots, p\}$, where $\tilde{H}$ is a HSIC-estimator, is positive definite. This was needed to ensure that the KKT conditions provide an equivalent characterisation of the solution of the optimisation problem. In the original formulation of HSIC-Lasso with the biased HSIC-estimator, see Definition 3.18, this requirement is fulfilled because the function to be optimised is convex. However, there is no theoretical guarantee for other estimators.
For this reason, we use a positive definite approximation $\tilde{M}$ as proposed by (Higham 1988). The spectral decomposition of M is computed and all negative eigenvalues are replaced with a small positive value $\varepsilon > 0$. In many applications the approximation turns out to be very close to M.

High computational costs A strong point of using HSIC-Lasso is its applicability to problems where the number of covariates p exceeds the sample size n. However, treating high-dimensional or ultra high-dimensional data poses considerable computational costs. These can be traced back to the calculation of the HSIC-estimates where H grows as $\mathcal{O}(p)$ and M as $\mathcal{O}(p^2)$.
A straightforward solution to this problem is the introduction of a screening stage before the model selection procedure developed in the previous subsection is applied. In this upstream step, a subset of potentially influential covariates is determined so that the HSIC-Lasso procedure only has to deal with these variables. (Makoto Yamada, Umezu, et al. 2018) propose a simple selection procedure that can be used for screening. HSIC-estimates $\widehat{\text{HSIC}}(Y, X^{(k)}), k \in \{1, \ldots, p\}$, are computed and a pre-fixed number $P < p$ of the covariates having the highest HSIC-estimates is selected. In order to ensure valid inference results, we have to adjust for the screening step as well because it is a type of model selection.
Two solutions come to mind. (Makoto Yamada, Umezu, et al. 2018) show that the selection in the screening step allows for an affine linear representation. Consequently, we can apply the polyhedral lemma 3.1 to derive additional truncation points that possibly restrict an inference target. This approach is based on the same idea as (Jason. D Lee and J. E. Taylor 2014)'s work where a combination of marginal screening and post-selection inference for a linear regression setting is investigated. The presented procedure enjoys the advantage of using the entire data for the screening step. Yet, one has to compromise on the detection capabilities of the HSIC-estimates as only asymptotically normal estimators can be used.
Another, simple approach is splitting the data into two folds, one dedicated to screening, the other dedicated to HSIC-Lasso selection among the screened variables, cf. (Cox 1975). Thus, we avoid taking the screening step into account for post-selection inference. For this reason, we can use the unbiased HSIC-estimator which is more precise than the block

or incomplete U-statistics estimator. However, both screening and HSIC-Lasso selection can only operate on a part of the data.

Hyperparameter choice Until now we treated the regularisation parameter λ and the weight vector w as given. In practice, however, a good choice of these quantities is indispensable for meaningful results. Since we do not know the data generating process, we have to make a data-dependent choice, or rather estimate the hyperparameters. Similar to screening, this is a type of model selection which we consequently have to account for to guarantee valid inference results.
(Loftus 2015) addresses this issue for cross-validation by deriving a conditioning set of quadratic constraints that describes the selection procedure of λ. Yet, this does not fit our framework of affine linear inequalities and is thus more suitable for truncated χ^2-distributions than Gaussians. (Markovic et al. 2017) are concerned with post-selection inference for a family of hyperparamter selection methods, including cross-validation, see for example (Stone 1974), and the Akaike information criterion put forward by (Akaike 1974), and shows how to obtain an asymptotic pivot. Yet, this approach is intricate and relies on assumptions that might not be fulfilled in some applications.
Continuing the idea of sample splitting for the screening step, we can use the first fold for hyperparameter selection as well. In doing so, we do not have to restrict ourselves to certain selection methods for the regularisation parameter and get an easy to implement and valid procedure.
Moreover, we can employ (Zou 2006)'s adaptive Lasso penalty, that uses the weight vector $w = 1/|\hat{\beta}|^{\gamma}$. γ is typically set to 1.5 or 2 and $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator, e.g. the ordinary least square estimator. Contrary to the vanilla Lasso, this method is shown to be asymptotically consistent for variable selection because the adaptive nature leads to less penalisation for influential covariates and more penalisation for non-influential variables. As major drawback, this property was only proven for a covariance matrix $\Sigma = \sigma^2\text{Id}$. Hence, we have to evaluate the usefulness of the adaptive Lasso in simulations.

We summarise our proposed PSI method for a normal (weighted) HSIC-Lasso selection procedure in Algorithm 3. For the sake of brevity, we only state the procedure for one-sided hypothesis testing (3.5) and the HSIC- and partial target. In addition, several choices, like the covariance estimator or the hyperparameters of estimators, are suppressed.

3.4 Investigation on artificial data

In this subsection we examine the performance of Algorithm 3 for different parameters on artificial data and, ultimately, compare it with other methods for model-free PSI. The source code for the experiments carried out was implemented in Python and is available on Github. It harnesses (Lim et al. 2020)'s `mskernel`-package for the calculation of the HSIC-estimates and makes heavy use of the Lasso optimisation routines of `scikit-learn`. The latter package uses the cyclical gradient descent algorithm, see for example (Friedman et al. 2007), and the least angle regression algorithm (LARS) proposed by (Efron, Hastie, et al. 2004).

Algorithm 3: Post-selection inference for HSIC-Lasso selection

```
Input: data $(\mathbf{X}^n, \mathbf{Y}^n)$, level $\alpha$, inference target $t$, split ratio $s$, number of
       screened variables $P$, screen-, $M$- and $H$-estimators $\mathrm{e}_s, \mathrm{e}_M, \mathrm{e}_H$
Output: significant variables $I_{sig}$
$(\mathbf{X}^{n,1}, \mathbf{Y}^{n,1}), (\mathbf{X}^{n,2}, \mathbf{Y}^{n,2}) \leftarrow \mathrm{split}((\mathbf{X}^n, \mathbf{Y}^n), s)$
                                                                              // 1st fold
$H^1 \leftarrow \mathrm{estimate}_H(\mathbf{X}^{n,1}, \mathbf{Y}^{n,1}, \mathrm{e}_s)$
$I_{sc} \leftarrow \mathrm{screening}(H^1, P)$
$M^1 \leftarrow \mathrm{estimate}_M(\mathbf{X}^{n,1}_{I_{sc}}, \mathrm{e}_s)$
$\tilde{M}^1 \leftarrow \text{positive-definite-approximation}(M^1)$
$U_1 \leftarrow \mathrm{cholesky}(\tilde{M}^1);\ Y_1 \leftarrow U_1^{-T} H^1_{I_{sc}}$
$\lambda \leftarrow \text{cross-validation}(U_1, Y_1)$ // or AIC
$w \leftarrow \mathrm{weights}(U_1, Y_1)$
                                                                              // 2nd fold
$H^2 \leftarrow \mathrm{estimate}_H(\mathbf{X}^{n,2}_{I_{sc}}, \mathbf{Y}^{n,2}_{I_{sc}}, \mathrm{e}_H)$
$M^2 \leftarrow \mathrm{estimate}_M(\mathbf{X}^{n,2}_{I_{sc}}, \mathrm{e}_M)$
$\tilde{M}^2 \leftarrow \text{positive-definite-approximation}(M^2)$
$\hat{\Sigma} \leftarrow \mathrm{estimate}_\Sigma(H^2)$
$U_2 \leftarrow \mathrm{cholesky}(\tilde{M}^2);\ Y_2 \leftarrow U_2^{-T} H^2$
$\hat{\beta} \leftarrow \text{lasso-opimisation}(Y_2, U_2, \lambda, w)$
$S \leftarrow \text{non-zero-indices}(\hat{\beta})$
$I_{sig} \leftarrow \emptyset$
if $t$ is partial target then
    $A \leftarrow -\lambda^{-1} \begin{pmatrix} (\tilde{M}^2_{SS})^{-1} & | & 0 \\ \tilde{M}^2_{S^cS}(\tilde{M}^2_{SS})^{-1} & | & \mathrm{Id} \end{pmatrix};\quad b \leftarrow \begin{pmatrix} -(\tilde{M}^2_{SS})^{-1} w_S \\ w_{S^c} - \tilde{M}^2_{S^cS}(\tilde{M}^2_{SS})^{-1} w_S \end{pmatrix}$
foreach $j \in S$ do
    if $t$ is HSIC-target then
        $\eta \leftarrow \mathrm{e}_j$
        $C \leftarrow (\eta^T \hat{\Sigma} \eta)^{-1} \hat{\Sigma}\, \eta;\quad Z \leftarrow (\mathrm{Id} - C\eta^T) H^2$
        $\mathcal{V}_j^- \leftarrow (\mathrm{e}_j^T C)^{-1} \left[ \mathrm{e}_j^T \tilde{M}^2 \hat{\beta}_{-j} - \mathrm{e}_j^T Z + \lambda w_j \right];\quad \mathcal{V}_j^+ \leftarrow \infty$
    else if $t$ is partial target then
        $\eta \leftarrow \mathrm{e}_j^T ((\tilde{M}^2_{SS})^{-1} \,|\, 0)$
        $\mathcal{V}^-, \mathcal{V}^+ \leftarrow \text{truncation-points}(A, b, \eta, \hat{\Sigma})$
    ... // full and carved target
    $p \leftarrow 1 - F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0, \eta^T \hat{\Sigma} \eta}(\eta^T H)$
    if $p \leq \alpha$ then
        $I_{sig} \leftarrow I_{sig} \cup \{j\}$
return $I_{sig}$
```

Throughout the experiments, we use the toy model

$$\begin{aligned} Y &= (X_1 - 1)\tanh(X_2 + X_3) + \operatorname{sign}(X_4) + \varepsilon, \\ X &\sim \mathcal{N}(0_{500}, \Xi), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3.20}$$

where $0_{500} \in \mathbb{R}^{500}$ and $\Xi \in \mathbb{R}^{500\times 500}$, to generate data. This is a clearly non-linear and high-dimensional setting which is prone to overfitting due to the large number of non-influential covariates. We choose σ^2 such that the variance of ε is a fifth of the variance of the X-dependent terms of Y amounting to a noise-to-signal ratio of 0.2. For the covariance matrix Ξ three different cases are considered: we either set $\Xi = \mathrm{Id}$, use decaying correlation, i.e. $\Xi_{ij} = 0.3^{|i-j|}$ for $i, j \in \{1, \dots, 500\}$, or constant correlation, that is $\Xi_{ij} = 0.1 + 0.9\,\delta_{ij}$ for $i, j \in \{1, \dots, 500\}$.
In the following, we use Gaussian kernels where we choose the bandwidth parameter according to the median heuristic, cf. (Schölkopf and A. J. Smola 2018). For Algorithm 3, we fix $\alpha = 0.05$, use a quarter of the data for the first fold ($s = 0.25$) and screen 50 covariates using the unbiased HSIC-estimator. For the estimation of the matrix M on the second fold we use the block estimator with $B = 10$ as it is computationally less expensive than the unbiased estimator and leads to similar results. We estimate the covariance matrix Σ of H based on the summands of the block (3.8) and incomplete U-statistics (3.9) estimator respectively. To this end, we use the oracle approximating shrinkage (OAS) estimator, which was presented in (Y. Chen et al. 2010) and is particularly tailored for high-dimensional Gaussian data. Running the following experiments with the empirical covariance estimator instead, however, leads to very similar results.

The first experiment investigates different methods of selecting the hyperparameter λ and the weight vector w. To begin with, we assure ourselves that choosing only a quarter of the data for the first fold and thus for screening is actually sufficient to include the influential covariates into the set that is further considered with high probability. Figure 10 shows that for all three considered covariance settings, identity, decaying correlation and constant correlation, the screening procedure is usually successful. This justifies choosing $s = 0.25$.

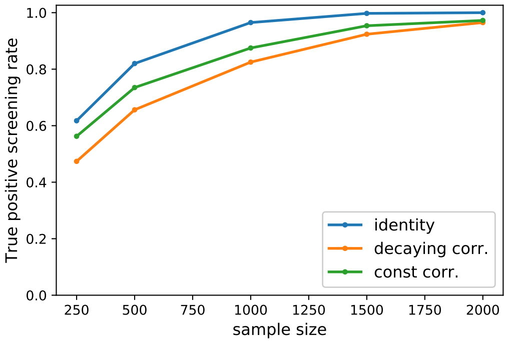

Figure 10: Rate of influential covariates detected by screening. For each of the three covariance settings and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ the true positive screening rate was calculated on 200 simulated datasets.

In this experiment we calculate H with the block estimator with size $B = 10$. For the selection of the hyperparameter λ, we investigate two widely used methods, the Akaike information criterion and cross-validation. The former approach relies on the number of degrees of freedom which is estimated according to (Zou et al. 2007). For the latter approach, we use standard 10-fold cross-validation. Moreover, we can set the entries of w to 1 resulting in a standard Lasso-penalty or we can choose them adaptively. In this case, we use the ordinary least squares estimator $\hat{\beta}$ and $\gamma = 1.5$ to calculate $w = 1/|\hat{\beta}|^{\gamma}$. Figure 11 shows that the average number of selected covariates decreases with increasing sample size across all considered covariance settings and methods for choosing λ and w. Moreover, we observe that setting w in an adaptive way leads to an overly sparse selection as fewer than the actually four influential variables are selected on average. This finding disqualifies the adaptive Lasso penalty. In contrast, non-adaptive AIC and cross-validation entail reasonable selection and often behave similarly; only for the constant correlation setting a major difference can be observed.

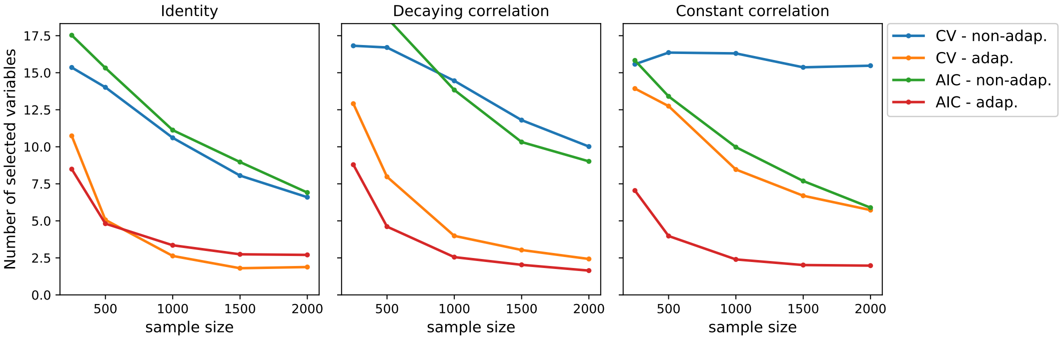

Figure 11: Average number of variables selected where λ is either chosen by cross-validation or AIC and w is either non-adaptive or adaptive. For each of the three covariance settings and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

In the following experiments, we use non-adaptive cross-validation as it is the most common technique for hyperparameter selection and the estimation of degrees of freedom for the Akaike information criterion is subject to some debate, see further (Kaufman and Rosset 2014) and (Janson et al. 2015).

The second experiment is concerned with assessing the performance of different estimators for H. To this end, we use the set-up of the first experiment but now consider block estimators of size 5 and 10 and incomplete U-statistics estimators with $l = 1$ and $l = 5$. In order to evaluate the estimators we use the false positive rate (FPR) and true positive rate (TPR) that they exhibit for the HSIC-, partial, full and carved inference targets. Let S be the set of selected indices, let t be any of the stated inference targets and consider testing $\mathrm{H}_0\colon\ t_j = 0$ against $\mathrm{H}_1\colon\ t_j > 0$ for $j \in S$. We define $I_0 := \{j \in S\colon t_j = 0\}$, $I_1 := \{j \in S\colon t_j > 0\}$ and $R := \{j \in S\colon \mathrm{H}_0 \text{ rejected}\}$. Then, the false positive and true positive rate are given by

$$\mathrm{FPR} = \mathrm{E}\left[\frac{|I_0 \cap R|}{|I_0|}\right], \quad \mathrm{TPR} = \mathrm{E}\left[\frac{|I_1 \cap R|}{|I_1|}\right].$$

Since FPR quantifies the rejection of true null hypotheses, we expect to observe values close to the type-I error α. In contrast, TPR indicates the power of a procedure, or rather the probability of avoiding type-II errors. Figure 12 shows that, even for small sample sizes, all estimators lead to false positive rates close to the desired value of 0.05 for the partial inference target. Noticeably, the incomplete U-statistics estimator with $l = 5$ exhibits the most volatile FPR. We do not present the plots for the other three targets because they show very similar results.

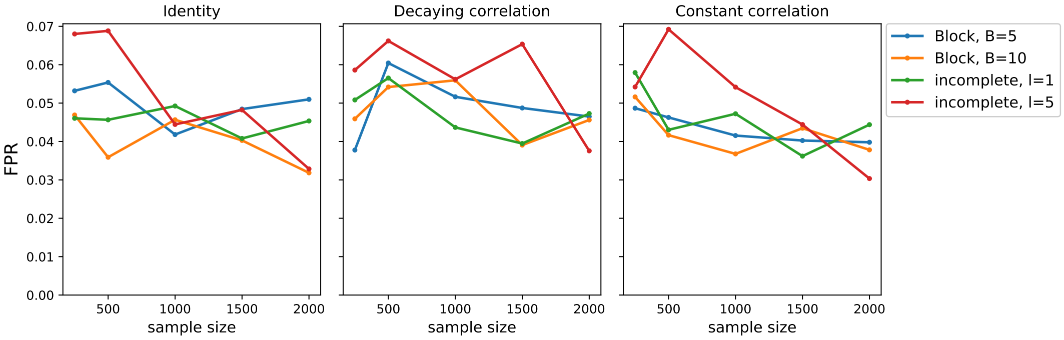

Figure 12: FPRs of different estimators for the partial target. For each of the three covariance settings and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

Proceeding to the analysis of the true positive rate, we consider the Figures 13 and 14. For the sample size $n = 250$, we notice that some data points are missing which can be attributed to influential variables not being selected and a lack of null hypothesis rejections. Both figures show a clear ranking of the examined estimators in terms of their power. The incomplete U-statistics estimator with $l = 5$ performs best by a considerable margin but is also costly in its computation. For this reason, the block estimator of size 10 emerges as a good choice in order to strike a balance between power and computational efficiency. We omit the plots for the full and carved inference targets as they are in line with the results for the HSIC- and partial target respectively.

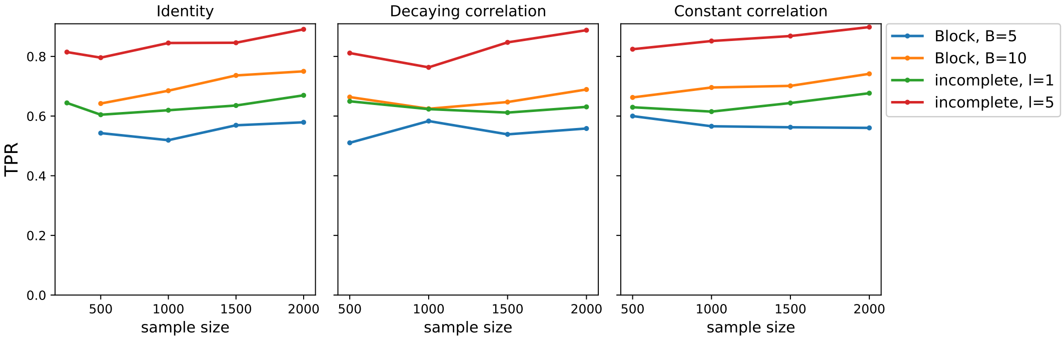

Figure 13: TPRs of different estimators for the HSIC-target. For each of the three covariance settings and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

The third experiment takes a closer look at the functioning of the HSIC-Lasso selection procedure which is designed to punish picking covariates that are dependent. To this end, we use the identity and constant correlation setting from the previous experiments and additionally correlate certain pairs of variables by setting the respective entries of Ξ to 0.7. Figure 15 depicts the selection rates of individual covariates when $\Xi = \mathrm{Id}$.

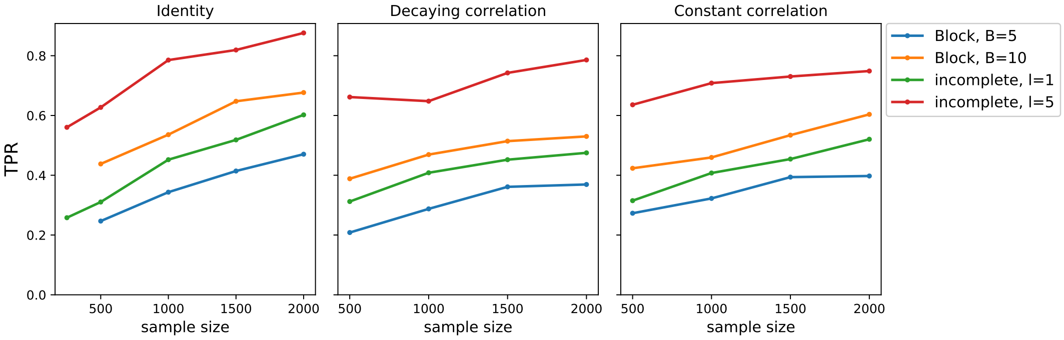

Figure 14: TPRs of different estimators for the partial target. For each of the three covariance settings and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

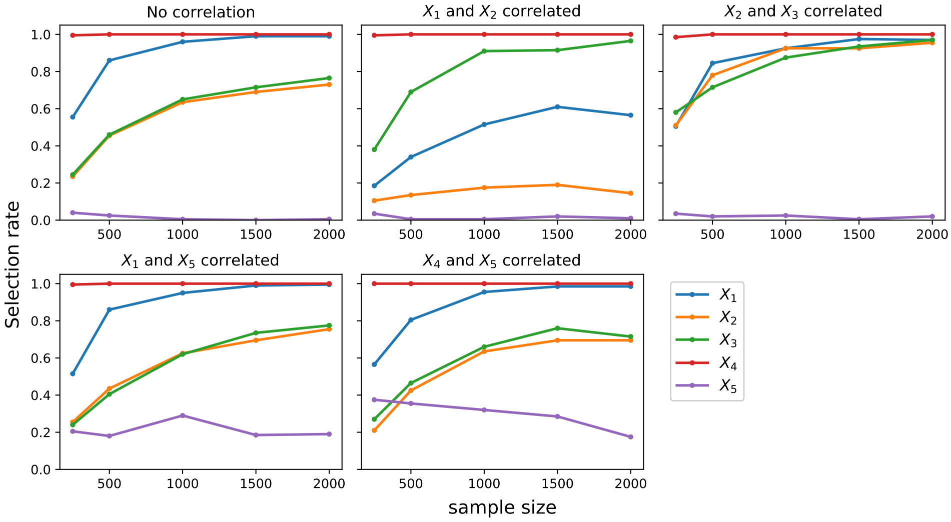

Figure 15: Selection rates of single variables for $\Xi = \mathrm{Id}$ and different correlated pairs. For each of them and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

Correlating two actually influential variables, X_1 and X_2, or X_2 and X_3, we get a seemingly ambiguous result. In the first case, both selection rates drop compared to the uncorrelated setting. In the second case, the rates for both X_2 and X_3 rise. This can be explained by the way how X_2 and X_3 influence the response. Since they appear in the term $X_2 + X_3$, positively correlating both summands induces the sum to take more extreme values and thus leads to stronger influence on Y. This is reflected in the increased selection rate. Similarly, the structure of (3.20) leads to lower selection rates of X_1 and X_2 when they are positively correlated.

Correlating an influential variable, such as X_1 and X_4, with the non-influential variable X_5, we observe that the selection rates of X_1 and X_4 are unaffected, whereas the selection rate of X_5 is noticeably larger compared to the uncorrelated setting. Yet, X_5 is selected less often than influential covariates.

In our experiments we find that the structure of the underlying data generating process outweighs the punishing effect of HSIC-Lasso selection among influential variables. However, unwanted picking of uninfluential covariates is moderately subdued. Figure 16 corroborates these results.

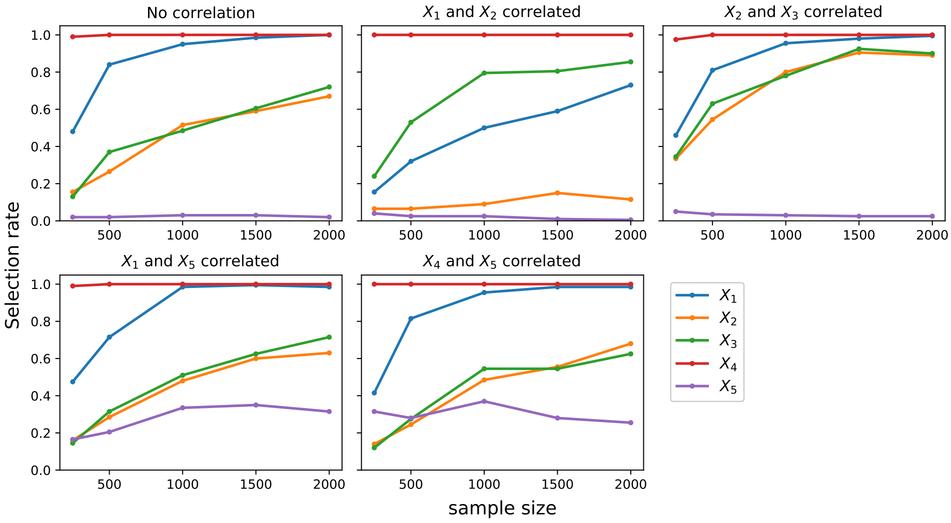

Figure 16: Selection rates of single variables for $\Xi_{ij} = 0.1 + 0.9\,\delta_{ij}$ and different correlated pairs. For each of them and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

Next, we examine testing outcomes for different targets against the backdrop of correlated pairs of covariates. The null hypothesis H_0 states that the respective target is zero; the alternative assumes a positive value. In the interest of brevity, we only show the plots for the setting $\Xi = \mathrm{Id}$ and omit the graphics for $\Xi_{ij} = 0.1 + 0.9\,\delta_{ij}$ as the latter corroborate the findings of the former in every situation. The H_0-rejection rate of the HSIC-targets matches the respective selection rates of the associated covariates as Figure 17 shows. This finding is not surprising as the HSIC-target captures the dependence between a certain covariate and the response, and the selection is mainly driven by this relationship. Therefore, this target is highly susceptible to confounding.

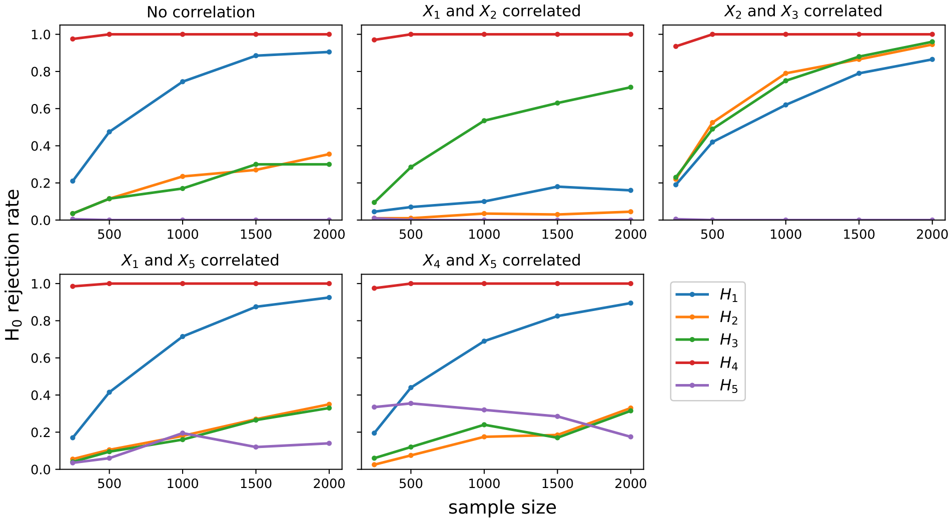

Figure 17: H_0-rejection rates of single HSIC-targets for $\Xi = \mathrm{Id}$ and different correlated pairs. For each of them and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

To the contrary, the partial, full and carved targets take the dependence structure of the

covariates into account which render them more natural choices for inference than the HSIC-target as the HSIC-Lasso selection procedure was developed to reduce confounding. As a note of caution, similarly to partial regression coefficients in a standard linear setting, the partial and carved targets are only sensible within the frame of the chosen model. Hence, they cannot be interpreted across models. For instance, $\hat{\beta}^{\text{par}}_{j,S} > 0$ might hold true in a correctly specified model S, but $\hat{\beta}^{\text{par}}_{j,\tilde{S}}$ can be zero in a model $\tilde{S}$ that does not include some influential variables. For this reason, the rates presented in the following should be interpreted with a grain of salt.
Considering correlated pairs of influential covariates, namely X_1 and X_2 as well as X_2 and X_3, Figure 18 shows that for the latter pair the dependence does not find expression in higher H$_0$-rejection rates of the corresponding partial targets. In the former case, the rates of $\hat{\beta}^{\text{par}}_{1,S}$ and $\hat{\beta}^{\text{par}}_{2,S}$ drop compared to the setting where no correlation is present. Moreover, the null hypothesis rejection rate of $\hat{\beta}^{\text{par}}_{3,S}$ rises in this situation. In the scenario where an influential and a non-influential variable are correlated, the picture is clearer. The H$_0$-rejection rates of $\hat{\beta}^{\text{par}}_{4,S}$ and $\hat{\beta}^{\text{par}}_{3,S}$ do not decrease when X_5 is correlated with X_1 and X_4 respectively. Meanwhile, despite being selected in roughly 20% of the cases when correlated, the partial target associated with X_5 is almost always not considered significant. Since the plots for the full and carved target yield the same results, they are omitted.
Ergo, we have found that, on the stage of hypothesis testing, targets considering the dependence structure among the covariates are effective in detecting uninfluential covariates that are correlated with influential ones. However, the manner how H$_0$-rejection rates materialise in the case of two influential covariates that are correlated is not so obvious.

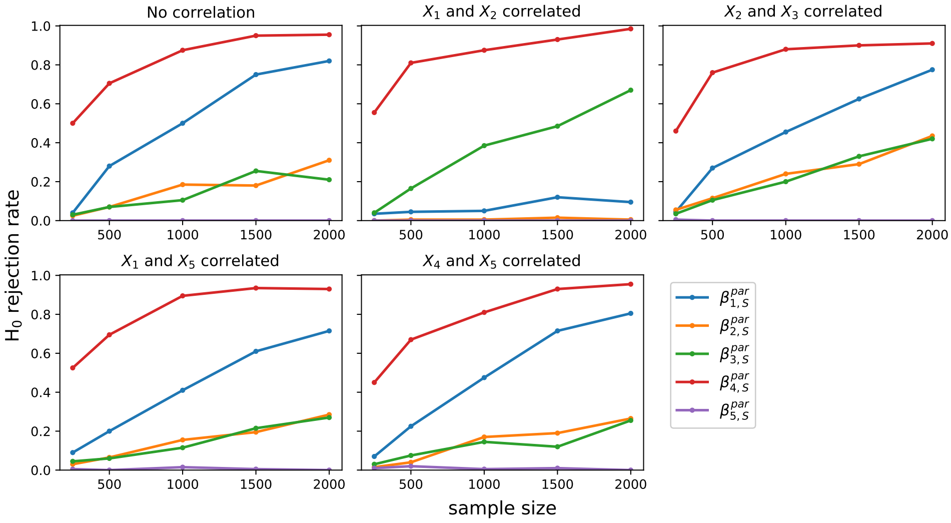

Figure 18: H$_0$-rejection rates of single partial targets for $\Xi = \text{Id}$ and different correlated pairs. For each of them and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 200 datasets were simulated.

The fourth experiment compares the developed approach with other methods for post-selection inference. To this end, we additionally use the logistic toy model

$$Y \sim \text{Bernoulli}\left(\frac{\exp(Z)}{1+\exp(Z)}\right), \quad Z = \sum_{i=1}^{5} X_i, \quad X \sim \mathcal{N}(0_{500}, \text{Id}), \tag{3.21}$$

and a standard linear model

$$Y = \sum_{i=5}^{5} X_i + \varepsilon, \quad X \sim \mathcal{N}(0_{500}, \text{Id}), \quad \varepsilon \sim \mathcal{N}(0, 1.5) \tag{3.22}$$

for data generation.

(Makoto Yamada, Umezu, et al. 2018) were the first to present a PSI-method for a selection procedure based on HSIC-estimates. An asymptotically normal estimator is used to calculate the values $H_1, \dots, H_p$, they are ordered by size and a predefined number k of the covariates with the largest HSIC-estimates is selected. In the following, we refer to this selection procedure as HSIC-ordering. Then, the pivotal quantity (3.4) is used to determine which of the chosen H-values are significant. Hence, the HSIC-targets of HSIC-Lasso and the inference targets of HSIC-ordering are identical. Building on the latter selection procedure, (Lim et al. 2020) apply a multiscale bootstrap method to carry out inference. This resampling approach was introduced by (Shimodaira 2004) and adapted for post-selection inference by (Terada and Shimodaira 2017, 2019). Ergo, multiscale bootstrapping is an alternative to the framework of truncated Gaussians and shown to be more powerful for HSIC-ordering, however, computationally expensive as well. In the following, we abbreviate this approach by Multi.

Moreover, when treating data from the linear toy model, it is justified to use a linear regression model, select a subset of covariates via Lasso and test the partial regression coefficients for significance with truncated Gaussians. This is precisely the set-up of (Jason D. Lee et al. 2016). According to (Negahban et al. 2012), we determine the regularisation parameter as follows

$$\lambda = 3\,\sigma\,\mathrm{E}\left[\|X^T\tilde{\varepsilon}\|_\infty\right], \quad \tilde{\varepsilon} \sim \mathcal{N}(0, \text{Id}_{n\times n}),$$

where $\sigma^2 = 1.5$ and n denotes the sample size. In doing so, we provide the selection procedure of the linear regression model with information about the variance of the error term that HSIC-Lasso and HSIC-ordering do not have. For this reason, the linear regression model, as used here, cannot be easily applied in practice and, therefore, merely serves as a benchmark.

In the following we consider four different data generating mechanisms: the non-linear problem (3.20) with $\Xi = \text{Id}$ and $\Xi_{ij} = 0.1 + 0.9\,\delta_{ij}$ for $i, j \in \{1, \dots, 500\}$, the logistic (3.21) and the linear (3.22) toy model. In each of these settings, we apply Algorithm 3 with both a block estimator, $B = 10$, and an incomplete U-statistics estimator, $l = 1$, and set the remaining parameters as in the previous experiments. Moreover, we use the Multi procedure with a block and incomplete U-statistics estimator and set $k = 30$. For linear data, we additionally assess the performance of the linear regression model.

In our simulations we find that the false positive rates of the HSIC-target are close to the envisaged value of $\alpha = 0.05$ for Algorithm 3 and lower for Multi, see Figure 19. For other targets of the HSIC-Lasso selection procedure as well as for the partial regression

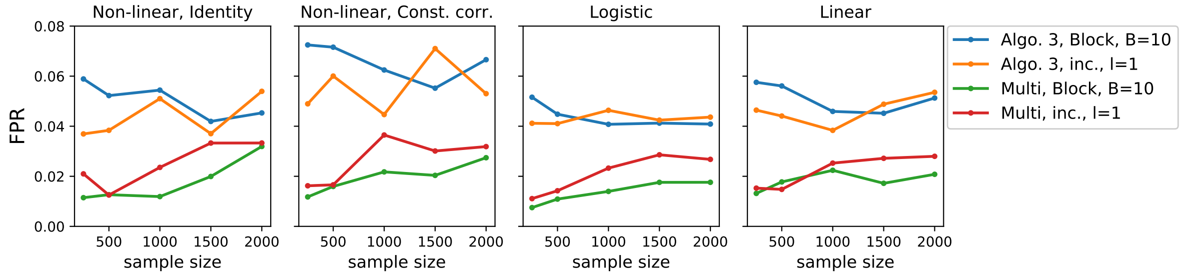

Figure 19: False positive rates of different selection procedures for the HSIC-target. For each of the different data-generating processes and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 100 datasets were simulated.

coefficient of the linear model, FPR is close to 0.05.

Turning to true positive rates, we observe that Algorithm 3 and Multi show a similar performance for the HSIC-target, albeit, in the case of logistic or linear data, the former is slightly superior for small sample sizes. The simulation results are depicted in Figure 20. Considering TPRs of the β-targets (partial, full and carved) of Algorithm 3, we find that the block estimator with $B = 10$ yields better results than the incomplete U-statistics estimator with $l = 1$ across all data generating models, see Figure 21. Moreover, the full target has a higher true positive rate than the partial or carved target. In the case of logistic and linear data, the TPR considerably grows with the sample size, whereas

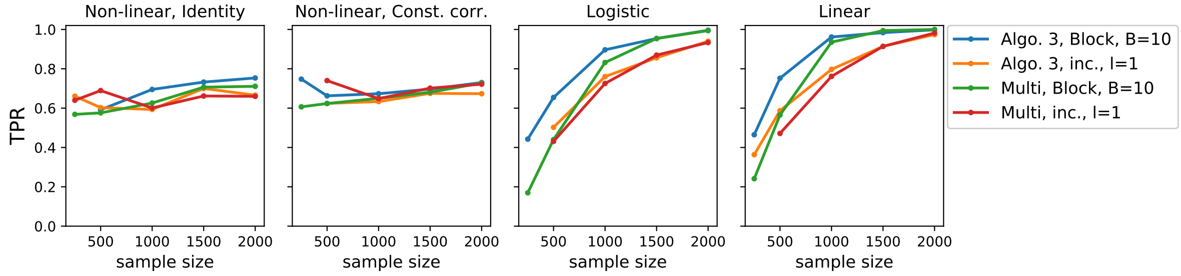

Figure 20: True positive rates of different selection procedures for the HSIC-target. For each of the different data-generating processes and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 100 datasets were simulated.

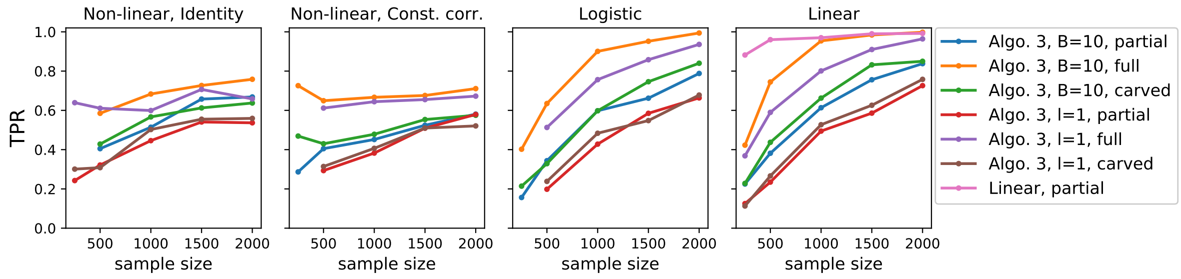

Figure 21: True positive rates of different selection procedures and β-targets. For each of the different data-generating processes and each sample size $n \in \{250, 500, 1000, 1500, 2000\}$ 100 datasets were simulated.

for the non-linear experiment this effect is feeble. In the setting of linear data, the TPR of the partial regression coefficient in a linear model is almost at 100%, even for small sample sizes. However, as n grows, all targets of Algorithm 3 reach high TPR-values as well which underlines the power of the presented model-free approach.

3.5 Performance on benchmark data

In order to conclude the analysis of the approach developed in this work, we apply it to two benchmark datasets from the UCI Repository and compare its performance to the Multi procedure, described in the previous subsection.

Turkish student dataset This dataset contains 5820 course evaluation scores provided by students from Gazi University, Ankara, see further (Gunduz and Fokoue 2013). Each student answered 28 questions on a Likert scale, meaning that the values are taken from $\{1, 2, 3, 4, 5\}$. For our experiment we use the the perceived difficulty of the course taking values in $\{1, 2, 3, 4, 5\}$ as response variable.

This data was previously evaluated by (Makoto Yamada, Umezu, et al. 2018) where a block estimator of size 10 was used to select 10 covariates with HSIC-ordering. Contrary to Multi, the subsequent inference was not based on multiscale bootstrapping but on the familiar framework of the polyhedral lemma and truncated Gaussians. Therefore, we denote this procedure Poly and report the obtained p-values together with our results. We employ Multi with the same parameters as Poly and also use Algorithm 3 where we set $s = 0.2$ and choose the Lasso regularisation parameter with 10-fold cross-validation. Since the number of features ($p = 28$) is manageable, we do not carry out screening. Moreover, we use the unbiased HSIC-estimator to calculate M and a block estimator of size 10 to get H. Table 2 summarises our findings.

First, we notice that Multi and Poly pick different features despite sharing the same selection procedure. We suspect that this is due to some randomisation carried out in (Makoto Yamada, Umezu, et al. 2018) which leads to different values of H. Moreover, we observe that HSIC-Lasso chooses a very parsimonious model with only four covariates whose associated HSIC-targets are highly significant. Among the tested approaches, there is moderate agreement on the influential covariates where Q17 and Q28 stand out as they are unanimously chosen and found to be significant.

For feature selection with HSIC-Lasso, we can additionally examine the significance of the β-targets that are associated with the four chosen features. Table 3 depicts the respective p-values. We see that only the partial, full or carved target linked with Q17 is found significant where Q11 exhibits low p-values as well.

Divorce predictors dataset This dataset consists of 170 samples with 54 features each, was collected by (Yöntem et al. 2019) and previously analysed with HSIC-based methods in (Lim et al. 2020). Participants of this study rated statements about their marriage on a scale from zero to four based on which we want to predict divorces. Since the response is consequently categorical, Gaussian kernels typically exhibit a bad performance. For this reason we use the the delta kernel on the Y-data which is tailored for classification

Feature description	p-value		
	Algo. 3	Multi	Poly
Q2: The course aims and objectives were clearly stated at the beginning of the period.	**0.021**	-	0.452
Q3: The course was worth the amount of credit assigned to it.	-	0.782	-
Q11: The course was relevant and beneficial to my professional development.	**0.004**	-	-
Q13: The Instructor's knowledge was relevant and up to date.	-	-	**0.018**
Q14: The Instructor came prepared for classes.	-	**0.001**	-
Q15: The Instructor taught in accordance with the announced lesson plan.	-	0.095	-
Q17: The Instructor arrived on time for classes.	$<$ **0.001**	$<$ **0.001**	**0.033**
Q18: The Instructor has a smooth and easy to follow delivery/speech.	-	-	0.186
Q19: The Instructor made effective use of class hours.	-	$<$ **0.001**	-
Q20: The Instructor explained the course and was eager to be helpful to students.	-	**0.004**	0.463
Q21: The Instructor demonstrated a positive approach to students.	-	**0.032**	**0.033**
Q22: The Instructor was open and respectful of the views of students about the course.	-	$<$ **0.001**	**0.042**
Q23: The Instructor encouraged participation in the course.	-	-	**0.037**
Q25: The Instructor responded to questions about the course inside and outside of the course.	-	**0.002**	-
Q26: The Instructor's evaluation system effectively measured the course objectives.	-	-	0.176
Q28: The Instructor treated all students in a right and objective manner.	**0.004**	**0.041**	$<$ **0.001**

Table 2: p-values of the HSIC-target for selected features of the Turkish student dataset calculated with different selection and inference procedures. (A hyphen signifies that a certain feature was not select.)

Associated feature	p-value		
	partial	full	carved
Q2	0.557	0.169	0.119
Q11	0.121	0.057	**0.008**
Q17	**0.008**	**0.044**	$<$ **0.001**
Q28	0.94	0.190	0.994

Table 3: p-values of different targets for features of the Turkish student dataset which were selected by HSIC-Lasso.

problems and given by

$$l(y, y') := \begin{cases} 1/n, & \text{if } y = y', \\ 0, & \text{otherwise,} \end{cases}$$

where n denotes the sample size. More details on kernels for classification problems can be found in (Song et al. 2012). Opposed to the Turkish student data, the analysis of the divorce predictors dataset is much harder as the sample size is rather low and the number of covariates comparatively high. We try the incomplete U-statistics estimator with a comparatively large size of $l = 15$ in order to increase detection power, and a block estimator with a low size of $B = 5$ to ensure that asymptotic normality can be assumed. Apart from this change, we use the same set-up of Algorithm 3 as for the Turkish student data and increase the number of selected variables for Multi from 10 to 15. Since the block estimator yields more plausible results for HSIC-Lasso, we only report these findings, see Table 4.

Considering HSIC-targets, we see that the features selected by Algorithm 3 and Multi respectively have a small overlap. Only the covariates 'We share the same views about being happy in our life with my spouse.' and 'We're just starting a discussion before I know what's going on.' are chosen and found significant by both procedures. The covariate 'My spouse and I have similar ideas about how roles should be in marriage.', which is also selected and found significant by the same PSI methods with the incomplete estimator, however, is not selected by Algorithm 3 with the block estimator. Among the β-targets, only the partial and carved target associated with 'When I discuss with my spouse, to contact him will eventually work.' reject the null hypothesis.

The diverging results can be explained by the inherent complexity of sociological data as well as the low sample size. It is likely that further hyperparameter tuning of Algorithm 3 and including domain-specific knowledge will improve the results.

Feature description	p-value	
	Algo. 3	Multi
When I discuss with my spouse, to contact him will eventually work.	**< 0.001**	-
The time I spent with my spouse is special for us.	-	**0.029**
We don't have time at home as partners.	0.398	-
I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.	-	**< 0.001**
We share the same views about being happy in our life with my spouse.	**< 0.001**	**0.002**
My spouse and I have similar ideas about how marriage should be.	-	0.671
My spouse and I have similar ideas about how roles should be in marriage.	-	**< 0.001**
I know my spouse very well.	-	0.210
I feel aggressive when I argue with my spouse.	**< 0.001**	-
I can use negative statements about my spouse's personality during our discussions.	-	**0.003**
I can insult my spouse during our discussions.	-	**< 0.001**
I can be humiliating when we have discussions.	-	**< 0.001**
My discussion with my spouse is not calm.	-	0.077
I hate my spouse's way of opening a subject.	-	**0.002**
Our discussions often occur suddenly.	-	0.320
We're just starting a discussion before I know what's going on.	**< 0.001**	**< 0.001**
When I talk to my spouse about something, my calm suddenly breaks.	-	**0.004**
Sometimes I think it's good for me to leave home for a while.	**< 0.001**	0.058
When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.	**0.001**	-

Table 4: p-values of the HSIC-target for selected features of the Divorce predictors dataset calculated with Algorithm 3 and Multi based on block estimates, $B = 5$. (A hyphen signifies that a certain feature was not select.)

Appendix 3.A U-statistics

U-statistics are a broad class of estimators pioneered by (Hoeffding 1948) that provide a framework to establish useful properties for a multitude of estimators. In the following, we use (A. J. Lee 1990) as a reference.

Definition 3.24. Let $X_1, \dots, X_n$ be independent random variables, which take values in a measurable space $(A, \mathcal{A})$ and share the same distribution, and let $h \colon A^k \to \mathbb{R}$ be a symmetric function. We denote $\mathcal{S}_{n,k}$ as the set of all k-subsets of $\{1, \dots, n\}$. For $n \geq k$,

$$U_n = \binom{n}{k}^{-1} \sum_{(i_1,\dots,i_k) \in \mathcal{S}_{n,k}} h(X_{i_1}, \dots, X_{i_k})$$

is a *U-statistic of degree k with kernel h.*

In order to prove the second statement of Theorem 3.17, we use an adaptation of the one-dimensional proof of asymptotic normality for an incomplete U-statistics estimator using random subset selection. Before commencing the proof, we state an auxiliary lemma.

Lemma 3.25. *Let $(a_i)_{i \in \mathbb{N}}$ be a sequence having the properties $lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} a_i = 0$ and $lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} a_i^2 = \sigma^2$ and the let the random variables $Z_1, \dots, Z_N$ have a multinomial distribution* $\mathrm{Mult}(m; N^{-1}, \dots, N^{-1})$. *Then as $m, N \to \infty$*

$$m^{-\frac{1}{2}} \sum_{i=1}^{N} a_i (Z_i - m/N) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

Proof of (3.11). In order to prove multidimensional convergence, we fall back on the Cramér-Wold device, i.e. it suffices to prove that

$$\sqrt{m}\, \nu^T \big(H_{\text{inc}} - H\big)$$

converges to a one-dimensional Gaussian distribution as $m \to \infty$ for any $\nu \in \mathbb{R}^p$.
We introduce the independent random vectors $Z^{(j)}, j \in \{1, \dots, p\}$ and index their entries with $\mathcal{S}_{n,4}$; hence, their elements are $\{Z_S^{(j)} \colon S \in \mathcal{S}_{n,4}\}$. All of them follow a multinomial distribution $\mathrm{Mult}(m; N^{-1}, \dots, N^{-1})$ where $N = \binom{n}{4}$. Hence, we can write

$$m^{\frac{1}{2}}\, \nu^T \big(H_{\text{inc}} - H\big) = m^{-\frac{1}{2}}\, \nu^T \sum_{S \in \mathcal{S}_{n,4}} Z_S \big(h(S) - H\big), \tag{3.23}$$

where the sum as well as the product within is to be understood componentwise, and $Z = (Z^{(1)}, \dots, Z^{(p)})$ as well as h are used in a vectorised way, slightly abusing notation. In order to derive the asymptotic distribution of (3.23), we consider its characteristic function ϕ_n. In the following manipulations we drop the indices for the summation $\sum_{S \in \mathcal{S}_{n,4}}$, introduce the notation $X^{(j)} = (X_1^{(j)}, \dots, X_n^{(j)}), j \in \{1, \dots, p\}$, and Y accordingly, and denote the p-dimensional vector of (complete) U-statistics by U_n, that is the vector of unbiased HSIC-estimators.

$$\begin{aligned}
\phi_n(t) &= \mathrm{E}\left[\exp\left(it\, m^{-\frac{1}{2}}\, \nu^T \sum Z_S\big(h(S) - H\big)\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\exp\left(it\, m^{-\frac{1}{2}}\, \nu^T \sum Z_S\big(h(S) - H\big)\right)\middle| X^{(1)}, \dots, X^{(p)}, Y\right]\right] \\
&= \mathrm{E}\Bigg[\exp\left(it\, m^{\frac{1}{2}} \sum\nolimits_{j=1}^{p} \nu_j U_n^{(j)}\right) \\
&\qquad \times \mathrm{E}\left[\exp\left(it\, m^{-\frac{1}{2}} \sum\nolimits_{j=1}^{p} \nu_j \sum \big(Z_S^{(j)} - m/N\big)\big(h_j(S) - H_j\big)\right)\middle| X^{(1)}, \dots, X^{(p)}, Y\right]\Bigg] \\
&= \mathrm{E}\Bigg[\exp\left(it\, m^{\frac{1}{2}} \sum\nolimits_{j=1}^{p} \nu_j U_n^{(j)}\right) \\
&\qquad \times \prod\nolimits_{j=1}^{p} \mathrm{E}\left[\exp\left(it\, m^{-\frac{1}{2}} \nu_j \sum \big(Z_S^{(j)} - m/N\big)\big(h_j(S) - H_j\big)\right)\middle| X^{(1)}, \dots, X^{(p)}, Y\right]\Bigg]
\end{aligned}$$

In the manipulations above we used the tower law of conditional expectation and the independence of the $Z_S^{(j)}, j \in \{1, \dots, p\}$. Moreover, we inserted $m\, U_n = m/N \sum h(S)$. Having separated the randomness coming from the data and the subset selection, we treat the second factor in the product above. Standard U-statistics theory implies that

$$\lim_{N\to\infty} N^{-1} \sum_{S\in\mathcal{S}_{n,4}} (h_j(S) - H_j) = 0 \quad \text{and} \quad \lim_{N\to\infty} N^{-1} \sum_{S\in\mathcal{S}_{n,4}} (h_j(S) - H_j)^2 = \sigma_j^2$$

almost surely where (Song et al. 2012) state a formula for σ_j^2. Ergo, the requirements of Lemma 3.25 are fulfilled and applying it together with the dominated convergence theorem yields

$$\begin{aligned}
\lim_{n\to\infty} \phi_n(t) &= \lim_{n\to\infty} \mathrm{E}\left[\exp\left(it\, m^{\frac{1}{2}} \sum\nolimits_{j=1}^{p} \nu_j U_n^{(j)}\right)\right] \prod_{j=1}^{p} \exp\big(-(\sigma_j \nu_j)^2 t^2/2\big) \\
&= \lim_{n\to\infty} \mathrm{E}\left[\exp\left(it\sqrt{m/n}\, \nu^T \big(\sqrt{n}\, U_n\big)\right)\right] \prod_{j=1}^{p} \exp\big(-(\sigma_j \nu_j)^2 t^2/2\big).
\end{aligned}$$

Using the multidimensional Central Limit Theorem for U-statistics, cf. (Korolyuk and Borovskikh 1994), and Slutsky's theorem 2.25, we arrive at

$$\lim_{n\to\infty} \phi_n(t) = \exp\big(-(\sqrt{r}\, \nu^T \Sigma \nu)\, t^2/2\big) \prod_{j=1}^{p} \exp\big(-(\sigma_j \nu_j)^2 t^2/2\big),$$

where Σ is a positive definite matrix. The limit of ϕ_n is clearly a Gaussian characteristic function which proves asymptotic normality. □

Appendix 3.B Karush-Kuhn-Tucker conditions

In order to find a suitable representation for selection events, the Karush-Kuhn-Tucker (KKT) conditions, which were independently found by (Karush 1939) and (Kuhn and Tucker 1951), are an essential tool. In the following, we use (Boyd and Vandenberghe 2004) as a reference and consider a reduced set-up that suffices for our intended application.

Definition 3.26. Let $f_0, \ldots, f_m \colon D \to \mathbb{R}$, $D \subset \mathbb{R}^n$, $int(D) \neq \emptyset$, be differentiable functions and consider the optimisation problem

$$\begin{aligned} &\text{minimise} && f_0(x), \\ &\text{subject to} && f_i(x) \leq 0, \quad i \in \{1, \ldots, m\}. \end{aligned} \tag{3.24}$$

The *Karush-Kuhn-Tucker (KKT) conditions* are given by

$$\begin{gathered} \nabla f_0(x) + \textstyle\sum_{i=1}^m u_i \nabla f_i(x) = 0 \\ f_i(x) \leq 0, \quad u_i \geq 0, \quad u_i f_i(x) = 0, \quad \forall\, i \in \{1, \ldots, m\}. \end{gathered}$$

The KKT conditions provide a handy characterisation of the solution of the optimisation problem (3.24).

Theorem 3.27. *In the situation of Definition 3.26, assume that f_0 is convex and Slater's condition holds, i.e. there exists $\tilde{x} \in int(D)$ such that $f_i(\tilde{x}) < 0$ for all $i \in \{1, \ldots, m\}$. Then the Karush-Kuhn-Tucker conditions are sufficient and necessary for optimality.*

References

Achard, Sophie, Dinh Tuan Pham, and Christian Jutten (2003). “Quadratic dependence measure for nonlinear blind source separation”. In: *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003*, pp. 263–268.

Akaike, Hirotugu (1974). “A new look at the statistical model identification.” In: *IEEE Trans. Autom. Control* 19, pp. 716–723.

Aronszajn, N. (1950). “Theory of reproducing kernels.” In: *Trans. Am. Math. Soc.* 68, pp. 337–404.

Bach, Francis R. and Michael I. Jordan (2003). “Kernel independent component analysis.” In: *J. Mach. Learn. Res.* 3.1, pp. 1–48.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao (2013). “Valid post-selection inference.” In: *Ann. Stat.* 41.2, pp. 802–837.

Borgwardt, Karsten M., Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola (2006). “Integrating structured biological data by Kernel Maximum Mean Discrepancy”. In: *Bioinformatics* 22.14, e49–e57.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex optimization.* Cambridge University Press, pp. xiii + 716.

Bühlmann, Peter (2013). “Statistical significance in high-dimensional linear models.” In: *Bernoulli* 19.4, pp. 1212–1242.

Chen, Yilun, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero (2010). “Shrinkage algorithms for MMSE covariance estimation.” In: *IEEE Trans. Signal Process.* 58.10, pp. 5016–5029.

Climente-González, Héctor, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada (2019). “Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data”. In: *Bioinformatics* 35.14, pp. i427–i435.

Cox, D. R. (1975). “A note on data-splitting for the evaluation of significance levels.” In: *Biometrika* 62, pp. 441–444.

Efron, Bradley (1979). “Bootstrap methods: another look at the jackknife.” In: *Ann. Stat.* 7, pp. 1–26.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004). “Least angle regression. (With discussion).” In: *Ann. Stat.* 32.2, pp. 407–499.

Efron, Bradley and Robert J. Tibshirani (1993). *An introduction to the bootstrap.* Vol. 57. New York, NY: Chapman & Hall, pp. xvi + 436.

Fisher, Sir Ronald (1956). “On a test of significance in Pearson’s Biometrika Tables (No. 11).” In: *J. R. Stat. Soc., Ser. B* 18, pp. 56–60.

Fithian, William, Dennis Sun, and Jonathan Taylor (2014). *Optimal Inference After Model Selection.* arXiv: `1410.2597 [math.ST]`.

Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani (2007). “Pathwise coordinate optimization”. In: *Ann. Appl. Stat.* 1.2, pp. 302–332.

Fukumizu, Kenji, Francis R. Bach, and Michael I. Jordan (2004). “Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces.” In: *J. Mach. Learn. Res.* 5, pp. 73–99.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola (2012). “A kernel two-sample test.” In: *J. Mach. Learn. Res.* 13, pp. 723–773.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf (2005). "Measuring statistical dependence with Hilbert-Schmidt norms." In: *Algorithmic learning theory. 16th international conference, ALT 2005, Singapore, October 8–11, 2005. Proceedings.* Berlin: Springer, pp. 63–77.

Gretton, Arthur, Alexander Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos Logothetis (2005). "Kernel Constrained Covariance for Dependence Measurement". In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 1–8.

Gunduz, Necla and Ernest Fokoue (2013). *UCI Machine Learning Repository.*

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical learning with sparsity. The Lasso and generalizations.* Vol. 143. Boca Raton, FL: CRC Press, pp. xv + 351.

Higham, Nicholas J. (1988). "Computing a nearest symmetric positive semidefinite matrix". In: *Linear Algebra and its Applications* 103, pp. 103–118.

Hocking, R. R. (1976). "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression". In: *Biometrics* 32.1, pp. 1–49.

Hoeffding, Wassily (1948). "A class of statistics with asymptotically normal distribution." In: *Ann. Math. Stat.* 19, pp. 293–325.

Hyun, Sangwon, Max G'sell, and Ryan J. Tibshirani (2018). "Exact post-selection inference for the generalized lasso path." In: *Electron. J. Stat.* 12.1, pp. 1053–1097.

Janson, Lucas, William Fithian, and Trevor J. Hastie (2015). "Effective degrees of freedom: a flawed metaphor". In: *Biometrika* 102.2, pp. 479–485.

Javanmard, Adel and Andrea Montanari (2014). "Confidence intervals and hypothesis testing for high-dimensional regression." In: *J. Mach. Learn. Res.* 15, pp. 2869–2909.

Karush, William (1939). "Minima of functions of several variables with inequalities as side conditions". MA thesis. Illinois, USA: Department of Mathematics, University of Chicago.

Kaufman, S. and S. Rosset (2014). "When does more regularization imply fewer degrees of freedom? Sufficient conditions and counterexamples". In: *Biometrika* 101.4, pp. 771–784.

Korolyuk, V. S. and Yu. V. Borovskikh (1994). *Theory of U-statistics. Updated and transl. from the Russian by P. V. Malyshev and D. V. Malyshev.* Dordrecht: Kluwer Academic Publishers, pp. ix + 552.

Kuhn, H. W. and A. W. Tucker (1951). "Nonlinear Programming". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, California: University of California Press, pp. 481–492.

Lee, A. J. (1990). *U-statistics. Theory and practice.* Vol. 110. New York etc.: Marcel Dekker, Inc., pp. xi + 302.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor (2016). "Exact post-selection inference, with application to the Lasso." In: *Ann. Stat.* 44.3, pp. 907–927.

Lee, Jason. D and Jonathan E Taylor (2014). "Exact Post Model Selection Inference for Marginal Screening". In: *Advances in Neural Information Processing Systems 27.* Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., pp. 136–144.

Leeb, Hannes and Benedikt M. Pötscher (2005). "Model selection and inference: facts and fiction." In: *Econom. Theory* 21.1, pp. 21–59.

— (2006). "Can one estimate the conditional distribution of post-model-selection estimators?" In: *Ann. Stat.* 34.5, pp. 2554–2591.

Lim, Jen Ning, Makoto Yamada, Wittawat Jitkrittum, Yoshikazu Terada, Shigeyuki Matsui, and Hidetoshi Shimodaira (2020). "More Powerful Selective Kernel Tests for Feature Selection". In: ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 820–830.

Liu, Keli, Jelena Markovic, and Robert Tibshirani (2018). *More powerful post-selection inference, with application to the Lasso.* arXiv: `1801.09037 [stat.ME]`.

Loftus, Joshua R. (2015). *Selective inference after cross-validation.* arXiv: `1511.08866 [stat.ME]`.

Markovic, Jelena, Lucy Xia, and Jonathan Taylor (2017). *Unifying approach to selective inference with applications to cross-validation.* arXiv: `1703.06559 [stat.ME]`.

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu (2012). "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Stat. Sci.* 27.4, pp. 538–557.

Park, Trevor and George Casella (2008). "The Bayesian Lasso." In: *J. Am. Stat. Assoc.* 103.482, pp. 681–686.

Poignard, Benjamin and Makoto Yamada (2020). "Sparse Hilbert-Schmidt Independence Criterion Regression". In: ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 538–548.

Pötscher, Benedikt M. (1991). "Effects of Model Selection on Inference". In: *Econometric Theory* 7.2, pp. 163–185.

Schölkopf, Bernhard and Alexander J. Smola (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: MIT Press.

Shimodaira, Hidetoshi (2004). "Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling". In: *Ann. Stat.* 32.6, pp. 2616–2641.

Smola, Alex, Arthur Gretton, Le Song, and Bernhard Schölkopf (2007). "A Hilbert space embedding for distributions." In: *Algorithmic learning theory. 18th international conference, ALT 2007, Sendai, Japan, October 1–4, 2007. Proceedings.* Berlin: Springer, pp. 13–31.

Song, Le, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt (2012). "Feature selection via dependence maximization." In: *J. Mach. Learn. Res.* 13, pp. 1393–1434.

Steinwart, Ingo (2002). "On the influence of the kernel on the consistency of support vector machines." In: *J. Mach. Learn. Res.* 2.1, pp. 67–93.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions. Discussion". In: *J. R. Stat. Soc., Ser. B* 36, pp. 111–147.

Takahashi, Yuta, Masao Ueki, Makoto Yamada, Gen Tamiya, Ikuko Motoike, Daisuke Saigusa, Miyuki Sakurai, Fuji Nagami, Soichi Ogishima, Seizo Koshiba, Kengo Kinoshita, Masayuki Tamamoto, and Hiroaki Tomita (2020). "Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection". In: *Translational Psychiatry* 10, p. 157.

Taylor, Jonathan and Robert Tibshirani (2018). "Post-selection inference for ℓ_1-penalized likelihood models." In: *Can. J. Stat.* 46.1, pp. 41–61.

Terada, Yoshikazu and Hidetoshi Shimodaira (2017). *Selective inference for the problem of regions via multiscale bootstrap.* arXiv: `1711.00949 [math.ST]`.

— (2019). *Selective inference after feature selection via multiscale bootstrap.* arXiv: `1905.10573 [stat.ME]`.

Tian, Xiaoying, Joshua R. Loftus, and Jonathan E. Taylor (2018). "Selective inference with unknown variance via the square-root lasso." In: *Biometrika* 105.4, pp. 755–768.

Tian, Xiaoying and Jonathan Taylor (2017). "Asymptotics of selective inference." In: *Scand. J. Stat.* 44.2, pp. 480–499.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso." In: *J. R. Stat. Soc., Ser. B* 58.1, pp. 267–288.

Tibshirani, Ryan J., Alessandro Rinaldo, Rob Tibshirani, and Larry Wasserman (2018). "Uniform asymptotic inference and the bootstrap after model selection." In: *Ann. Stat.* 46.3, pp. 1255–1287.

Tibshirani, Ryan J., Jonathan Taylor, Richard Lockhart, and Robert Tibshirani (2016). "Exact Post-Selection Inference for Sequential Regression Procedures". In: *Journal of the American Statistical Association* 111.514, pp. 600–620.

van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." In: *Ann. Stat.* 42.3, pp. 1166–1202.

Yamada, M., J. Tang, J. Lugo-Martinez, E. Hodzic, R. Shrestha, A. Saha, H. Ouyang, D. Yin, H. Mamitsuka, C. Sahinalp, P. Radivojac, F. Menczer, and Y. Chang (2018). "Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data". In: *IEEE Transactions on Knowledge and Data Engineering* 30.7, pp. 1352–1365.

Yamada, Makoto, Wittawat Jitkrittum, Leonid Sigal, Eric P. Xing, and Masashi Sugiyama (2014). "High-dimensional feature selection by feature-wise kernelized Lasso." In: *Neural Comput.* 26.1, pp. 185–207.

Yamada, Makoto, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi (2018). "Post Selection Inference with Kernels". In: ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, pp. 152–160.

Yöntem, Mustafa Kemal, Kemal Adem, Tahsin İlhan, and Serhat Kılıçarslan (2019). "Divorce prediction using correlation based feature selection and artificial neural networks". In: *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi* 9, pp. 259–273.

Zhang, Cun-Hui and Stephanie S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models." In: *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 76.1, pp. 217–242.

Zhang, Qinyi, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic (2018). "Large-scale kernel methods for independence testing." In: *Stat. Comput.* 28.1, pp. 113–130.

Zou, Hui (2006). "The adaptive lasso and its oracle properties." In: *J. Am. Stat. Assoc.* 101.476, pp. 1418–1429.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2007). "On the "degrees of freedom" of the lasso". In: *Ann. Stat.* 35.5, pp. 2173–2192.