Stability of Linear Structural Equation Models of Causal Inference*

Karthik Abinav Sankararaman
University of Maryland, College Park
kabinav@cs.umd.edu

Anand Louis
Indian Institute of
Science, Bangalore
anandl@iisc.ac.in

Navin Goyal
Microsoft Research, India
navingo@microsoft.com

Abstract

We consider numerical stability of the parameter recovery problem in Linear Structural Equation Model (LSEM) of causal inference. Numerical stability is essential for the recovered parameters to be reliable. A long line of work starting from Wright (1920) has focused on understanding which sub-classes of LSEM allow for efficient parameter recovery. Despite decades of study, this question is not yet fully resolved. The goal of the present paper is complementary to this line of work: we want to understand the stability of the recovery problem in the cases when efficient recovery is possible. Numerical stability of Pearl's notion of causality was first studied in Schulman and Srivastava (2016) using the concept of condition number where they provide ill-conditioned examples. In this work, we provide a condition number analysis for the LSEM. First we prove that under a sufficient condition, for a certain sub-class of LSEM that are *bow-free* (Brito and Pearl (2002)), parameter recovery is numerically stable. We further prove that *randomly* chosen input parameters for this family satisfy the condition with a substantial probability. Hence for this family, on a large subset of parameter space, recovery is stable. Next we construct an example of LSEM on four vertices with *unbounded* condition number. We then corroborate our theoretical findings via simulations as well as real-world experiments for a sociology application. Finally, we provide a general heuristic for estimating the condition number of any LSEM instance.

1 Introduction

Inferring *causality*, *i.e.,* whether a group of events causes another group of events is a central problem in a wide range of fields from natural to social sciences. A common approach to inferring causality is Randomized controlled trials (RCT). Here the experimenter intervenes on a system of variables (often called stimulus variables) such that it is not affected by any confounders with the variables of interest (often called response variables) and observes the probability distributions on the response variables. Unfortunately, in many cases of interest performing RCT is either costly or impossible due to practical or ethical or legal reasons. A common example is the age-old debate [32] on whether smoking causes cancer. In such scenarios RCT is completely out of the question due to ethical issues. This necessitates new inference techniques.

The causal inference problem has been extensively studied in statistics and mathematics (*e.g.,* [18, 20, 23, 24]) where decades of research has led to rich mathematical theories and a framework for conceptualizing and analyzing causal inference. One such line of work is the *Linear Structural Equation Model* (or LSEM in short) for formalizing causal inference (see the monograph [4] for a survey of classical results). In fact, this is among the most commonly used models of causality in social sciences [2, 4] and some natural sciences [31]. In this model, we are given a mixed graph on n (observable) variables[1] of the system containing both directed and bi-directed edges (see Figure 1 for an example). We will assume that the directed edges in the mixed graph form a directed acyclic graph (DAG). A directed edge from vertex u to vertex v represents the presence of causal effect of variable u on variable v, while the bi-directed edges represent the presence of confounding effect (modeled as noise) which we next explain.[2] In the LSEM, the fol-

*Full version of the paper [25]

[1] In this paper we are interested in properties for large n.

[2] We also interchangeably refer to the directed edges as *observable* edges since they denote the direct causal effects and

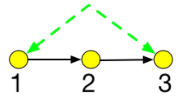

Figure 1: Mixed Graph: Black solid edges represent causal edges. Green dotted edges represent covariance of the noise.

lowing extra assumption is made (see Equation (1)): the value of a variable v is determined by a (weighted) linear combination of its parents' (in the directed graph) values added with a zero-mean *Gaussian* noise term (η_v). The bi-directed graph indicates dependencies between the noise variables (*i.e.,* lack of an edge between variables u and v implies that the covariance between η_u and η_v is 0). We use $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ to represent the matrix of edge weights of the DAG, $\mathbf{\Omega} \in \mathbb{R}^{n \times n}$ to represent the covariance matrix of the Gaussian noise and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ to represent the covariance matrix of the observation data (henceforth called *data covariance matrix*). Let $\mathbf{X} \in \mathbb{R}^{n \times 1}$ denote a vector of random variables corresponding to the observable variables in the system. Let $\eta \in \mathbb{R}^{n \times 1}$ denote the vector of corresponding noises whose covariance matrix is given by $\mathbf{\Omega}$. Formally, the LSEM assumes the following relationship between the random variables in $\mathbf{X}$:

$$\mathbf{X} = \mathbf{\Lambda}^T \mathbf{X} + \eta. \tag{1}$$

From the *Gaussian* assumption on the noise random variable η, it follows that $\mathbf{X}$ is also a multi-variate Gaussian with covariance matrix given by

$$\mathbf{\Sigma} = (\mathbf{I} - \mathbf{\Lambda})^{-T} \, \mathbf{\Omega} \, (\mathbf{I} - \mathbf{\Lambda})^{-1} \ . \tag{2}$$

In a typical setting, the experimenter estimates the joint-distribution by estimating the covariance matrix $\mathbf{\Sigma}$ over the observable variables obtained from finitely many samples. The experimenter also has a causal hypothesis (represented as a mixed graph for the causal effects and the covariance among the noise variables, which in turn determines which entries of $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are required to be 0, referred to as the *zero-patterns* of $\mathbf{\Lambda}$ and $\mathbf{\Omega}$). One then wants to solve the inverse problem of recovering $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ given $\mathbf{\Sigma}$. This problem is solvable for some special types of mixed graphs using parameter recovery algorithms, such as the one in [12, 16].

Thus a central question in the study of LSEM is for which mixed graphs (specified by their zero patterns) and which values of the parameters ($\mathbf{\Lambda}$ and $\mathbf{\Omega}$) is the inverse problem above solvable; in other words, which parameters are identifiable. Ideally one would like all values of the parameters to be identifiable. However, identifiability is often too strong a property to expect to hold for all parameters and instead we are satisfied with a slightly weaker property, namely *generic identifiability* (GI): here we require that identifiability holds for all parameter values except for a measure 0 set (according to some reasonable measure). (The issue of identifiability is a very general one that arises in solving inverse problems in statistics and many other areas of science.) A series of works (*e.g.,* [6, 14, 13, 16, 22]) have made progress on this question by providing various classes of mixed graphs that do allow generic identifiability. However, the general problem of generic identifiability has not been fully resolved. This problem is important since it is a version of a central question in science: what kind of causal hypotheses can be validated purely from observational data as opposed to the situations where one can do RCT. Much of the prior work has focused on designing algorithms with the assumption that the *exact* joint distribution over the variables is available. However, in practice, the data is noisy and inaccurate and the joint distribution is generated via *finite* samples from this noisy data.

While theoretical advances assuming exact joint distribution have been useful it is imperative to understand the effect of violation of these assumptions rigorously. Algorithms for identification in LSEMs are *numerical* algorithms that solve the inverse problem of recovering the underlying parameters constructed from noisy and limited data (a common heuristic is the RICF algorithm [12]). Such models and associated algorithms are useful only if they solve the inverse problem in a *stable* fashion: if the data is perturbed by a small amount then the recovered parameters change only a small amount. If the recovery is unstable then the recovered parameters are unlikely to tell us much about the underlying causal model as they are inordinately affected by the noise. We say that *robust identifiability* (RI) holds for parameter values $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ if even after perturbing the corresponding $\mathbf{\Sigma}$ to $\mathbf{\Sigma}'$ (by a small noise), the recovered parameter values $\mathbf{\Lambda}'$ and $\mathbf{\Omega}'$ are close to the original ones. It follows from the preceding discussion that to consider an inverse problem solved it is not sufficient for generic identifiability to hold; instead, we would like the stronger property of robust identifiability to hold for almost all parameter values (we call this generic robust identifiability; for now we leave the notions of "almost everywhere" and "close" informal). In addition, the problem should also be solvable by an efficient algorithm. The mixed graphs we consider all admit efficient parameter recovery algorithms. Note that GI and RI are properties of the problem and not of the algorithm used to solve the problem.

In the general context of inverse problems, the difference between GI and RI is quite common and important: *e.g.,* in solving a system of n linear equations in n variables

the bi-directed edges as *unobservable* edges since they indicate the unobserved common causes.

given by an $n \times n$ matrix M. For any reasonable distribution on $n \times n$ matrices, the set of singular matrices has measure 0 (an algebraic set of lower dimension given by $\det(M) = 0$). Hence, M is invertible with probability 1 and GI holds. This however does not imply that generic robust identifiability holds: for that one needs to say that the set of ill-conditioned matrices has small measure. To understand RI for this problem, one needs to resort to analyses of the minimum singular value of M which are non-trivial in comparison to GI (e.g., see [7, Sec 2.4]). In general, proving RI almost everywhere turns out to be harder and remains an open problem in many cases even though GI results are known. One recent example is rank-1 decomposition of tensors (for $n \times n \times n$ random tensors of rank up to $n^2/16$, GI is known, whereas generic robust identifiability is known only up to rank $n^{1.5}$); see, *e.g.,* [3]. The problem of tensor decomposition is just one example of a general recent concern for RI results in the theoretical computer science literature and there are many more examples. Generic robust identifiability remains an open problem for semi-Markovian models for which efficient GI is known, *e.g.,* [23].

In the context of causality, the study of robust identifiability was initiated in [28] where the authors construct a family of examples in the so-called Pearl's notion of causality on semi-Markovian graphs[3] (see *e.g.,* [23]) and show that for this family there exists an *adversarial* perturbation of the input which causes the recovered parameters to be drastically different (under an appropriate metric described later) from the actual set of parameters. However this result has the following limitation. Their adversarial perturbations are carefully crafted and this worst case scenario can be alleviated by modifying just a few edges (in fact just deleting some edges randomly suffices which can also be achieved without changing the zero-pattern by choosing the parameters appropriately). This leads us to ask: *how prevalent are such perturbations?* Since there is no canonical notion of what a typical LSEM model is (i.e. the graph structure and the associated parameters), we will assume that the graph structure is given and the parameters are randomly chosen according to some reasonable distribution. Thus we would like to answer the following question[4].

Question 1.1 (Informal)**.** For the class of LSEMs that are uniquely identifiable, does robust identifiability hold for most choices of parameters?

The question above is informal mainly because "most choices of parameters" is not a formally defined notion. We can quantify this notion by putting a probability measure on the space of all parameters. This is what we will do.

Notation. Throughout this paper, we will use the following notation. Bold fonts represent matrices and vectors.

Given matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$, we define the *relative distance*, denoted by $\mathrm{Rel}(\mathbf{A}, \mathbf{B})$ as the following quantity.

$$\mathrm{Rel}(\mathbf{A}, \mathbf{B}) \overset{\text{def}}{=} \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant m: \left|A_{i,j}\right|\neq 0} \frac{|A_{i,j} - B_{i,j}|}{\left|A_{i,j}\right|}.$$

In this paper we use the notion of condition number (see [7] for a detailed survey on condition numbers in numerical algorithms) to quantitatively measure the effect of perturbations on data in the parameter recovery problem. The specific definition of condition number we use is a natural extension of the ℓ_∞-condition number studied in [28] to matrices.

Definition 1.2 (Relative ℓ_∞-condition number)**.** Let $\mathbf{\Sigma}$ be a given data covariance matrix and $\mathbf{\Lambda}$ be the corresponding parameter matrix. Let a γ-perturbed family of matrices be denoted by $\mathcal{F}_\gamma$ (*i.e.,* set of matrices $\tilde{\mathbf{\Sigma}}_\gamma$ such that $\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}}_\gamma) \leqslant \gamma$). For any $\tilde{\mathbf{\Sigma}}_\gamma \in \mathcal{F}_\gamma$ let the corresponding recovered parameter matrix be denoted by $\tilde{\mathbf{\Lambda}}_\gamma$. Then the relative ℓ_∞-condition number is defined as,

$$\kappa(\mathbf{\Lambda}, \mathbf{\Sigma}) \overset{\text{def}}{=} \lim_{\gamma\to 0^+} \sup_{\tilde{\mathbf{\Sigma}}_\gamma \in \mathcal{F}_\gamma} \frac{\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}}_\gamma)}{\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}}_\gamma)}. \tag{3}$$

We confine our attention to the stability of recovery of $\mathbf{\Lambda}$ as once $\mathbf{\Lambda}$ is recovered approximately, $\mathbf{\Omega} = (\mathbf{I} - \mathbf{\Lambda})^T \mathbf{\Sigma} (\mathbf{I} - \mathbf{\Lambda})$ can be easily approximated in a stable manner. In this paper, we restrict our theoretical analyses to causal models specified by *bow-free* paths which are inspired by bow-free graphs [5][5] In [5] the authors show that the bow-free property of the mixed graph underlying the causal model is a sufficient condition for unique identifiability. We now define bow-free graphs and bow-free paths; Figure 2 has an example of bow-free paths.

Definition 1.3 (Bow-free graphs)**.** Bow-free (mixed) graphs are those where for any pair of vertices, both directed and undirected edges are never present simultaneously.

[3] Unlike LSEM, this is a non-parametric model. The functional dependence of a variable on the parents' variables is allowed to be fully general, in particular it need not be linear. This of course comes at the price of making the inference problem computationally and statistically harder.

[4] The question of understanding the condition number (a measure of stability) of LSEMs was raised in [28], though presumably in the worst-case sense of whether there are identifiable instances for which recovery is unstable. As mentioned in their work, when the model is *not* uniquely identifiable, the authors in [11] show an example where uncertainty in estimating the parameters can be unbounded.

[5] See footnote 4 in [5] for the significance of bow-free graphs in LSEM.

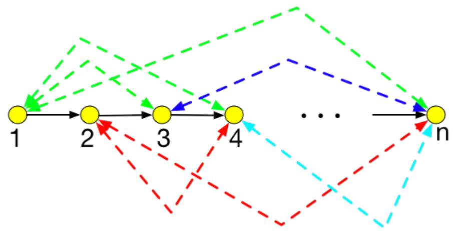

Figure 2: Illustration of a bow-free path. Solid lines represent directed causal edges and dotted lines represent bi-directed covariance edges.

Definition 1.4 (Bow-free Paths)**.** A causal model is called a *bow-free path* if the underlying DAG forms a directed path and the mixed graph is bow-free [5]. Consider n vertices indexed $1, 2, \ldots, n$. The (observable) DAG forms a path starting at 1 with directed edges going from vertex i to vertex $i+1$. The bi-directed edges can exist between pairs of vertices (i, j) only if $|i - j| \geqslant 2$. Thus the only potential non-zero entries in $\mathbf{\Lambda}$ are in the diagonal immediately above the principal diagonal. Similarly, the diagonal immediately above and below the principal diagonal in $\mathbf{\Omega}$ is always zero.

We further assume the following conditions on the parameters of the *bow-free* paths for studying the numerical stability of *any* parameter recovery algorithm. Later in Section 4 we show that a natural random process of generating $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ satisfies these assumptions with high-probability. Informally, the model can be viewed as a *generalization* of *symmetric diagonally dominant* (SDD) matrices.

Model 1.5. Our model satisfies the following conditions on the data-covariance matrix and the perturbation with the underlying graph structure given by bow-free paths. It takes parameters α and λ.

1. **Data properties.** $\mathbf{\Sigma} \succeq 0$ is an $n \times n$ symmetric matrix satisfying the following conditions[6] for some $0 \leqslant \alpha \leqslant \frac{1}{5}$.

   $$\left|\Sigma_{i-1,i}\right|, \left|\Sigma_{i,i+1}\right|, \left|\Sigma_{i-1,i+1}\right| \leqslant \alpha\, \Sigma_{i,i} \qquad \forall i \in [n-1].$$

   Additionally, the *true* parameters $\mathbf{\Lambda}$, corresponding to $\mathbf{\Sigma}$, satisfy the following. For every i, j such that there is a directed edge from i to j, we have $\frac{1}{n^2} \leqslant \frac{1}{\lambda} \leqslant \left|\Lambda_{i,j}\right| \leqslant 1$, where λ is a parameter.

2. **Perturbation properties.** For each $i, j \in [n]$ and a fixed $\gamma \leqslant \frac{1}{n^6}$, let $\varepsilon_{i,j} = \varepsilon_{j,i}$ be arbitrary numbers satisfying $0 \leqslant \left|\varepsilon_{i,j}\right| = \left|\varepsilon_{j,i}\right| \leqslant \gamma \left|\Sigma_{i,j}\right|$ for all i, j such that that there exists a pair $i^*, j^* \in [n]$ with $\left|\varepsilon_{i^*,j^*}\right| = \left|\varepsilon_{j^*,i^*}\right| = \gamma \left|\Sigma_{i^*,j^*}\right|$. Note that we eventually consider γ in the limit going to 0, hence some small but fixed γ suffices. Let $\tilde{\Sigma}_{i,j} \overset{\text{def}}{=} \Sigma_{i,j} + \varepsilon_{i,j}$ for every pair (i, j).

Remark 1.6. The covariance matrix has errors arising from three sources: (1) measurement errors, (2) numerical errors, and (3) error due to finitely many samples. The combined effect is modeled by Item 2 (perturbation properties) above of Model 1.4 which is very general as we only care about the max error.

Remark 1.7. We show in the full version that the constant $\frac{1}{5}$ is an *approximation* to the following. Let $\tau = 1 + 5n\gamma$. Then we want $[1 - (\tau + 2)\,\alpha - (\tau + 1)\,\alpha^2] > 0$ and $\alpha \leqslant \frac{1}{4}$.

With the definitions and model in place, we can now pose Question 1.1 formally:

Question 1.8 (Formal)**.** For the class of LSEMs represented by bow-free paths and Model 1.5, can we characterize the behavior of the ℓ_∞-condition number?

1.1 Our Contributions

Our key contribution in this paper is a step towards understanding Question 1.1 by providing an answer to Question 1.8. In particular, we first prove that when Model assumptions 1.5 hold, the ℓ_∞-condition number of the parameter recovery problem is upper-bounded by a polynomial in n. Formally, we prove Theorem 1.9. This implies that the loss in precision scales *logarithmically* and hence we need at most $O(\log d)$ additional bits to get a precision of d bits[7]. See [7] and [28] for further discussion on the relation between condition number and the number of bits needed for d-bit precision in computation as well as other implications of small condition number.

Theorem 1.9 (Stability Result)**.** *Under the assumptions in Model 1.5, we have the following bound on the condition number for bow-free paths with n vertices.*

$$\kappa(\mathbf{\Lambda}, \mathbf{\Sigma}) \leqslant O(n^2).$$

More specifically, the condition number is upper-bounded by $\kappa(\mathbf{\Lambda}, \mathbf{\Sigma}) \leqslant O(\lambda)$ and for the parameters chosen in this model we have the bound in Theorem 1.9.

Furthermore, in Section 4 we show that a natural generative process to construct matrices $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ satisfies the Model assumptions 1.5. Hence this implies that a *large* class of instances are well-conditioned and hence ill-conditioned models are not prevalent under our generative assumption. Moreover, as described in the model

[6] We note that the important part is that α is bounded by a *constant* independent of n. The precise constant is of secondary importance.

[7] Note we already need $O(\mathsf{poly}(n) \log n)$ bits to represent the graph and the associated matrices.

preliminaries, all data covariance matrices that are SDD and have the property that every row has at least 8 entries that are not arbitrarily close to 0, satisfy the assumptions in Model 1.5. Thus an important corollary of this theorem is that when the data covariance matrix is in this class of SDD matrices, it is well-conditioned.

Next we show that there exist examples for LSEMs with *arbitrarily* high condition number. This implies that on such examples it is unreasonable to expect any form of accurate computation. Formally, we prove Theorem 1.10. It shows that the recovery problem itself has a bad condition number and does not depend on any particular algorithm that is used for parameter recovery. This theorem follows easily using the techniques of [16] and we include it for completeness.

Theorem 1.10 (Instability Result)**.** *There exists a bow-free path of length four and data covariance matrix* $\mathbf{\Sigma}$ *such that the parameter recovery problem of obtaining* $\mathbf{\Lambda}$ *from* $\mathbf{\Sigma}$ *has an arbitrarily large condition number.*

We perform numerical simulations to corroborate the theoretical results in this paper. We verify our theorems using numerical simulations. Furthermore, we consider general graphs (*e.g.,* clique of paths, layered graphs) and show that a similar behavior on the condition number holds. Finally, we consider a real-world dataset used in [29] and perform condition number analysis experimentally.

In the full-version of the paper, we additionally give a general heuristic for practitioners to determine the condition number of any given instance, which may be of independent interest. This heuristic can detect *bad* instances with high-probability.

1.2 Related work

A standard reference on Structural Causal Models is Bollen [4]. Identifiability and robust identifiability of LSEMs has been studied from various viewpoints in the literature: Robustness to model misspecification, to measurement errors, *etc.* (*e.g.,* Chapter 5 of Bollen [4] and [19, 26, 21]). These works are not directly related to this paper, since they focus on the identifiability problem under erroneous model specification and/or measurement. See the sub-section on measurement errors below for further details.

Identifiability. The problem of (generic) identification in LSEMs is well-studied though still not fully understood. We give a brief overview of the known results. All works in this section assume that the data-covariance matrix is exact and do not consider robustness aspects of the problem. These works do not directly relate to the problem studied in this paper. This problem has a rich history (see [23] for some classical results) and we only give an overview of recent results. Brito and Pearl [6] gave a sufficient condition called G-criterion which implies linear independence on a certain set of equations. The variables involved in this set is called the auxiliary variables. The main theorem in their paper is that if for every variable there is a corresponding "auxiliary" variable, then the model is identifiable. Followed by this, Foygel *et al.* [16] introduced the notion of half-trek criterion which gives a graphical criterion for identifiability in LSEM. This criterion strictly subsumes the G-criterion framework. Both [6, 16] supply efficient algorithms for identification. Chen *et al.* [10] tackle the problem of identifying "overidentifying constraints", which leads to identifiability on a class of graphs strictly larger than those in [16] and [6]. Ernest *et al.* [15] consider a variant of the problem called "Partially Linear Additive SEM" (PLSEM) with gaussian noise. In this variant the value for a random variable X is determined both by an additive linear factor of some of its parents (called linear parents) and additive factor of the other parents (called non-linear parents) via a non-linear function. They give characterization for identifiability in terms of graphical conditions for this more general model. Chen [8] extends the half-trek criterion of [16] and give a c-component decomposition (Tian [33]) based algorithm to identify a broader class of models. Drton and Weihs [14] also extend the half-trek criterion ([16]) to include the ancestor decomposition technique of Tian [33]. Chen *et al.* [9] approach the parameter identification problem for LSEM via the notion of *auxiliary variables*. This method subsumes all the previous works above and identifies a strictly larger set of models.

Measurement errors. Another recent line of work [27, 34] studies the problem of identification under *measurement errors*. Both our work and these works share the same motivation: causal inference in real-world is usually performed on noisy data and hence it is important to understand how noise affects the causal identification. However their focus differs significantly from the question we tackle in this paper. They pose and answer the problem of causal identification when the variables used are not identical to the ones that were intended to be measured. This leads to different conditional independences and hence a different causal graph; they study the identification problem in this setting and characterize when such identification is possible. Recently [17] looked at sample complexity of identification in LSEM. They consider the special case when $\mathbf{\Omega}$ is the Identity matrix and give a new algorithm for identifiability with optimal sample complexity.

In this paper we are interested in the question of robust identifiability, along the lines of aforementioned work of

Schulman and Srivastava [28]. While the work of [28] was direct inspiration for the present paper, since we work with LSEMs and [28] work with the semi-Markovian causal models of Pearl, the techniques involved are completely different. Moreover, as previously remarked, another important difference between [28] and our work is that our main result is positive: the parameter recovery problem is well-conditioned for most choices of the parameters for a well-defined notion of most choices.

2 Bow-free Paths — Stable Instances

In this section, we show that for bow-free paths under Model 1.5, the condition number is small. More precisely, we prove Theorem 1.9.

To prove the main theorem, we set-up some helper lemmas. Using Foygel *et al.* [16], we obtain the following recurrence to compute the values of $\mathbf{\Lambda}$ from the correlation matrix $\mathbf{\Sigma}$. The base case is $\Lambda_{1,2} = \frac{\Sigma_{1,2}}{\Sigma_{1,1}}$. For every $i \geqslant 2$ we have,

$$\Lambda_{i,i+1} = \frac{-\Lambda_{i-1,i}\Sigma_{i-1,i+1} + \Sigma_{i,i+1}}{-\Lambda_{i-1,i}\Sigma_{i-1,i} + \Sigma_{i,i}}\,. \tag{4}$$

We first show that the model assumptions 1.5 imply that the parameters recovered from the perturbed version are not too large.

Lemma 2.1. *For each $i \in [n-1]$, we have $\left|\tilde{\Lambda}_{i,i+1}\right| \leqslant \tau_i \ (:= \tau_{i-1} + 5\gamma) \leqslant \tau\ (:= 1 + 5n\gamma)$.*

Next, we show that the relative distance between the real parameter $\mathbf{\Lambda}$ and the recovered parameter from the perturbed instance $\tilde{\mathbf{\Lambda}}$ is not too large.

Lemma 2.2. *Let $\tau = 1 + 5n\gamma$. Define $\beta_c := \frac{[(3+3\tau)\alpha + (\tau+1)]}{1-(\tau+2)\alpha-(\tau+1)\alpha^2-4n\gamma}$. Then for each $i \in [n-1]$ we have that,*

$$\left|\Lambda_{i,i+1} - \tilde{\Lambda}_{i,i+1}\right| \leqslant \beta_c \cdot \frac{\gamma}{1-\gamma}\,.$$

2.1 Proof of Theorem 1.9

We are now ready to prove the main Theorem 1.9. From Lemma 2.2 and the model assumptions we have $\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}}) = \frac{|\Lambda_{i,i+1}-\tilde{\Lambda}_{i,i+1}|}{|\Lambda_{i,i+1}|} \leqslant \beta_c \cdot \frac{\gamma}{1-\gamma} \cdot \lambda$. From perturbation properties in the Model 1.5, and the definition of i^*, j^*, we have $\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}}) = \frac{\gamma|\Sigma_{i^*,j^*}|}{|\Sigma_{i^*,j^*}|} = \gamma$. Therefore,

$$\frac{\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})}{\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}})} \leqslant \frac{\lambda \cdot \beta_c}{1-\gamma}.$$

This implies that,

$$\lim_{\gamma \to 0^+} \frac{\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})}{\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}})} \leqslant \lambda\beta_c \leqslant O(\lambda) \leqslant O(n^2).$$

The second last inequality used the fact that $\beta_c \leqslant O(1)$ and the last inequality used the fact that $\frac{1}{\lambda} \geqslant \Omega(n^{-2})$.

3 Instability Example

In this section, we show that there exist simple *bow-free path* examples where the condition number can be arbitrarily large. We consider a point on the measure 0 set of unidentifiable parameters and apply a small perturbation (of magnitude ϵ). This instance in the parameter space is identifiable (by using [16]). We then apply a perturbation (of magnitude γ) to this instance and compute the condition number. By making ϵ arbitrarily close to 0, we obtain as large a condition number as desired. Any point on the singular variety could be used for this purpose and the proof of Theorem 1.10 provides a concrete instance. The example we construct is as follows. Consider a path of four vertices. Fix a small value ϵ. Define the matrices $\mathbf{\Omega}$ and $\mathbf{\Lambda}$ as follows.

$$\mathbf{\Omega} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 0 & 1/2 \\ 1/2 & 0 & 1+\epsilon & 0 \\ 1/2 & 1/2 & 0 & 1 \end{bmatrix} \succeq 0$$

and $\Lambda_{1,2} = \sqrt{2}, \Lambda_{2,3} = -\sqrt{2}, \Lambda_{3,4} = \frac{1}{2}$.

Thus, we have the following.

$$(\mathbb{1} - \mathbf{\Lambda})^{-1} = \begin{bmatrix} 1 & \sqrt{2} & -2 & -1 \\ 0 & 1 & -\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Multiplying with $\mathbf{\Omega}$ we get $(\mathbb{1} - \mathbf{\Lambda})^{-T}\mathbf{\Omega}$ is,

$$\begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ \sqrt{2} & 1 & 1/\sqrt{2} & 1/\sqrt{2} + 1/2 \\ -3/2 & -\sqrt{2} & \epsilon & -1 - 1/\sqrt{2} \\ -1/4 & -1/\sqrt{2} + 1/2 & \epsilon/2 & 1/2(1 - 1/\sqrt{2}) \end{bmatrix}$$

Therefore, the resulting matrix $\mathbf{\Sigma}$, obtained by $(\mathbb{1} - \mathbf{\Lambda})^{-T}\mathbf{\Omega}(\mathbb{1} - \mathbf{\Lambda})^{-1}$ is,

$$\begin{bmatrix} 1 & \sqrt{2} & -\frac{3}{2} & -\frac{1}{4} \\ \sqrt{2} & 3 & -\frac{5}{\sqrt{2}} & \frac{1}{2} - \frac{3}{2\sqrt{2}} \\ -\frac{3}{2} & -\frac{5}{\sqrt{2}} & 5+\epsilon & \frac{3}{2} - \frac{1}{\sqrt{2}} + \frac{\epsilon}{2} \\ -\frac{1}{4} & \frac{1}{2} - \frac{3}{2\sqrt{2}} & \frac{3}{2} - \frac{1}{\sqrt{2}} + \frac{\epsilon}{2} & \frac{5}{4} - \frac{1}{\sqrt{2}} + \frac{\epsilon}{4} \end{bmatrix}$$

Perturb every entry of $\mathbf{\Sigma}$ additively by γ to obtain $\tilde{\mathbf{\Sigma}}$. This will ensure that $\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}}) = 4\gamma$ since all entries in $\mathbf{\Sigma}$ are at least $1/4$. Now we show that in the reconstruction of $\tilde{\mathbf{\Lambda}}$ from $\mathbf{\Sigma}_\gamma$, the entry $\tilde{\mathbf{\Lambda}}_{3,4} = 1$. This implies that the condition number $\kappa(\mathbf{\Lambda}, \mathbf{\Sigma}) = O\left(\frac{1}{\gamma}\right)$ and since γ can be made arbitrarily close to 0, this implies that the condition number is unbounded.

The denominator in the expression for $\tilde{\Lambda}_{3,4}$ in (4) is, $-\Lambda_{2,3}\tilde{\Sigma}_{2,3} + \tilde{\Sigma}_{3,3} = \epsilon + \left(1 + \sqrt{2}\right)\gamma$.

Likewise the numerator in the expression for $\tilde{\Lambda}_{3,4}$ evaluates to, $-\Lambda_{2,3}\tilde{\Sigma}_{2,4} + \tilde{\Sigma}_{3,4} = \epsilon/2 + \left(1 + \sqrt{2}\right)\gamma$.

Therefore when $\epsilon \to 0$ we have $\tilde{\Lambda}_{3,4} \to 1$ and hence $\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}}) = O(1) \neq 0$.

4 When do random instances satisfy model assumptions?

In this section, we prove theoretically that randomly generated $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ satisfy the Model 1.5 for a natural generative process, albeit with slightly weaker constants.

Generative model for LSEM **instances.** We consider the following generative model for the LSEM instances. $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ is generated by choosing each non-zero entry to be a sample from the uniform $\mathcal{U}[-h, h]$ distribution. $\mathbf{\Omega} \in \mathbb{R}^{n \times n}$ is chosen by first sampling n-dimensional vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^d$ from a d-dimensional unit sphere, such that $\mathbf{v}_i$ is a uniform sample in the subspace perpendicular to $\mathbf{v}_{i-1}$ for every i (*i.e.,* $\langle \mathbf{v}_i, \mathbf{v}_{i-1} \rangle = 0$) and then letting $\Omega_{i,j} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. Note, in the scenario when $\mathbf{\Omega}$ need not follow a specified *zero-pattern* then first generating vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ independently and uniformly from a d-dimensional unit sphere and then setting $\Omega_{i,j} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ gives an uniform distribution over PSD matrices. Hence, the above generative procedure is a natural extension to randomly generating PSD matrices with a specified zero-patterns.

4.1 Regime when the generative model satisfy the Model 1.5 properties

First, we prove that the random generative process satisfies the bound $\alpha \leqslant \frac{1}{5}$ in Theorem 4.1 below.

Theorem 4.1. *Consider* $\mathbf{\Lambda}$ *and* $\mathbf{\Omega}$ *generated using the above random model with* $0 \leqslant h < \frac{1}{\sqrt{2}}$*. Then there exists a* $0 \leqslant \zeta_h \leqslant h + o(1)$ *and* $\Omega(n^{-2}) \leqslant \chi_h$*, such that for every sufficiently large n, there exists* $0 \leqslant \delta_{1,n} < \frac{1}{2}$ *and* $0 \leqslant \delta_{2,n} < \frac{1}{2}$ *such that,*

$$\mathbb{P}\left[\forall\ i\ \left|\Sigma_{i-1,i}\right|, \left|\Sigma_{i,i+1}\right|, \left|\Sigma_{i-1,i+1}\right| \leqslant \zeta_h \Sigma_{i,i}\right] \geqslant 1 - \delta_{1,n}. \quad (5)$$

$$\mathbb{P}\left[\forall\ i\ \left|\Lambda_{i,i-1}\right| \geqslant \chi_h\right] \geqslant 1 - \delta_{2,n}. \quad (6)$$

Moreover, we have that $\lim_{n \to \infty} \delta_{1,n} = 0$ *and* $\lim_{n \to \infty} \delta_{2,n} = 0$*. Thus, the above statements hold with high-probability.*

Remark 4.2. Note that in this theorem, the constant ζ_h for α is at most h. When $0 \leqslant h < \frac{1}{\sqrt{2}}$, this is at most $\frac{1}{\sqrt{2}}$ where the maximum is achieved when $h = \frac{1}{\sqrt{2}}$. Moreover, when $h < 0.2$, we have $h \leqslant \frac{1}{5}$. Therefore, this satisfies the *exact* requirement in data properties of Model 1.5 when $h < 0.2$ and in the regime $0.2 \leqslant h < \frac{1}{\sqrt{2}}$, it satisfies the property for a slightly larger value of α, namely, $\alpha \leqslant \frac{1}{\sqrt{2}}$. Nonetheless in Section 5 we show experimentally that even in this regime the instance is well-conditioned.

4.2 What happens to the random model when $h > 1$?

We now briefly show that when $h > 1$, the conditions of Model 1.5 *do not* hold. More specifically, we can show that the value of α is larger than 1 with non-negligible probability. Note that the following theorem implies that in this regime we cannot hope to expect $\zeta_h < 1$.

Theorem 4.3. *For every* $h = (1 + z)$*, where* $z > 0$ *is a constant and a large enough n, there exists an* $i \in [n]$ *such that for a constant (dependent on z),* $0 < C_z \leqslant 1$ *we have that,*

$$\mathbb{P}\left[\left|\Sigma_{i,i+1}\right| \geqslant \Sigma_{i,i}\right] \geqslant C_z. \quad (7)$$

5 Experiments

In this section, we will describe our numerical and real-world experimental results. For the purposes of this section, we define the following quantity *randomized condition number* as follows. Consider a given data correlation matrix $\mathbf{\Sigma}$ and the corresponding parameter matrix $\mathbf{\Lambda}$. Let $\mathbf{\Sigma}_\epsilon$ be a matrix obtained by adding $\mathcal{N}(0, \epsilon^2)$ independent random variable to each entry in $\mathbf{\Sigma}$ and let $\tilde{\mathbf{\Lambda}}$ be the corresponding parameter matrix. Then the randomized condition number is $\mathbb{E}\left[\frac{\mathrm{Rel}(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})}{\mathrm{Rel}(\mathbf{\Sigma}, \tilde{\mathbf{\Sigma}}_\epsilon)}\right]$. We use randomized condition number as a proxy for studying the ℓ_∞-condition number.

5.1 Good instances

We start off by showing that on most random instances, the randomized condition number is small (*i.e.,* instances are stable). In particular, it is stable *because* it satisfies the Model 1.5. Thus, a large fraction of random instances satisfies the condition and hence the stability proof directly implies low condition number on these instances.

The first experiment is as follows. We start with an instance of $(\mathbf{\Lambda}, \mathbf{\Omega})$ and the bow-free path. We construct the matrix $\mathbf{\Sigma}$ using the recurrence (4). We then plot the values of $|\Sigma_{i,i}|$, $|\Sigma_{i,i+1}|$, $|\Sigma_{i-1,i}|$, $|\Sigma_{i-1,i+1}|$. $\mathbf{\Lambda}, \mathbf{\Omega}$ are generated exactly as described in the generative model discussed in the Section 4 with $h = 1$. Figure 3 shows the plots for three random runs. Note that in all three cases it satisfies the data properties in Model 1.5.

Next, we explicitly analyze the effect of *random* perturbations on these instances. As predicted by our theory, the instances are fairly stable. We do this as follows. Given a $(\mathbf{\Lambda}, \mathbf{\Omega})$ pair, we can generate $\mathbf{\Sigma}$ in two ways. We can either (1) use the recurrence (4), which can introduce numerical precision errors in $\mathbf{\Sigma}$, or (2) we can generate samples of $\boldsymbol{X}$ (observational data) and estimate $\mathbf{\Sigma}$ by taking the average over samples of $\boldsymbol{X}\boldsymbol{X}^T$, which can introduce sampling errors in $\mathbf{\Sigma}$.

We add, to each non-zero entry of the obtained $\mathbf{\Sigma}$ (*e.g.,* perturbations), an independent $\mathcal{N}(0, \epsilon^2)$ random variable. The value of ϵ we chose for this experiment is 10^{-6}. We then recover $\tilde{\mathbf{\Lambda}}$ from this perturbed $\tilde{\mathbf{\Sigma}}$ using the recurrence (4). Hence, the "noise" entering the system is through two sources, namely, sampling errors and perturbations. Figure 4 shows the effect of both sampling and perturbations.

In the full-version we consider additional experiments. First, we show that both sampling errors and perturbations equally affect the randomized condition number. Moreover, we consider general DAGs and show that similar results holds even for more general graphs.

5.2 Bad Instances

In the next experiment, we start off with the bad example on a bow-free path of length 4. We show that as confirmed in the theory, this instance is highly unstable. We then perturb this instance slightly and show that in a small enough ball around this instance, it continues to remain unstable. Finally, we plot a graph showing the variation of the randomized condition number as a function of the radius of the ball around this instance (see next paragraph for precise definition).

In particular, we start with the bad $\mathbf{\Lambda}$ and $\mathbf{\Omega}$. We then consider a region around these matrices as follows. To $\mathbf{\Lambda}$ (and likewise to $\mathbf{\Omega}$) add an independent $\mathcal{N}(0, \epsilon^2)$ random variable to each non-zero entry. We then consider the effect of random perturbations starting from this new $(\tilde{\mathbf{\Lambda}}, \mathbf{\Omega})$ pair. Figure 5 shows the effect of random perturbations in the region around the matrix $\mathbf{\Lambda}$ and Figure 6 shows the same in the region around $\mathbf{\Omega}$. As evident from the figures, the randomized condition number continues to remain large even when slightly perturbed. Hence, this implies that there are infinitely many pairs $(\mathbf{\Lambda}, \mathbf{\Omega})$ which produce very large condition numbers.

5.3 Real-world data

In this experiment, we analyze a sociology dataset on 6 vertices, taken from a public Internet repository [1]. This was used in [29] which almost resembles our model of LSEM with the exception that we consider Gaussian noise while they don't. Nonetheless, we experiment with the matrices returned by their algorithm, and compute the randomized condition number.

The dataset and pre-processing is the same as [29][8]. Note that the causal graph given by domain experts is *bow-free* therefore, from the theorem in [5], this graph is identifiable. We constructed the $\mathbf{\Sigma}$ matrix from the observational data. We then use the algorithm of [16] to recover the parameter $\mathbf{\Lambda}$. We then *perturb* the data as follows. To each entry in the observational data, we add an additive $\mathcal{N}(0, \epsilon^2)$ noise independently. Compared to the magnitudes of the actual data, this additive noise is insignificant. We recompute $\tilde{\mathbf{\Sigma}}$, from this perturbed observational data. Using the algorithm of [16], we once again recover the parameter $\tilde{\mathbf{\Lambda}}$. For a given ϵ, we run 100 independent runs and take the average. Figure 7 shows the variation of the randomized condition number as a function of ϵ. As the value of ϵ becomes very small, the randomized condition number remains almost constant and approaches 10^{-1}. This number is very close to our *well-behaved* instances implying that this dataset is fairly robust when modeled as a LSEM.

6 Conclusion and Future Directions

In this paper, we initiate the study of condition number analysis for LSEMs. We build theory and experiments to analyze the condition number for a class of instances. Further we give a heuristic that can be used in practice to identify if a given instance is *well-conditioned*. This work opens a number of future directions. An immediate direction is to extend our understanding to other instances. In particular, can we prove condition number for *every* bow-free graph? More generally, a series of recent works has extended the identifiability criteria beyond *bow-free* graphs. Can we analyze the condition number of these larger class of instances? The other direction is to compute the condition number algorithmically. This can be approached by either trying to prove the correctness of the heuristic proposed in this paper, or coming up with a new algorithm. What happens for causal models other than LSEMs? In particular, are the ill-conditioned instances rare in a well-defined sense for Pearl's semi-Markovian causal model?

[8] Obtained via a private communication with the authors of [29].

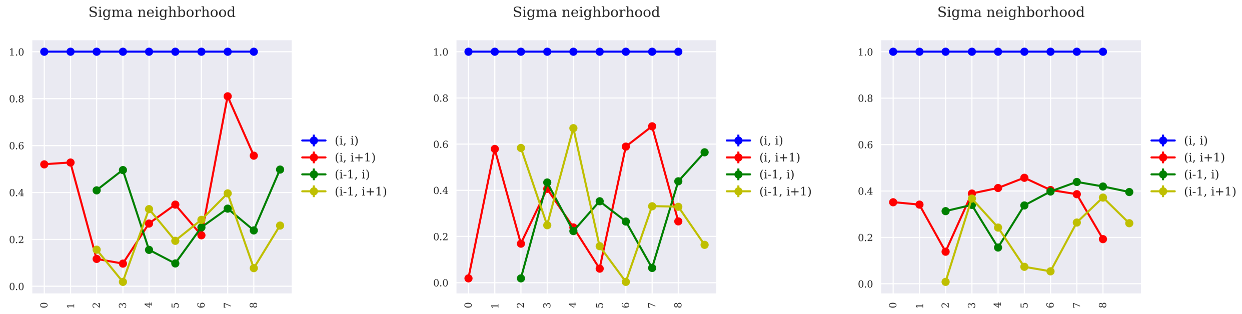

Figure 3: Local values of Σ for three random runs. **x-axis**: Value of i (0-indexed). **y-axis**: Value of $\mathbf{\Sigma}$. Thus, a point on the red-line with x-axis label 2 represents the value $\Sigma_{2,3}$.

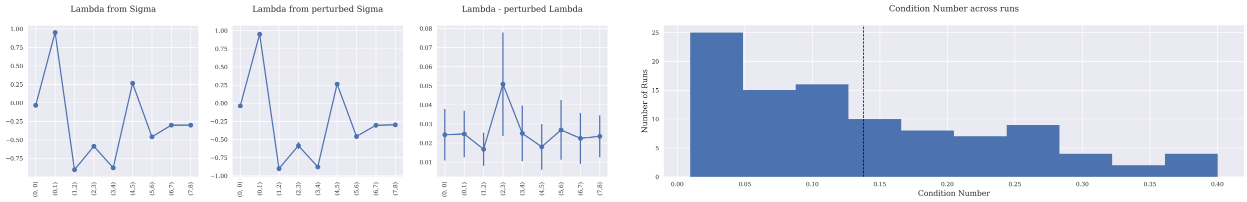

Figure 4: Both sampling and perturbation errors. (First) Denotes the actual values of $\mathbf{\Lambda}$. (Second) Denotes the values of $\mathbf{\Lambda}$ obtained from perturbing $\mathbf{\Sigma}$. (Third) Plots the difference between the two value of $\mathbf{\Lambda}$. (Fourth) Gives the histogram of the randomized condition number in various runs. (First three plots): **x-axis**: index of $\Lambda_{i,i+1}$ (0-indexed). **y-axis**: Value of $\mathbf{\Lambda}$.

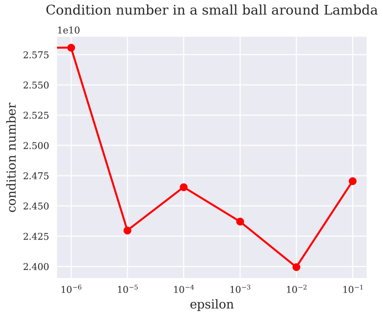

Figure 5: Randomized condition number in a small region around the bad $\mathbf{\Lambda}$. Scale of y-**axis**: 10^{10}.

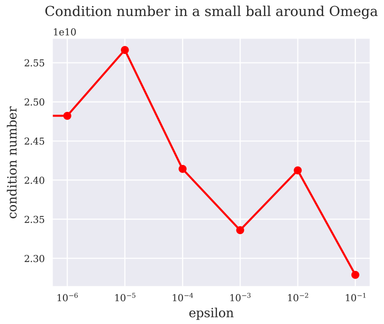

Figure 6: Randomized condition number in a small region around the bad $\mathbf{\Omega}$. Scale of y-**axis**: 10^{10}.

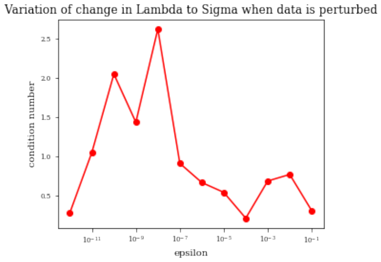

Figure 7: Variation of the randomized condition number as a function of the perturbation error on the sociology dataset

Finally, our results in this paper can be summarized as suggesting that while arbitrarily ill-conditioned instances exist they are rare and unlikely to arise in practice. Moreover, one may check how well-conditioned the instance at hand is. However, this statement depends on the underlying probability distribution on the instances. While our probability distribution is natural, it is desirable to achieve a more "robust" result. Smoothed analysis [30] seems to be well-suited for this purpose. In particular, one could begin with smoothed analysis of ill-conditioned instances identified in the present paper.

Acknowledgements

The authors thank Amit Sharma, Ilya Shpitser and Piyush Srivastava for useful discussions on causality and Shohei Shimizu for providing us with the sociology dataset.

Part of this work was done when KAS was visiting Indian Institute of Science, Bangalore. KAS was supported in part by NSF Awards CNS 1010789, CCF 1422569, CCF-1749864 and research awards from Adobe, Amazon, and Google. Anand Louis was supported in part by SERB Award ECR/2017/003296. AL is also grateful to Microsoft Research for supporting this collaboration.

References

[1] General social survey http://gss.norc.org/Get-The-Data. 2018.

[2] Peter M Bentler and David G Weeks. Linear structural equations with latent variables. *Psychometrika*, 45(3):289–308, 1980.

[3] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Symposium on Theory of Computing, STOC 2014*, pages 594–603, 2014.

[4] Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley-Interscience, 1989.

[5] Carlos Brito and Judea Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002.

[6] Carlos Brito and Judea Pearl. Graphical Condition for Identification in recursive SEM. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 47–54, 2006.

[7] Peter Bürgisser and Felipe Cucker. *Condition: The geometry of numerical algorithms*, volume 349. Springer Science & Business Media, 2013.

[8] Bryant Chen. Identification and Overidentification of Linear Structural Equation Models. *Advances in Neural Information Processing Systems (NIPS)*, pages 1587—-1595, 2016.

[9] Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables. *International Conference on Machine Learning (ICML)*, pages 757—-766, 2017.

[10] Bryant Chen, Jin Tian, and Judea Pearl. Testable Implications of Linear Structural Equation Models. *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 2424–2430, 2014.

[11] N Cornia and JM Mooij. Type-II errors of independence tests can lead to arbitrarily large errors in estimated causal effects: An illustrative example. In *CEUR Workshop Proceedings*, volume 1274, pages 35–42, 2014.

[12] Mathias Drton, Michael Eichler, and Thomas S Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(Oct):2329–2348, 2009.

[13] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, pages 865–886, 2011.

[14] Mathias Drton and Luca Weihs. Generic Identifiability of Linear Structural Equation Models by Ancestor Decomposition. *Scandinavian Journal of Statistics*, 43(4):1035–1045, 2016.

[15] Jan Ernest, Dominik Rothenhausler, and Peter Buhlmann. Causal inference in partially linear structural equation models: identifiability and estimation. *arXiv preprint arXiv:1607.05980*, 2016.

[16] Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Annals of Statistics*, 40(3):1682–1713, 2012.

[17] Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475, 2018.

[18] Isabelle Guyon, Dominik Janzing, and Bernhard Scholkopf. Causality: Objectives and assessment. In *Causality: Objectives and Assessment*, pages 1–42, 2010.

[19] Gregory R. Hancock. Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, 2 (2): 293, 2003.

[20] Paul W Holland, Clark Glymour, and Clive Granger. Statistics and causal inference. *ETS Research Report Series*, 1985(2), 1985.

[21] Xianzheng Huang, Leonard A. Stefanski, and Marie Davidian. Latent-model robustness in structural measurement error models. *Biometrika*, 93(1):53–64, 2006.

[22] Roderick P McDonald. What can we learn from the path equations?: Identifiability, constraints, equivalence. *Psychometrika*, 67(2):225–249, 2002.

[23] Judea Pearl. Causality. *Cambridge university press*, 2009.

[24] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017.

[25] Karthik Abinav Sankararaman, Anand Louis, and Navin Goyal. Stability of linear structural equation models of causal inference. *CoRR*, abs/1905.06836, 2019.

[26] Albert Satorra. Robustness issues in structural equation modeling: A review of recent developments. 24:367–386, 01 1990.

[27] Richard Scheines and Joseph Ramsey. Measurement error and causal discovery. In *CEUR workshop proceedings*, volume 1792, page 1. NIH Public Access, 2016.

[28] Leonard J Schulman and Piyush Srivastava. Stability of Causal Inference. *Uncertainty in Artificial Intelligence (UAI)*, 2016.

[29] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

[30] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[31] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010.

[32] Luther Terry. Smoking and health. *The Reports of the Surgeon General*, 1964.

[33] Jin Tian. Tian's PhD thesis, 2012.

[34] Kun Zhang, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*, 2017.