An Interventional Perspective on Identifiability in Gaussian LTI Systems with Independent Component Analysis

Goutham Rajendran*
Machine Learning Dept., Carnegie Mellon University, Pittsburgh, USA

Patrik Reizinger*
Max Planck Institute for Intelligent Systems, Tübingen, Germany
International Max Planck Research School for Intelligent Systems (IMPRS-IS)
European Laboratory for Learning and Intelligent Systems (ELLIS)

Wieland Brendel
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Pradeep Ravikumar
Machine Learning Dept., Carnegie Mellon University, Pittsburgh, USA

Editors: Francesco Locatello and Vanessa Didelez

Abstract

We investigate the relationship between system identification and intervention design in dynamical systems. While previous research demonstrated how identifiable representation learning methods, such as Independent Component Analysis (ICA), can reveal cause-effect relationships, it relied on a passive perspective without considering how to collect data. Our work shows that in Gaussian Linear Time-Invariant (LTI) systems, the system parameters can be identified by introducing diverse intervention signals in a multi-environment setting. By harnessing appropriate diversity assumptions motivated by the ICA literature, our findings connect experiment design and representational identifiability in dynamical systems. We corroborate our findings on synthetic and (simulated) physical data. Additionally, we show that Hidden Markov Models, in general, and (Gaussian) LTI systems, in particular, fulfil a generalization of the Causal de Finetti theorem with continuous parameters. The project's repository is at `github.com/rpatrik96/lti-ica`.

Keywords: Independent Component Analysis, identifiability, interventions, experiment design, LTI, dynamical systems

1. Introduction

Dynamical systems model temporal phenomena and are prevalent in physics and engineering. They are often Linear Time-Invariant (LTI), e.g., electronic circuits consisting of resistors, capacitors, and inductors, or even some hydraulic and electromechanical systems (Borutzky, 2011). Due to their practical relevance, control theory focuses on LTI systems to understand and control them.

LTI system identification (Åström and Eykhoff, 1971; Ljung, 1998; Pintelon and Schoukens, 2004), i.e., learning the model parameters, has been intensely studied since the 1960s, starting with Rudolf Kálmán's seminal work (Kalman, 1960b) —nonetheless, this field is still quite active, e.g. the ICML 2022 outstanding paper award went to a recent theoretical work on learning mixtures of linear dynamical systems (Chen and Poor (2022)). Dynamical system identification has two main approaches: 1) in the *temporal* domain regression is used to minimize the Mean Squared

* Equal Contribution

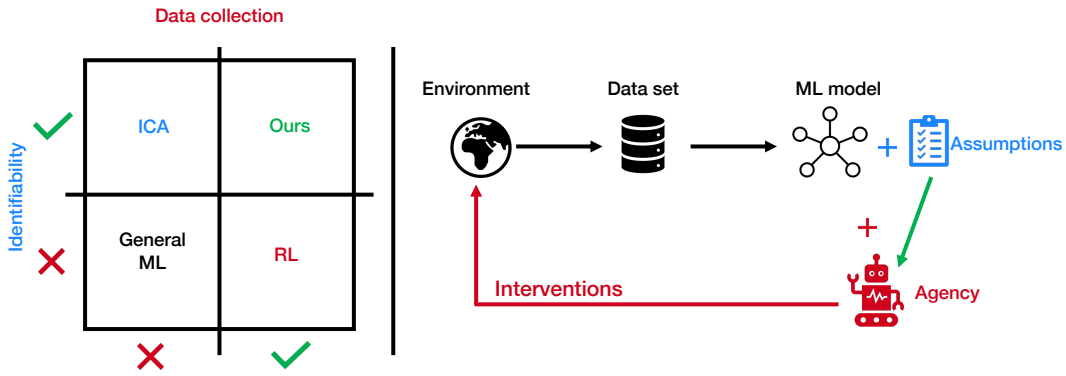


Figure 1: **Left:** Machine Learning (ML) methods categorized based on active data collection (interventions) and identifiability. **Right:** components of the training pipeline for each method on the left. General ML methods use pre-collected data to learn a representation (black components only); Reinforcement Learning (RL) additionally leverages interventions via agency (i.e., interactions with the world; black+red); Independent Component Analysis (ICA) uses pre-collected data with underlying assumptions to achieve identifiability (blue+black); whereas our method uses assumptions about the system to design interventions, i.e., actively collecting data to achieve identifiability (red+blue+black+green)

Error (MSE), which yields Maximum Likelihood Estimation (MLE) for Gaussian random variables (RVs) (Ljung, 1998); 2) in the *frequency* domain the (discrete) Fourier transform is deployed (Ljung, 1998; Pintelon and Schoukens, 2004).

On the other hand, causality (Spirtes et al., 2000; Pearl, 2009) and Independent Component Analysis (ICA) (Comon, 1994; Hyvärinen and Oja, 2000) developed independently from dynamical systems theory, though all three fields attempt to explain natural phenomena via *identifiable* statistical models. Here, identifiability means that a unique parameter set fits the model, and it can be unambiguously recovered and used for downstream tasks such as explanation, planning, or generalization. Dynamical systems are inextricably linked to causality since the arrow of time prescribes causality. Despite their similarities, these fields use different perspectives: in control theory, interventions and control signals (e.g., by applying force to contract a spring) provide an active perspective: i.e., system identifiability results from interactions. On the other hand, ICA studies *passive* identifiability from pre-collected data by imposing distributional and/or functional assumptions.

Our work connects these perspectives: we provide an active data collection strategy—relying on sufficiently varying environments—with identifiability guarantees in Gaussian LTI systems. Our results suggest that equipping ICA with active data collection can yield interventional identifiability in Causal Representation Learning (CRL), as illustrated in Fig. 1[1]. Our learning method maximizes the control signals' log-likelihood, by only assuming knowledge of the control signal distribution (a zero-mean factorized Gaussian with known diagonal covariance) but not the control–observation

1. Grouping all other ML methods into one category is obviously a simplification; we do this to stress that, in general, for most practical problems, a data set is given, and the world (the data generating process) is not explicitly modeled; though data-specific inductive biases (e.g., using Convolutional Neural Networks (CNNs) for images) are used

pairs (also called trajectories), which relaxes the assumptions of conventional regression-based algorithms. We further emphasize the causality connection by showing that the recently proposed Causal de Finetti (CdF) theorem (Guo et al. (2022)) is satisfied in Hidden Markov Models (HMMs) (as a superset of LTI systems), and provide an example for Gaussian LTI systems. This work relies on (Reizinger et al., 2023; Guo et al., 2022) and recent developments in the ICA and CRL literatures (Hyvarinen and Morioka, 2016; Schölkopf et al., 2021). Our contributions are:

- We prove formally and demonstrate that diverse control signals across multiple environments suffice for identifying a Gaussian LTI system. Therefore, active, diverse data collection can enable system identification, giving a strategy for practitioners for data collection.
- We propose an estimation method based on log-likelihood maximization for system identification in the multi-environment setting.
- We show that Hidden Markov Model in general, and (Gaussian) LTI systems in particular, fulfil a generalization of the Causal de Finetti theorem with continuous parameters.

2. Background

Linear Time-Invariant Systems We focus on learning the system parameters of *discrete* LTI systems, which are first-order auto-regressive dynamical systems modeling temporal data.

Definition 1 (Discrete LTI System) *For time step t with a hidden state $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_\mathbf{x}}$, an observed state $\mathbf{y}_t \in \mathcal{Y} \subseteq \mathbb{R}^{d_\mathbf{y}}$, and a (hidden) control signal $\mathbf{u}_t \in \mathcal{U} \subseteq \mathbb{R}^{d_\mathbf{u}}$ with system parameters $\mathbf{A} \in \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$, $\mathbf{B} \in \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{u}}$ and $\mathbf{C} \in \mathbb{R}^{d_\mathbf{y} \times d_\mathbf{x}}$, a discrete LTI system's dynamics is given by*

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\varepsilon}^{\mathrm{x}} \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \boldsymbol{\varepsilon}^{\mathrm{y}}, \end{aligned} \tag{1}$$

where $\boldsymbol{\varepsilon}^{\mathrm{x}}, \boldsymbol{\varepsilon}^{\mathrm{y}}$ are independent noise variables referred to as the process and observation noise, modeling epistemic ($\boldsymbol{\varepsilon}^{\mathrm{x}}$) and aleatoric ($\boldsymbol{\varepsilon}^{\mathrm{y}}$) uncertainty. $\mathbf{A}$ is the state transition, $\mathbf{B}$ the control, and $\mathbf{C}$ the observation matrix. We make standard assumptions on the LTI system as follows:

Assumption 2 (LTI system properties) *We assume that the LTI system of Defn. 1 satisfies:*

(i) *The system is controllable and observable; i.e., the controllability matrix $\mathbf{M}_c$ (Defn. 13) and the observability matrix $\mathbf{M}_o$ (Defn. 15) are full rank;*

(ii) *the control signal $\mathbf{u}$ has a zero-mean factorized Gaussian distribution, $\boldsymbol{\varepsilon}^{\mathrm{x}}, \boldsymbol{\varepsilon}^{\mathrm{y}}$ are Gaussian and all three are independent.*

(iii) *The system is stable, i.e., $\mathbf{A}$ has eigenvalues with magnitude less than 1.*

Observability and *controllability* ensure that the entire state space can be observed and controlled, i.e., we can collect information about the whole of $\mathcal{X}$. *Gaussianity* is a common distributional assumption, and *system stability* is necessary to prevent the system from exploding—i.e., a finite control signal induces a finite system output. Next, we define the *transfer function* of an LTI system, which characterizes the system in frequency domain and is widely used in engineering—it also elucidates the sufficient equivalence class of system parameters we need to identify (cf. Lem. 20).

Definition 3 (Transfer function) *The transfer function $\mathbf{H}(z)$ of a noiseless LTI system relates the control signal and (scalar) output components in the discrete frequency domain (z is the discrete complex frequency variable):*

$$\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{2}$$

The transfer function is the z-transform of the impulse response, which is a theoretical construct describing the system output for a Dirac-delta excitation (Ljung, 1998). Practitioners often use the transfer function for analysis and design, therefore identifiability guarantees for transfer functions are highly desirable. Learning the system parameters from observed data is traditionally estimated via the Markov parameter matrix given by:

Definition 4 (Markov parameter matrix) *For an LTI system and horizon $T \geq 0$, the Markov parameter matrix is*

$$\mathbf{G} = [\mathbf{I}, \mathbf{CB}, \mathbf{CAB}, \dots, \mathbf{CA}^{T-1}\mathbf{B}] \tag{3}$$

Once the Markov parameter matrix is estimated, the Ho-Kálmán algorithm (Ho and Kálmán, 1966) can be used for system identification. Our approach is similar, though working with multiple environments poses additional complexity.

Structural Equation Models (SEMs). We exploit the inherent connection of dynamical systems to causality (Spirtes et al., 2000; Pearl, 2009) and focus on the linear case (Peters et al., 2017; Rajendran et al., 2021; Squires and Uhler, 2022), where the causal relationships among d observed variables $\mathbf{x} = [\mathrm{x}_1, \dots, \mathrm{x}_d]$ are given as $\mathbf{x} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$, where the matrix $\mathbf{A}$ encodes a Directed Acyclic Graph (DAG) via its non-zero entries. This model is closely related to LTI systems but without the temporal component: non-temporal SEMs only model instantaneous effects, e.g., when the discrete time steps are longer than the propagation time of a change within a system; though some extensions consider both instantaneous and temporal effects (Hyvarinen et al., 2010; Lippe et al., 2022b).

Independent Component Analysis (ICA). ICA (Comon, 1994; Hyvarinen et al., 2001) models observed variables via a deterministic function $\boldsymbol{f}$ and independent source (latent) variables $\boldsymbol{\varepsilon}$, i.e. $\mathbf{x} = \boldsymbol{f}(\boldsymbol{\varepsilon})$. ICA studies identifiable models where $\boldsymbol{\varepsilon}$ can be recovered up to indeterminacies, e.g., scaling, permutation, and coordinate transformations. Recent work has generalized this to latent variable models with potentially dependent sources (Kivva et al., 2021; Hyvärinen et al., 2023) and CRL (Schölkopf et al., 2021). However, the connection to LTI systems has not been fully realized.

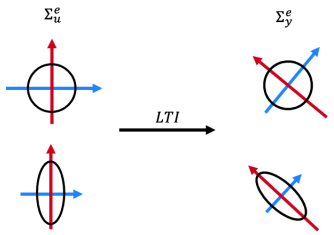
Identifiability. Identifiability postulates the uniqueness of system parameters that fit the data, e.g., $\boldsymbol{f}, \boldsymbol{\varepsilon}$ in ICA and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ in LTI systems. If multiple parameter sets generate the same observed data, then it is impossible to uniquely learn the ground-truth parameters. Since non-identifiability causes problems during learning (D'Amour et al., 2022; Wang et al., 2021), identifiability is crucial for provable system identification. For LTI systems practice, perfect identifiability is impossible since we do not directly observe the raw control signals. Even having access to the true Markov parameters could only guarantee system parameter identifiability up to similarity transformations (Oymak and Ozay, 2019). However, this is sufficient for LTI systems since the transfer function is invariant to similarity transformations (cf. Lem. 20).

3. Main Results

We prove that Gaussian LTI systems can be identified by actively designing control signals to form a sufficiently diverse set of environments (cf. § 3.2 for details). This is inspired by previous works on multi-environmental identifiability in causality and ICA, where data from multiple environments is passively observed (Hyvarinen and Morioka, 2016; Gresele et al., 2019) and then used for learning the underlying parameters. However, in several physical systems, we can apply agency (control) to design experiments.

3.1. Intuition

Our main result provides a *sufficient* condition for identifying Gaussian LTI systems from multiple environments. and also suggests how to design the experiments (data collection) to yield identifiability. Our claim hinges on a sufficient variability condition. The technical details are in § 3.2, whereas our main theorem in § 3.3. Now, we provide an intuition. For this, let us assume that the state $\mathbf{x}_t$, the control signal $\mathbf{u}_t$, and the observed signal $\mathbf{y}_t$ are two-dimensional. We know that for Gaussian $\mathbf{x}_t$, $\mathbf{u}_t$ (and noise variables), $\mathbf{y}_t$ will also be Gaussian, which can be expressed in closed form.



The figure on the right shows the relationship (described by the LTI system equations; cf. Appxs. B.1 and B.2) between the covariances of $\mathbf{u}_t$ and $\mathbf{y}_t$ for two environments. For simplicity, assume that applying the system dynamics is an isometry, i.e., it will only rotate the covariance of $\mathbf{u}_t$ ($\mathbf{\Sigma}_{\mathbf{u}}^e$) into the covariance of $\mathbf{y}_t$ ($\mathbf{\Sigma}_{\mathbf{y}}^e$; this assumption is only for the intuition). In this case, if $\mathbf{u}_t$ has an isotropic Gaussian distribution, then there will be a rotation indeterminacy, since any two axes yield independent components for $\mathbf{y}_t$. However, by adding a new experiment, where $\mathbf{u}_t$ is not isotropic, reduces this indeterminacy to permutations.

3.2. Setting

Assume access to $|E|$ environments with index e, where we observe trajectories from the LTI system

$$\mathbf{x}_{t+1}^e = \mathbf{A}\mathbf{x}_t^e + \mathbf{B}\mathbf{u}_t^e \qquad \mathbf{y}_t^e = \mathbf{C}\mathbf{x}_t^e + \boldsymbol{\varepsilon}_t^e, \qquad \text{where } e \in E = \{0; \dots; |E|-1\} \tag{4}$$

For Gaussian control signals $\mathbf{u}_t^e$ (see below), we can absorb the state noise into $\mathbf{u}_t^e$ (we also dropped the superscript $\mathbf{y}$ from the observation noise). We actively select the control signals for each environment in the form

$$\mathbf{u}_t^e \sim \prod_{i=1}^{d_{\mathbf{u}}} \mathcal{N}\left(0; (\boldsymbol{\sigma}_i^e)^2\right) \tag{5}$$

where $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}}; \mathbf{\Sigma}_{\mathbf{u}})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}_{\mathbf{u}}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{u}} = \operatorname{diag}\left(\boldsymbol{\sigma}_1^2, \dots, \boldsymbol{\sigma}_{d_{\mathbf{u}}}^2\right)$—Gaussianity is a standard assumption for LTI systems. W.l.o.g, we assume a zero initial state $\mathbf{x}_0^e = \mathbf{0}$ and zero mean control signals, but our techniques directly extend to non-zero initial states and mean-shifted control signals with almost no modifications (simply by centering the data via the empirical mean, see Lem. 22); therefore, we focus on the zero-mean case. For identifiability, we need to observe a sufficiently diverse set of environments, quantified via:

Definition 5 (Environment variability matrix) *For an arbitrary base environment (we use $0 \in E$), we define the environment variability matrix $\mathbf{\Delta} \in \mathbb{R}^{|E| \times d_{\mathbf{u}}}$ as*

$$\forall e \in E, i \leq d_{\mathbf{u}} : \mathbf{\Delta}_{e,i} = \frac{1}{(\boldsymbol{\sigma}_i^e)^2} - \frac{1}{(\boldsymbol{\sigma}_i^0)^2}. \tag{6}$$

To achieve sufficient variability, we require that $|E| > d_{\mathbf{u}}$ and $\mathbf{\Delta}$ has full column rank.

Assumption 6 (Environment Variability) $\mathbf{\Delta}$ *has column rank $d_{\mathbf{u}}$.*

Intuitively, this assumption captures that the control signals should be "different" across environments. We only design and observe the variances $(\boldsymbol{\sigma}_i^e)^2$, but not the raw control signals $\mathbf{u}_t^e$. If we had access to $\mathbf{u}_t^e$, then correlation computations would suffice to identify the system (Bakshi et al., 2023a; Oymak and Ozay, 2019). Assum. 6 is not a restrictive assumption because, for instance, if the practitioner chooses the variances from reasonable distributions, e.g., uniformly from a nonempty bounded interval, then well-known results from random matrix theory show that this assumption holds with high probability (Rudelson and Vershynin, 2009, 2008)—to see this, we use the well-known fact that such distributions have bounded sub-Gaussian norm (Vershynin, 2018).

3.3. Main identifiability result

We state our identifiability result for observations with a fixed horizon $T > 0$ from $|E|$ environments.

Theorem 7 *[LTI system identifiability with sufficient variability] For LTI systems satisfying Assums. 2 and 6, the Markov parameter matrix* $\mathbf{G}$ *is identifiable up to permutations and diagonal scaling.*

Proof [Sketch] Intuitively, each independent environment controls a distinct rank-1 subspace of the underlying parameters. If the environments capture $d_{\mathbf{u}}$ linearly independent facets, we can probe the entire space of the system parameters and learn them up to similarity transformations.

Formally, we use change of variables to express the observational density as a function of the control signal parameters. Then we compute the log-odds for each environment w.r.t. an arbitrary base environment. This yields an equation system involving the environment variability matrix $\boldsymbol{\Delta}$, with coefficients being quadratic functions of the control signals. We then compute second derivatives to arrive at a linear equation system. Assuming a full-rank environment variability matrix yields the identifiability of the Markov parameter matrix. The proof is deferred to Appx. B.1. ■

Thm. 7 suggests an active data collection scheme for identifying Gaussian LTI systems and gives an active (intervention-based) view of identifiability theory instead of a passive (relying on pre-collected data samples) view: i.e., the control signal $\mathbf{u}$ should be specified such that Assum. 6 holds (e.g., Gaussians with variances sampled from uniform distributions on nonempty bounded intervals). Thm. 7 proves identifiability of the Markov parameter matrix $\mathbf{G}$, from which the system parameters can be recovered.

For the sake of completeness, we state *how* to do this next. After identifying $\mathbf{G}$, standard techniques (Ho and Kálmán, 1966; Oymak and Ozay, 2019) can extract the underlying system parameters, provided the system identification problem is well-conditioned. For the final corollary, we define the Hankel matrix and assume it to be full-rank.

Definition 8 (Hankel matrix) *For integer parameters $T_1, T_2 \geq 0$, define the (T_1, T_2) Hankel matrix* $\mathbf{H}$ *to be the $T_1 \times T_2$ block matrix with the (i, j) block being* $\mathbf{C}\mathbf{A}^{i+j-2}\mathbf{B}$.

Assumption 9 *There exist integers $T_1, T_2 \geq 0$ such that $T_1 + T_2 \leq T$ and the associated (T_1, T_2) Hankel Matrix* $\mathbf{H}$ *has rank $d_{\mathbf{x}}$.*

Corollary 10 *[Identifiability of LTI systems under sufficient variability] For LTI systems satisfying Assums. 2, 6 and 9, the matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *are identifiable up to a similarity transformation and diagonal scaling.*

Proof By Thm. 7, we recover the Markov parameter matrix. Then, Assum. 9 guarantees that the Hankel matrix is full-rank. Thus, we can use standard system identification results (Oymak and Ozay, 2019; Ho and Kálmán, 1966) to recover $\mathbf{A}, \mathbf{B}, \mathbf{C}$ up to a similarity transformation (which includes permutations) and diagonal scaling. ■

Thm. 7 also implies the identifiability of the practically important transfer function $\mathbf{H}(z)$:

Corollary 11 *For LTI systems satisfying Assums. 2, 6 and 9, the transfer function is identifiable up to permutations and diagonal scaling.*

Proof Using Thm. 7, the system parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are identifiable up to a similarity transformation and diagonal scaling. By Lem. 18, the transfer function is invariant to similarity transformations, completing the proof. ■

3.4. Learning method

Our learning method relies on maximizing the multi-environmental data log-likelihood. The following lemma shows that this objective leads to identifiability:

Lemma 12 *[Identifiability via the multi-environmental log-likelihood] Under Assums. 2, 6 and 9, the parameters that maximize the log-likelihood of a Gaussian LTI system relate to the ground truth via a linear transformation; or, equivalently, the corresponding transfer function is equivalent to the ground truth up to permutations and scalings.*

The proof is deferred to Appx. B.2 and builds on Thm. 7. We formulate the log-likelihood of the control signals (cf. (65)) and optimize it. The model parameters are shared between environments; thus, by conditioning on the model parameters, we have a multivariate Gaussian log-likelihood. Assuming $\mathbf{u}_t^e$ has zero mean and that the environment-dependent covariance $\mathbf{\Sigma}_{\mathbf{u}}^e$ is known, the multivariate Gaussian log-likelihood becomes a weighted least squares problem , which emphasizes that ICA-based and regression-based methods are connected. The resulting loss is (up to constants):

$$\mathcal{L} \propto \sum_e \sum_t (\mathbf{u}_t^e)^\top \mathbf{\Sigma}_{\mathbf{u}}^e \mathbf{u}_t^e,$$

where e indexes the environments, t the time steps. Using this formulation, we learn the matrix $\mathbf{T}^{-1}$ (see Appx. B.2) numerically via gradient descent on the negative log-likelihood.

3.5. Causal de Finetti connection

Roweis and Ghahramani (1999) unified ICA, the Kalman filter, and factor analysis for linear Gaussian systems; however, without a discussion on causality—Gaussianity is generally a prohibitive condition for causal discovery in linear systems (Shimizu et al., 2006). In this work, we provide a causal perspective on Gaussian LTI systems in the multienvironmental case; however, we need to elucidate why and how this fits into the literature. Guo et al. (2022) proved a causal version of the de Finetti theorem, showing that for binary and categorical variables, the cause and effect mechanisms are parameterized by independent parameters, statistically formulating the Independent Causal Mechanisms (ICM) principle (Peters et al., 2018). However, this theoretical result is prohibitive in practice due to requiring exponentially many independence tests. Reizinger et al. (2023)

provided the first insight that contrastive nonlinear Independent Component Analysis (NLICA) can be thought of as a practical realization of the CdF theorem (cf. Appx. C for details).

Here, we show that trajectories in HMMs (LTI systems are a special case of HMMs) satisfy the conditions of the CdF theorem and provide an example for Gaussian LTI systems. Thus, we confirm the conjecture of Guo et al. (2022): at least in a special case, the CdF theorem extends to continuous CdF parameters. Gaussian LTI systems are HMMs, which, by definition, satisfy the following conditional independences for any time steps $l < t < k$:

$$\mathbf{x}_{l<t} \perp \mathbf{x}_{k>t} | \mathbf{x}_t; \qquad \mathbf{y}_t \perp \mathbf{x}_{k>t} | \mathbf{x}_t, \tag{7}$$

i.e., the joint density factorizes (Roweis and Ghahramani, 1999, (3.3)) for a fixed e

$$p\left(\mathbf{y}_1^e, \dots, \mathbf{y}_T^e; \mathbf{x}_1^e, \dots, \mathbf{x}_T^e; \theta^e, \psi^e\right) = \prod_{t=1}^{T} p\left(\mathbf{x}_t^e | \boldsymbol{Pa}_t^e; \theta^e\right) p\left(\mathbf{y}_t^e | \mathbf{x}_t^e; \psi^e\right), \tag{8}$$

where in (8) we included the distributional parameters (to make the correspondence to the CdF theorem easier to see); T is the length of each trajectory, and $\boldsymbol{Pa}_1^e = \emptyset; \boldsymbol{Pa}_{t\neq 1}^e = \mathbf{x}_{t-1}^e$ the parents of $\mathbf{x}_t^e$. The likelihood also factorizes over all environments, which are independent. The CdF theorem posits that the multi-environmental joint density is a mixture of i.i.d. RVs, where for each e the density factorizes as in (8)—assuming the exchangeability (Defn. 24) of pairs $(\mathbf{x}_t, \mathbf{y}_t)_{t\in\mathbb{N}}$ and wo conditional indepencies in the underlying DAG, which, it turns out, are satisfied in HMMs. The CdF theorem states that this factorization is possible with *independent* CdF parameters, i.e., $\psi \perp \theta$, with corresponding measures μ, ν $(\boldsymbol{Pa}_1^e = \emptyset; \boldsymbol{Pa}_{t\neq 1}^e = \mathbf{x}_{t-1}^e)$.

$$p\left(\{\mathbf{y}_1^e, \dots, \mathbf{y}_T^e; \mathbf{x}_1^e, \dots, \mathbf{x}_T^e\}_{e=1}^{|E|}\right) = \int \prod_{e=1}^{|E|} \prod_{t=1}^{T} p\left(\mathbf{y}_t^e | \mathbf{x}_t^e; \psi\right) p\left(\mathbf{x}_t^e | \boldsymbol{Pa}_t^e; \theta\right) d\mu(\theta) d\nu(\psi). \tag{9}$$

If there exist unique $\theta = \theta_0$ and $\psi = \psi_0$ CdF parameters with corresponding Dirac measures, (9) and (8) become equivalent. Moreover, HMMs can be defined with continuous ψ, θ; thus, showing that the there is a version of the CdF theorem for continuous-valued parameters, which we demonstrate for Gaussian LTI systems in the next example.

Example 1 (CdF parameters in Gaussian LTI systems) *The Markov factorization of Gaussian LTI systems (cf. (8)) consists of the conditional distributions describing the state dynamics and the observations:*

$$p\left(\mathbf{x}_{t+1} | \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{A}\mathbf{x}_t; \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^{\mathrm{x}}}\right); \qquad p\left(\mathbf{y}_t | \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{C}\mathbf{x}_t; \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}_t}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^{\mathrm{y}}}\right) \tag{10}$$

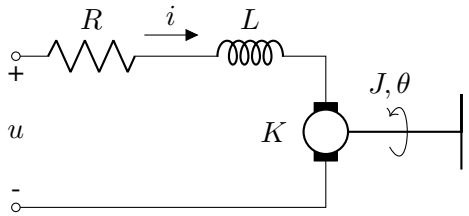
Now we collect the parameters of $p\left(\mathbf{x}_{t+1} | \mathbf{x}_t\right)$ and $p\left(\mathbf{y}_t | \mathbf{x}_t\right)$ into θ and ψ; thus and make their presence explicit by writing $p\left(\mathbf{x}_{t+1} | \mathbf{x}_t; \theta\right)$ and $p\left(\mathbf{y}_t | \mathbf{x}_t; \psi\right)$, where

$$\theta = \left\{\mathbf{A}; \mathbf{B}; \boldsymbol{\Sigma}_{\mathbf{u}}; \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^{\mathrm{x}}}\right\}; \qquad \psi = \left\{\mathbf{C}; \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^{\mathrm{y}}}\right\}. \tag{11}$$

If we know θ, ψ then we can construct the Markov factorization in (8). Then, by defining the corresponding measures to include the indeterminacies (i.e., similarity transformations that stay within the same equivalence class; cf. Lem. 18), then we get (9), showing that Gaussian LTI systems satisfy a CdF theorem over continuous CdF parameters.

4. Experiments

Real-world example (DC motor). We start the experiments section by describing a real-world LTI system and demonstrate that our method can successfully learn the model parameters (we measure this by comparing whether the control signals could be reconstructed; for the exact setting and metrics used, refer to the paragraphs "Setup" and "Metric" below). Assume we have a DC motor, depicted in the figure on the right with the voltage u as the control signal, R and L the armature resistance and conductance, K the electromotive force constant, J and D the rotor inertia and damping coefficient. The states are the armature current i and the rotor angle θ.



The DC motor is a physical system with a *continuous* state-space representation, i.e., an ODE system specifying the time derivative of the states:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} i \\ \theta \end{bmatrix} = \begin{bmatrix} -R/L & K/L \\ -K/J & -D/J \end{bmatrix}\begin{bmatrix} i \\ \theta \end{bmatrix} + \begin{bmatrix} 1/L \\ 0 \end{bmatrix} u; \qquad y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} i \\ \theta \end{bmatrix} \tag{12}$$

For our learning procedure, we need to convert this to a discrete representation in form of (4). This entails two choices[2]: selecting the discretization 1) method and 2) step. The discretization method determines how the system dynamics is modelled between time steps; we choose the widely-used Zero-Order Hold (ZOH) method, which assumes that between time steps, the state value remains the same. The discretization time step is important since it determines the stability of the simulation. That is, even if the system is stable in the sense of Assum. 2, the simulation might diverge if the step size is too big. Notably, this is a property of the numerical ODE solver (Atkinson et al., 2011).

The question is how we should choose the voltage distribution to identify the system parameters. We selected three environments satisfying Assum. 6 (here, the control signal is one-dimensional) and maximized the multi-environmental log-likelihood (65). We used $1\mathrm{e}{-2}$ as the learning rate and time step, 50 epochs with 3 segments with $5,000$ data points each, and a batch size of 8. Measuring the performance with the Mean Correlation Coefficient (MCC) score, we achieved 0.999 on both the training and validation sets, indicating our method's successful application to real-world problems.

Setup. We run additional experiments as follows. We generate data from a controllable and observable Gaussian LTI system, as defined in (4) with observation noise set to zero and a factorized control signal (5) with zero or non-zero-mean—In the latter case, we assume the mean to be known and include that in the log-likelihood (65) to center the reconstructed control signal. We experiment with unknown intervention targets (i.e., $\mathbf{B}$ is full-rank and non-diagonal), and also with observing either $\mathbf{x}_t$ ($\mathbf{C} = \mathbf{I}$) or $\mathbf{y}_t$ ($\mathbf{C} \neq \mathbf{I}$)—the choice of neither $\mathbf{B}$ or $\mathbf{C}$ is used as an inductive bias during training. We also compare performance when, in each environment, only one component of $\mathbf{u}_t$ has a different variance, yielding the minimal condition number (i.e., one) in $\mathbf{\Delta}$—this option is further discussed in Appx. D.1. The discretization time step is set to $3\mathrm{e}{-3}$[3]. We learn the map from $(\mathbf{y}_{t+1}; \mathbf{y}_t) \mapsto \mathbf{u}_t$ as a single matrix (with orthogonally initialized weights) via Stochastic Gradient

2. In practice, we discretize the continuous system with the `cont2discrete` method in `scipy`
3. The discrete time step, together with the chosen ODE solver (i.e., the algorithm that turns the continuous system into a discrete one (in our case, the forward Euler method) affects the stability of the *simulated* system. I.e., too large a time step could lead to divergence even if the modeled physical system is stable

Descent (SGD) with a learning rate of $3\text{e}{-3}$ and batch size of 64, optimizing (65). We use $(d_{\mathbf{u}}+1)$ environments with $12,000$ data points each and train for $4,000$ epochs. To ensure stability, we clip the gradient norms to 0.5. Explicitly parameterizing $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ (when applicable) yields inferior results; thus, we do not explore this approach in our experiments.

Metric. We report the Mean Correlation Coefficient (MCC) (Hyvarinen and Morioka, 2016) to measure the correlation between the learned and true control signals (for training, we do not use knowledge of the control signal, only of its covariance). MCC has been used in prior works (Khemakhem et al., 2020b; Kivva et al., 2022) to quantify identifiability; it measures linear correlations up to permutation of the components. To compute the best permutation, a linear sum assignment problem is solved and finally, the correlation coefficients are computed and averaged.

Table 1: Validation of our identifiability claim, i.e., learning $\mathbf{u}$ from observations with different $\mathbf{C}$ and $\mathbf{B}$, and (non-)zero mean $\boldsymbol{\mu}^e_{\mathbf{u}}$ for $\mathbf{u}_t$. We use the minimal $(d_{\mathbf{u}}+1)$ number of environments. In the rightmost column, $\mathbf{B} \neq \mathbf{I}$, $\mathbf{C} \neq \mathbf{I}$, $\boldsymbol{\mu}^e_{\mathbf{u}} \neq \mathbf{0}$, and the e^{th} variance component to 0.9999, the others to 0.0001, yielding a well-conditioned $\boldsymbol{\Delta}$ (cf. Appx. D.1). Mean and standard deviation are reported across 5 runs. $d_{\mathbf{u}}$ is the dimensionality of $\mathbf{u}$ ($d_{\mathbf{x}} = d_{\mathbf{u}} = d_{\mathbf{y}}$), $|E|$ is the number of environments, Mean Correlation Coefficient (MCC) measures identifiability in $[0; 1]$ (higher is better)

$d_{\mathbf{u}}$	$\lvert E\rvert$	MCC ↑								
		$\boldsymbol{\mu}^e_{\mathbf{u}} \neq \mathbf{0}$				$\boldsymbol{\mu}^e_{\mathbf{u}} = \mathbf{0}$				$\boldsymbol{\mu}^e_{\mathbf{u}} \neq \mathbf{0}$
		$\mathbf{B} = \mathbf{I}$		$\mathbf{B} \neq \mathbf{I}$		$\mathbf{B} = \mathbf{I}$		$\mathbf{B} \neq \mathbf{I}$		$\mathbf{B} \neq \mathbf{I}$
		$\mathbf{C} = \mathbf{I}$	$\mathbf{C} \neq \mathbf{I}$	$\mathbf{C} = \mathbf{I}$	$\mathbf{C} \neq \mathbf{I}$	$\mathbf{C} = \mathbf{I}$	$\mathbf{C} \neq \mathbf{I}$	$\mathbf{C} = \mathbf{I}$	$\mathbf{C} \neq \mathbf{I}$	$\mathbf{C} \neq \mathbf{I}$
2	3	$0.866_{\pm 0.033}$	$0.767_{\pm 0.158}$	$0.730_{\pm 0.191}$	$0.697_{\pm 0.090}$	$0.633_{\pm 0.104}$	$0.675_{\pm 0.132}$	$0.734_{\pm 0.158}$	$0.725_{\pm 0.115}$	$0.968_{\pm 0.055}$
3	4	$0.901_{\pm 0.054}$	$0.910_{\pm 0.061}$	$0.916_{\pm 0.044}$	$0.861_{\pm 0.062}$	$0.659_{\pm 0.201}$	$0.618_{\pm 0.241}$	$0.633_{\pm 0.110}$	$0.667_{\pm 0.109}$	$1.000_{\pm 0.000}$
5	6	$0.892_{\pm 0.057}$	$0.929_{\pm 0.025}$	$0.928_{\pm 0.026}$	$0.911_{\pm 0.045}$	$0.657_{\pm 0.116}$	$0.618_{\pm 0.079}$	$0.620_{\pm 0.025}$	$0.539_{\pm 0.078}$	$0.995_{\pm 0.009}$
8	9	$0.943_{\pm 0.006}$	$0.940_{\pm 0.008}$	$0.941_{\pm 0.009}$	$0.867_{\pm 0.039}$	$0.585_{\pm 0.144}$	$0.479_{\pm 0.129}$	$0.523_{\pm 0.016}$	$0.414_{\pm 0.031}$	$0.977_{\pm 0.011}$
10	11	$0.939_{\pm 0.011}$	$0.915_{\pm 0.023}$	$0.925_{\pm 0.017}$	$0.924_{\pm 0.042}$	$0.708_{\pm 0.042}$	$0.611_{\pm 0.043}$	$0.604_{\pm 0.063}$	$0.525_{\pm 0.097}$	$0.996_{\pm 0.006}$

Results. From our ablations, it is prevalent that when the control signal $\mathbf{u}_t$ has a non-zero mean, it makes the learning problem easier; this holds for both known ($\mathbf{B} = \mathbf{I}$) and unknown ($\mathbf{B} \neq \mathbf{I}$) intervention targets or whether we directly observe the state $\mathbf{x}_t$ ($\mathbf{C} = \mathbf{I}$) or not ($\mathbf{C} \neq \mathbf{I}$) (all but the rightmost column in Tab. 1). These MCCs are also comparable to the best MCCs in (Ahuja et al., 2022b; Kivva et al., 2022; Willetts and Paige, 2021). We also report the MCC when the environment variability matrix $\boldsymbol{\Delta}$ is well-conditioned, since it directly affects how diverse the environments are (the rightmost column in Tab. 1). A better conditioned $\boldsymbol{\Delta}$ (with a condition number of one) yields higher MCCs. This suggests that when the environments are more diverse (quantified by the condition number of $\boldsymbol{\Delta}$), we get better identifiability. Thus, we recommend practitioners that—while considering any constraints in the physical system—they should strive to design experiments with a $\boldsymbol{\Delta}$ matrix with the lowest possible condition number (cf. also Appx. D.1).

5. Related work

LTI systems. LTI systems are widely used in machine learning and science, e.g., (Grewal and Andrews, 2010; Schiff, 2009; Athans, 1974; Mesot and Barber, 2007; Kalman, 1960b), since they are convenient to model temporal systems. Learning the system parameters (system identification)

has a vast literature so we do not attempt to summarize them here, see (Åström and Eykhoff, 1971; Ljung, 1998, 2010; Galrinho, 2016) and references therein. Recent works on LTI systems include studying polynomial-complexity (in both time and samples) algorithms for system identification (Bakshi et al., 2023a; Dean et al., 2020; Simchowitz et al., 2019), prediction and estimation through the no-regret learning framework (Sarkar and Rakhlin, 2019; Hardt et al., 2016; Simchowitz et al., 2018) and learning mixtures of such systems (Chen and Poor, 2022; Bakshi et al., 2023b).

Nonlinear ICA. Independent Component Analysis (Comon, 1994; Hyvärinen and Oja, 2000) comprises statistical methods to identify latent variables and is now a fundamental primitive in SEM. Identifiability is impossible in the nonlinear case without specific assumptions (Darmois, 1951; Hyvärinen and Pajunen, 1999); even the linear case requires non-Gaussian source (latent) variables. Recent works on nonlinear ICA incorporate auxiliary variables (Hyvarinen et al., 2019; Gresele et al., 2019; Khemakhem et al., 2020a; Hälvä et al., 2021; Buchholz et al., 2023; Rajendran et al., 2024), exploit temporal structure in the data (Hyvarinen and Morioka, 2017, 2016; Hälvä and Hyvärinen, 2020; Morioka et al., 2021; Monti et al., 2020; Hyvarinen et al., 2010; Klindt et al., 2021; Zimmermann et al., 2021), or restrict the model class (Shimizu et al., 2006; Hoyer et al., 2008; Zhang and Hyvarinen, 2012; Gresele et al., 2021; Kivva et al., 2022). Several works related nonlinear ICA to SEM estimation (Gresele et al., 2021; Monti et al., 2020; Shimizu et al., 2006; von Kügelgen et al., 2021; Hyvärinen et al., 2023; Reizinger et al., 2023) by inverting the data generating process—i.e., estimating the inverse functional assignment with an inference model. Another approach towards identifiability is to assume access to multiple environments, where either the distributions or some property of the model class changes (Gresele et al., 2019). Our work is similar to the latter: we design interventions in multiple environments to aid identifiability.

Interventional and temporal models. Recent works studied identifiability under interventional data, e.g., (Brehmer et al., 2022; Ahuja et al., 2022a,b; Lachapelle et al., 2022; Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Jiang and Aragam, 2023; Liang et al., 2023; Rajendran et al., 2024; von Kügelgen et al., 2023). These works assume intervening on exactly one variable, or require paired counterfactual data—our result does not require such assumptions. Perhaps the most closely related works are CITRIS (Lippe et al., 2022a) and its variants (Lippe et al., 2022b,c), TDRL (Yao et al., 2022) and LEAP (Yao et al., 2021). These works consider representation learning from temporal data; however, there are differences: e.g., CITRIS considers interventions that are changing as a function of time, such as the sequence of frames in a video. Moreover, they assume that the intervention targets are known a priori. Due to such differences, neither of these results nor their methodology directly translates to our setting.

Multienvironmental Causal Discovery (CD). There are multiple works investigating CD from multiple environments. Peters et al. (2015) consider linear SEMs, where the marginal variances of the effect variables are the same across environments; Ghassami et al. (2017), on the other hand, assume the same weights across environments and also require that the ratio of the cause and effect variances are different. In a follow-up work, Ghassami et al. (2018) propose a linear regression based CD method for linear SEMs with independence tests. Wang et al. (2018) study the case when the linear SEM weights are different across environments, but either the cause or the effect marginal variances are the same in a pair of environments. Perry et al. (2022) investigate bivariate CD and relax the i.i.d. assumption to sparse distribution shifts across environments, i.e., only a subset and not all conditionals change in each environment. Our multienvironmental identifiability and causal

discovery result resembles to some extent the causal identifiability result in multimodal Contrastive Learning (CL) (Morioka and Hyvarinen, 2023), which can be thought of as a causal and multimodal generalization of Time-Contrastive Learning (TCL) (Hyvarinen and Morioka, 2016).

6. Discussion

Limitations. Our results concern an important and widely used model class; however, the assumptions of linearity and time-invariance may be restrictive in some applications, and one may need to constrain the control signal, e.g., for safety reasons—we leave this for future work. While our theory is general enough to handle noise via a denoising argument, our experimental log-likelihood formulation assumes noiselessness as is common in the ICA literature (Hyvärinen and Oja, 2000; Gresele et al., 2021)—this can be a limiting factor in practice where measurement noise can be arbitrary, though our preliminary experiments suggest some robustness even when observations are noisy (cf. Tab. 2 in Appx. D.2). Another technical aspect is that to model $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, we parametrize the linear map by a single matrix; however, this makes extracting the model parameters non-trivial. Future work could relax such assumptions.

Extensions to related works. We extend the ideas of Reizinger et al. (2023). In the context of causal inference, they show that identifiability via ICA also yields the underlying DAG, i.e., nonlinear ICA can be used for CD for causal data generating processes, generalizing results for linear models (Shimizu et al., 2006). Furthermore, they discuss how contrastive nonlinear ICA (Zimmermann et al., 2021) can be seen as a practical realization of the Causal de Finetti theorem (Guo et al., 2022); however, it remained an open question whether other ICA algorithms can be seen as such. By showing that a linear ICA method with sufficient environment variability (akin to TCL (Hyvarinen and Morioka, 2016)) can identify dynamical and thus causal systems, our work strengthens this connection between the ICA and causal perspectives. In other words, we show that assumptions derived from the ICA literature can be used to design interventions (control signals for the experiments), thereby conceptually introducing agency into the framework. Interestingly, our result does not prescribe the number of state variables that need to be intervened on, it only requires sufficient variability of the control signal. Furthermore, we show that HMMs naturally satisfy the conditions for the CdF theorem (Guo et al., 2022), showing a potential reason why system identification methods can succeed in this model class. Since HMMs can have continuous parameters, our contribution corroborates the conjecture of Guo et al. (2022) about the existence (at least in a restricted model class) of a CdF theorem for continuous CdF parameters.

Conclusion. In this work, we apply advances in the causality literature towards the practical application of LTI systems. While identifiability in Gaussian LTI systems has a long history in control theory, our work provides a different means of achieving it via an interventional and multi-environmental perspective. We show that with a precise environment variability condition on the control (intervention) signal, a Gaussian LTI system is identifiable in the multi-environment case—i.e., it does not require white noise, which can be problematic for physical systems. This can be interpreted as an equivalence of the passive Independent Component Analysis (ICA) perspective of identifiability, i.e., learning from a provided data set, to the agency-based (interventional) identifiability notion of Causal Representation Learning (CRL). Finally, we connect Hidden Markov Models (HMMs) to an extension of the CdF theorem (Guo et al., 2022) with continuous parameters, providing a potential reason for why system identification is possible in HMMs.

Acknowledgments

We thank anonymous reviewers for useful comments. The authors also thank Siyuan Guo for her insights regarding the Causal de Finetti connection, Felix Leeb for discussing the practical implications of this paper, Zsolt Kollár and András Retzler for fruitful discussions on system identification, and Sara Magliacane for suggesting improvements for the experiments. Goutham Rajendran and Pradeep Ravikumar acknowledge the support of AFRL and DARPA via FA8750-23-2-1015, ONR via N00014-23-1-2368, NSF via IIS-1909816, IIS-1955532, and also acknowledge the support of JPMorgan Chase & Co. AI Research. Wieland Brendel acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. Wieland Brendel is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Patrik Reizinger thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support and acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

References

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022a.

Kartik Ahuja, Yixin Wang, Divyat Mahajan, and Yoshua Bengio. Interventional Causal Representation Learning, September 2022b. URL http://arxiv.org/abs/2209.11924. arXiv:2209.11924 [cs, stat].

Karl Johan Åström and Peter Eykhoff. System identification—a survey. *Automatica*, 7(2):123–162, 1971.

Michael Athans. The importance of kalman filtering methods for economic systems. In *Annals of Economic and Social Measurement, Volume 3, number 1*, pages 49–64. NBER, 1974.

Kendall Atkinson, Weimin Han, and David E Stewart. *Numerical solution of ordinary differential equations*. John Wiley & Sons, 2011.

Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. A new approach to learning linear dynamical systems. *arXiv preprint arXiv:2301.09519*, 2023a.

Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. Tensor decompositions meet control theory: learning general mixtures of linear dynamical systems. In *International Conference on Machine Learning*, pages 1549–1563. PMLR, 2023b.

Wolfgang Borutzky, editor. *Bond graph modelling of engineering systems*. Springer, New York, NY, 2011 edition, June 2011.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

Yanxi Chen and H Vincent Poor. Learning mixtures of linear dynamical systems. In *International Conference on Machine Learning*, pages 3507–3557. PMLR, 2022.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022.

George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

Miguel Galrinho. *Least squares methods for system identification of structured models*. PhD thesis, KTH Royal Institute of Technology, 2016.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning Causal Structures Using Regression Invariance. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain Causal Structure Learning in Linear Systems. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-View Nonlinear ICA. *arXiv:1905.06642 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1905.06642. arXiv: 1905.06642.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *arXiv:2106.05200 [cs, stat]*, June 2021. URL http://arxiv.org/abs/2106.05200. arXiv: 2106.05200.

Mohinder S Grewal and Angus P Andrews. Applications of kalman filtering in aerospace 1960 to the present [historical perspectives]. *IEEE Control Systems Magazine*, 30(3):69–78, 2010.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. *arXiv:2203.15756 [cs, math, stat]*, March 2022. URL http://arxiv.org/abs/2203.15756. arXiv: 2203.15756.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *arXiv:1605.06336 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1605.06336. arXiv: 1605.06336.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, April 2017. URL http://proceedings.mlr.press/v54/hyvarinen17a.html. ISSN: 2640-3498.

Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*. J. Wiley, New York, 2001. ISBN 978-0-471-40540-5.

Aapo Hyvarinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. page 23, 2010. URL https://www.jmlr.org/papers/volume11/hyvarinen10a/hyvarinen10a.pdf.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. *arXiv:1805.08651 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1805.08651. arXiv: 1805.08651.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*, 2023.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, June 2000. ISSN 0893-6080. doi: 10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear, February 2023. URL http://arxiv.org/abs/2302.02672. arXiv:2302.02672 [cs, stat].

Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series. *arXiv:2006.12107 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2006.12107. arXiv: 2006.12107.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA. *arXiv:2106.09620 [cs, stat]*, June 2021. URL http://arxiv.org/abs/2106.09620. arXiv: 2106.09620.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. *arXiv preprint arXiv:2306.02899*, 2023.

Rudolf E Kalman. On the general theory of control systems. In *Proceedings First International Conference on Automatic Control, Moscow, USSR*, pages 481–492, 1960a.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960b.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, June 2020a. URL http://proceedings.mlr.press/v108/khemakhem20a.html. ISSN: 2640-3498.

Ilyes Khemakhem, Diederik P Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models. *NeurIPS2020*, 2020b.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *arXiv preprint arXiv:2106.15563*, 2021.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. *arXiv preprint arXiv:2206.10044*, 2022.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. *arXiv:2007.10930 [cs, stat]*, March 2021. URL http://arxiv.org/abs/2007.10930. arXiv: 2007.10930.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *arXiv preprint arXiv:2305.17225*, 2023.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences, June 2022a. URL http://arxiv.org/abs/2202.03169. Number: arXiv:2202.03169 arXiv:2202.03169 [cs, stat].

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. iCITRIS: Causal Representation Learning for Instantaneous Temporal Effects. July 2022b. URL https://openreview.net/forum?id=xeDKTZsZ7Z7.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Intervention Design for Causal Representation Learning. July 2022c. URL https://openreview.net/forum?id=TpVzjh4M2hd.

Lennart Ljung. *System identification*. Springer, 1998.

Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010.

Bertrand Mesot and David Barber. Switching linear dynamical systems for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1850–1858, 2007.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal Discovery with General Non-Linear Relationships using Non-Linear ICA. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, August 2020. URL http://proceedings.mlr.press/v115/monti20a.html. ISSN: 2640-3498.

Hiroshi Morioka and Aapo Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 3399–3426. PMLR, April 2023. URL https://proceedings.mlr.press/v206/morioka23a.html. ISSN: 2640-3498.

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvärinen. Independent Innovation Analysis for Nonlinear Vector Autoregressive Process. *arXiv:2006.10944 [cs, stat]*, February 2021. URL http://arxiv.org/abs/2006.10944. arXiv: 2006.10944.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-511-80316-1. doi: 10.1017/CBO9780511803161. URL http://ebooks.cambridge.org/ref/id/CBO9780511803161.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis, October 2022. URL http://arxiv.org/abs/2206.02013. arXiv:2206.02013 [cs, stat].

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv:1501.01332 [stat]*, November 2015. URL http://arxiv.org/abs/1501.01332. arXiv: 1501.01332.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. *Journal of Statistical Computation and Simulation*, 88(16): 3248–3248, November 2018. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2018.1505197. URL https://www.tandfonline.com/doi/full/10.1080/00949655.2018.1505197.

Rik Pintelon and Johan Schoukens. *System identification*. Wiley-IEEE Press, New York, NY, April 2004.

Goutham Rajendran, Bohdan Kivva, Ming Gao, and Bryon Aragam. Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. *Advances in Neural Information Processing Systems*, 34:18660–18672, 2021.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint*, 2024.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based Causal Discovery with Nonlinear ICA. *Transactions on Machine Learning Research*, April 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2Yo9xqR6Ab.

Sam Roweis and Zoubin Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345, February 1999. ISSN 08997667. doi: 10.1162/089976699300016674. URL http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3daph%26AN%3d1555006%26site%3dehost-live. Publisher: MIT Press.

Mark Rudelson and Roman Vershynin. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Steven J Schiff. Kalman meets neuron: the emerging intersection of control theory with neuroscience. In *2009 annual international conference of the IEEE engineering in medicine and biology society*, pages 3318–3321. IEEE, 2009.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. arXiv:2102.11107.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvarinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. page 28, 2006.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires and Caroline Uhler. Causal structure learning: a combinatorial perspective. *Foundations of Computational Mathematics*, pages 1–35, 2022.

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *arXiv preprint arXiv:2306.00542*, 2023.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style, June 2021. URL http://arxiv.org/abs/2106.04619. arXiv: 2106.04619.

Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34:5443–5455, 2021.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct Estimation of Differences in Causal Graphs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Matthew Willetts and Brooks Paige. I Don't Need $\mathbf{u}$: Identifiable Non-Linear ICA Without Side Information. *arXiv:2106.05238 [cs, stat]*, June 2021. URL http://arxiv.org/abs/2106.05238. arXiv: 2106.05238.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning Temporally Causal Latent Processes from General Temporal Data. *arXiv:2110.05428 [cs, stat]*, October 2021. URL http://arxiv.org/abs/2110.05428. arXiv: 2110.05428 version: 1.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally Disentangled Representation Learning, October 2022. URL http://arxiv.org/abs/2210.13647. arXiv:2210.13647 [cs, stat].

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.

Kun Zhang and Aapo Hyvarinen. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv:1205.2599 [cs, stat]*, May 2012. URL http://arxiv.org/abs/1205.2599. arXiv: 1205.2599.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive Learning Inverts the Data Generating Process. *arXiv:2102.08850 [cs]*, February 2021. URL http://arxiv.org/abs/2102.08850. arXiv: 2102.08850.

Appendix A. Linear Time Invariant systems

In this section, we review some standard concepts about LTI systems. The notions of controllability and identifiability were introduced by Kalman (Kalman, 1960a) and it is now widely accepted that they govern when an LTI system can be learnt.

A.1. Controllability

Definition 13 (Controllability matrix) *The controllability matrix* $\mathbf{M}_c \in \mathbb{R}^{d_{\mathbf{x}} \times (d_{\mathbf{x}} \cdot d_{\mathbf{u}})}$ *for Defn. 1 is defined as*

$$\mathbf{M}_c = \left[\mathbf{B}; \mathbf{AB}; \dots; \mathbf{A}^{d_{\mathbf{x}}-1}\mathbf{B}\right], \tag{13}$$

Because of our system dynamics, the controllability matrix intuitively captures the state space that can be reached eventually.

Lemma 14 *The similarity transformation* $\mathbf{PAP}^{-1}, \mathbf{PB}$ *does not change the rank of* $\mathbf{M}_c$*.*

Proof Since $\left[\mathbf{PAP}^{-1}\right]^i = \mathbf{PA}^i\mathbf{P}^{-1}$ and $\mathbf{P}$ are full-rank, we have

$$\mathbf{M}_c\left(\mathbf{PAP}^{-1}, \mathbf{PB}\right) = \left[\mathbf{PB}; \mathbf{PAB}; \dots; \mathbf{PA}^{d_{\mathbf{x}}-1}\mathbf{B}\right] = \mathbf{PM}_c \tag{14}$$

■

A.2. Observability

Definition 15 (Observability matrix) *The observability matrix* $\mathbf{M}_o \in \mathbb{R}^{(d_{\mathbf{x}} \cdot d_{\mathbf{y}}) \times d_{\mathbf{x}}}$ *for Defn. 1 is defined as*

$$\mathbf{M}_o = \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{d_{\mathbf{x}}-1} \end{pmatrix} \tag{15}$$

Similar to the controllability matrix, the observability matrix encapsulates the state space that we can observe eventually as the system evolves.

Lemma 16 *The similarity transformation* $\mathbf{PAP}^{-1}, \mathbf{CP}^{-1}$ *does not change the rank of* $\mathbf{M}_o$*.*

Proof Since $\left[\mathbf{PAP}^{-1}\right]^i = \mathbf{PA}^i\mathbf{P}^{-1}$ and $\mathbf{P}$ are full-rank, we have

$$\mathbf{M}_o\left(\mathbf{PAP}^{-1}, \mathbf{CP}^{-1}\right) = \begin{pmatrix} \mathbf{CP}^{-1} \\ \mathbf{CAP}^{-1} \\ \vdots \\ \mathbf{CA}^{d_{\mathbf{x}}-1}\mathbf{P}^{-1} \end{pmatrix} = \mathbf{M}_o\mathbf{P}^{-1} \tag{16}$$

■

A.3. Transformations of LTI systems

Definition 17 (Identifiability up to similarity transformations) *The LTI system parameters* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *are identifiable up to similarity transformations if for any other system parameters* $\mathbf{A}', \mathbf{B}', \mathbf{C}'$ *that fit the LTI system, there exists a full rank matrix* $\mathbf{P} \in \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$ *such that*

$$\mathbf{A}' = \mathbf{P}\mathbf{A}\mathbf{P}^{-1}, \quad \mathbf{B}' = \mathbf{P}\mathbf{B}, \quad \mathbf{C}' = \mathbf{C}\mathbf{P}^{-1} \tag{17}$$

Lemma 18 (Equivalence class of LTI systems) *Coordinate transformations of the state* $\mathbf{x}$ *by a full rank matrix* $\mathbf{P}$*, i.e.,* $\mathbf{x}' = \mathbf{P}\mathbf{x}$*, with the corresponding transformations of the systems matrices:*

$$\begin{aligned}\mathbf{A}' &= \mathbf{P}\mathbf{A}\mathbf{P}^{-1} \\ \mathbf{B}' &= \mathbf{P}\mathbf{B} \\ \mathbf{C}' &= \mathbf{C}\mathbf{P}^{-1},\end{aligned}$$

yield an equivalent transfer function.

Proof We start by the definition of the transfer function for the original and transformed systems:

$$\mathbf{H}(z) = \mathbf{C}\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}\right]^{-1}\mathbf{B} \tag{18}$$

$$\mathbf{H}'(z) = \mathbf{C}'\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}'\right]^{-1}\mathbf{B}' \tag{19}$$

We substitute the transformed matrices into $\mathbf{H}'(z)$ from the definition to get

$$\mathbf{H}'(z) = \mathbf{C}\mathbf{P}^{-1}\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{P}\mathbf{A}\mathbf{P}^{-1}\right]^{-1}\mathbf{P}\mathbf{B} \tag{20}$$

Then we rewrite $\mathbf{I}_{d_\mathbf{x}} = \mathbf{P}\mathbf{P}^{-1}$

$$= \mathbf{C}\mathbf{P}^{-1}\left[z\mathbf{P}\mathbf{P}^{-1} - \mathbf{P}\mathbf{A}\mathbf{P}^{-1}\right]^{-1}\mathbf{P}\mathbf{B} \tag{21}$$

$$= \mathbf{C}\mathbf{P}^{-1}\left[\mathbf{P}\left(z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}\right)\mathbf{P}^{-1}\right]^{-1}\mathbf{P}\mathbf{B} \tag{22}$$

$$= \mathbf{C}\mathbf{P}^{-1}\mathbf{P}\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}\right]^{-1}\mathbf{P}^{-1}\mathbf{P}\mathbf{B} = \mathbf{H}(z) \tag{23}$$

■

Remark 19 (Reasonable identifiability requirement for LTI systems) *Lem. 18 implies by the invariance of the transfer function that a reasonable identifiability result for LTI systems is one up to linear transformations.*

Lemma 20 (Equivalence of transfer functions LTI systems) *If two transfer functions* $\mathbf{H}(z) = \mathbf{H}'(z)$*, then the state is given up to a change of coordinate frame, and the relationship between the matrices of the two state-space representations will relate as:*

$$\begin{aligned}\mathbf{A}' &= \mathbf{P}\mathbf{A}\mathbf{P}^{-1} \\ \mathbf{B}' &= \mathbf{P}\mathbf{B} \\ \mathbf{C}' &= \mathbf{C}\mathbf{P}^{-1},\end{aligned}$$

where $\mathbf{P}$ *is a full rank matrix and* $\mathbf{x}' = \mathbf{P}\mathbf{x}$*.*

Proof Note that since we consider LTI systems, indeterminacies are only possible up to linear transformations. We assume that though $\mathbf{H}(z) = \mathbf{H}'(z)$ the matrices are transformed in the most general way, i.e. $\forall \mathbf{M} \leftarrow \mathbf{P}_l \mathbf{M} \mathbf{P}_r$

We start by the definition of the transfer function for the original and transformed systems:

$$\mathbf{H}(z) = \mathbf{C}\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}\right]^{-1}\mathbf{B} \tag{24}$$

$$\mathbf{H}'(z) = \mathbf{C}'\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{A}'\right]^{-1}\mathbf{B}' \tag{25}$$

We substitute the transformed matrices into $\mathbf{H}'(z)$ from the definition to get

$$\mathbf{H}'(z) = \mathbf{P}_l^C\mathbf{C}\mathbf{P}_r^C\left[z\mathbf{I}_{d_\mathbf{x}} - \mathbf{P}_l^A\mathbf{A}\mathbf{P}_r^A\right]^{-1}\mathbf{P}_l^B\mathbf{B}\mathbf{P}_r^B \tag{26}$$

$$= \mathbf{P}_l^C\mathbf{C}\mathbf{P}_r^C\left[\mathbf{P}_l^A\left(z(\mathbf{P}_l^A)^{-1}(\mathbf{P}_r^A)^{-1} - \mathbf{P}_l^A\mathbf{A}\right)\mathbf{P}_r^A\right]^{-1}\mathbf{P}_l^B\mathbf{B}\mathbf{P}_r^B \tag{27}$$

$$= \mathbf{P}_l^C\mathbf{C}\mathbf{P}_r^C(\mathbf{P}_r^A)^{-1}\left[z(\mathbf{P}_l^A)^{-1}(\mathbf{P}_r^A)^{-1} - \mathbf{A}\right]^{-1}(\mathbf{P}_l^A)^{-1}\mathbf{P}_l^B\mathbf{B}\mathbf{P}_r^B \tag{28}$$

For this expression to equal to $\mathbf{H}(z)$, it is it necessary to have $\mathbf{P}_l^C = \mathbf{P}_r^B = \mathbf{I}$, yielding

$$= \mathbf{C}\mathbf{P}_r^C(\mathbf{P}_r^A)^{-1}\left[z(\mathbf{P}_l^A)^{-1}(\mathbf{P}_r^A)^{-1} - \mathbf{A}\right]^{-1}(\mathbf{P}_l^A)^{-1}\mathbf{P}_l^B\mathbf{B} \tag{29}$$

Then left multiplying with the inverse of C, and doing the same from the right with that of B, we need to have $(z\mathbf{I} - A)^{-1}$, which requires that $\mathbf{P}_r^C(\mathbf{P}_r^A)^{-1} = \mathbf{I}$, $(\mathbf{P}_l^A)^{-1}\mathbf{P}_l^B = \mathbf{I}$, and $(\mathbf{P}_l^A)^{-1}(\mathbf{P}_r^A)^{-1} = \mathbf{I}$, which yields the coordinate transforms from the lemma. ■

Appendix B. Proofs

B.1. Proof of Theorem 7

Instate the setting of the Gaussian LTI system as described in § 3. In this section, we prove our main identifiability result Thm. 7. We start with the following lemma.

Lemma 21 *For each environment e, we have*

$$\mathbf{y}_t^e = \mathbf{C}\mathbf{A}^t\mathbf{x}_0^e + \sum_{i=1}^{t}\mathbf{C}\mathbf{A}^{i-1}\mathbf{B}\mathbf{u}_{t-i}^e + \boldsymbol{\varepsilon}_t^e \tag{30}$$

Proof For any fixed environment e, we simply repeatedly apply the LTI equations to get

$$\mathbf{y}_t^e = \mathbf{C}\mathbf{x}_t^e + \boldsymbol{\varepsilon}_t^e \tag{31}$$

$$= \mathbf{C}(\mathbf{A}\mathbf{x}_{t-1}^e + \mathbf{B}\mathbf{u}_{t-1}^e) + \boldsymbol{\varepsilon}_t^e \tag{32}$$

$$= \mathbf{C}\mathbf{A}(\mathbf{A}\mathbf{x}_{t-2}^e + \mathbf{B}\mathbf{u}_{t-2}^e) + \mathbf{C}\mathbf{B}\mathbf{u}_{t-1}^e + \boldsymbol{\varepsilon}_t^e \tag{33}$$

$$= \dots \tag{34}$$

$$= \mathbf{C}\mathbf{A}^t\mathbf{x}_0^e + \sum_{i=1}^{t}\mathbf{C}\mathbf{A}^{i-1}\mathbf{B}\mathbf{u}_{t-i}^e + \boldsymbol{\varepsilon}_t^e \tag{35}$$

■

Before proceeding to the proof, we emphasize that by centering both $\mathbf{u}_t$ and $\mathbf{y}_t$, we can assume, w.l.o.g., that $\mathbf{u}_t$ and, (by the linearity of the LTI system and the linearity of the expectation operator) $\mathbf{y}_t$ are zero-mean. By the same argument, we can also set $\mathbf{x}_0^e = \mathbf{0}$:

Lemma 22 (Zero-mean signals and zero initial state) *By the linearity of the system and the expectation operator, w.l.o.g., $\mathbf{u}_t$, $\mathbf{y}_t$, can be assumed to have zero means, and we can also set $\mathbf{x}_0 = \mathbf{0}$.*

Proof Lem. 21 describes the map from each $\mathbf{u}_i$ to $\mathbf{y}_t$—this is due to each $\mathbf{u}_i$ being an i.i.d. Gaussian sample; thus, their sum is also Gaussian. Denote $\mathbf{T}_t = \sum_{i=1}^{t} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}$. Since the Gaussian distribution is fully characterized by its mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we investigate how these two quantities are related:

$$\boldsymbol{\mu}_{\mathbf{u}} \mapsto \boldsymbol{\mu}_{\mathbf{y},t} := \mathbf{T}_t \boldsymbol{\mu}_{\mathbf{u}} + \mathbf{C}\mathbf{A}^t \mathbf{x}_0 \tag{36}$$

$$\boldsymbol{\Sigma}_{\mathbf{u}} \mapsto \boldsymbol{\Sigma}_{\mathbf{y},t} := \mathbf{T}_t \boldsymbol{\Sigma}_{\mathbf{u}} \mathbf{T}_t^{\top}. \tag{37}$$

When the initial state is non-zero, i.e., $\mathbf{x}_0 \neq \mathbf{0}$. Since $\mathbf{x}_0$ is a constant, it only affects the mean of $\mathbf{y}_t$. For a given t we can define $\hat{\mathbf{u}}_t := \mathbf{u}_t - \boldsymbol{\mu}_{\mathbf{u}}$ ($\boldsymbol{\mu}_{\mathbf{u}}$ is independent from t due to the i.i.d. assumption) and $\hat{\mathbf{y}}_t := \mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t}$. By assuming that we know $\boldsymbol{\mu}_{\mathbf{u}}$, we can calculate $\hat{\mathbf{u}}_t$, and we can use the empirical mean for calculating $\hat{\mathbf{y}}_t$. This means that instead of the original LTI system we conceptually use a modified LTI system, which has a zero-mean input and a zero-mean output (i.e., the transformations for calculating $\hat{\mathbf{u}}_t, \hat{\mathbf{y}}_t$ are considered part of the system). This is without loss of generality, since by the linearity of the system and the expectation operator, and since $\boldsymbol{\mu}_{\mathbf{u}}$ is known (i.e., it is a constant), we can always recover the mean of $\mathbf{y}_t$ by using a constant control signal $\mathbf{u}_t$, measuring the (constant) output, and adding that to $\hat{\mathbf{y}}_t$. ■

We restate our main theorem for convenience.

Theorem 7 *[LTI system identifiability with sufficient variability] For LTI systems satisfying Assums. 2 and 6, the Markov parameter matrix $\mathbf{G}$ is identifiable up to permutations and diagonal scaling.*

Proof [Proof of Thm. 7] Consider the dataset of observations $(\mathbf{y}_t^e)_{e\in E, t\leq T}$ for a horizon T. Suppose there exist two sets of parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}})$ that could have generated the dataset, we will now argue that they must be related by a similarity transformation.

Define $\mathbf{T}_t = \sum_{i=1}^{t} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}$, $\widetilde{\mathbf{T}}_t = \sum_{i=1}^{t} \widetilde{\mathbf{C}}\widetilde{\mathbf{A}}^{i-1}\widetilde{\mathbf{B}}$ and let $\boldsymbol{\Sigma}_{\mathbf{u}}^e = \mathrm{diag}((\boldsymbol{\sigma}_1^e)^2, \ldots, (\boldsymbol{\sigma}_{d_{\mathbf{u}}}^e)^2)$ be the diagonal matrix of variances. By assumption, we have $\mathbf{u}_i^e \sim (\boldsymbol{\Sigma}_{\mathbf{u}}^e)^{1/2}\mathcal{N}(\mathbf{0}; \mathbf{I})$ independently for all i. Also, by a standard denoising argument (Khemakhem et al., 2020a; Lachapelle et al., 2022; Kivva et al., 2022), we can set $\boldsymbol{\varepsilon}_t^e = 0$ by deconvolving the noise operator. Assuming $\mathbf{x}_0^e = \mathbf{0}$ via Lem. 22, using Lem. 21, we have

$$\mathbf{y}_t^e = \sum_{i=1}^{t} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}\mathbf{u}_{t-i}^e + \boldsymbol{\varepsilon}_t^e \sim \mathbf{T}_t(\boldsymbol{\Sigma}_{\mathbf{u}}^e)^{1/2}\mathcal{N}(\mathbf{0}; \mathbf{I}) \tag{38}$$

for all $e \in E, t \leq T$. Let the densities of $\mathbf{y}_t^e$ be denoted $p_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y})$ and $p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y})$ respectively in these two models under environment e. Let $p_{\mathcal{N}}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ denote the density of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\sigma}^2)$ evaluated at point x, so that

$$\ln p_{\mathcal{N}}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = -\frac{(x-\boldsymbol{\mu})^2}{2\boldsymbol{\sigma}^2} - \ln \boldsymbol{\sigma} - \ln \sqrt{2\pi} \tag{39}$$

Now, using a standard change of variables,

$$\ln p_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) = \ln|\det \mathbf{T}_t^{-1}| + \sum_{i\le d_{\mathbf{u}}} \ln p_{\mathcal{N}}((\mathbf{T}_t^{-1}\mathbf{y})_i; 0, (\boldsymbol{\sigma}_i^e)^2) \tag{40}$$

$$= \ln|\det \mathbf{T}_t^{-1}| - \sum_{i=1}^{d_{\mathbf{u}}} \left(\frac{(\mathbf{T}_t^{-1}\mathbf{y})_i^2}{2(\boldsymbol{\sigma}_i^e)^2} + \ln \boldsymbol{\sigma}_i^e + \ln\sqrt{2\pi} \right) \tag{41}$$

.

Similarly,

$$\ln p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) = \ln|\det \widetilde{\mathbf{T}}_t^{-1}| - \sum_{i=1}^{d_{\mathbf{u}}} \left(\frac{(\widetilde{\mathbf{T}}_t^{-1}\mathbf{y})_i^2}{2(\boldsymbol{\sigma}_i^e)^2} + \ln \boldsymbol{\sigma}_i^e + \ln\sqrt{2\pi} \right) \tag{42}$$

Now we will consider the log-odds $q_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) = \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) - \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},0,t}(\mathbf{y})$ of the e-th environment with respect to a fixed 0-th environment. Note that they match under the two models as per our assumptions. We use the analytic expression for the densities to obtain

$$q_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) = \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) - \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},0,t}(\mathbf{y}) \tag{43}$$

$$= \sum_{i=1}^{d_{\mathbf{u}}} \left(-\frac{(\mathbf{T}_t^{-1}\mathbf{y})_i^2}{2} \left(\frac{1}{(\boldsymbol{\sigma}_i^e)^2} - \frac{1}{(\boldsymbol{\sigma}_i^0)^2} \right) - \ln \frac{\boldsymbol{\sigma}_i^e}{\boldsymbol{\sigma}_i^0} \right) \tag{44}$$

Analogously defining $q_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) = \ln p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) - \ln p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},0,t}(\mathbf{y})$ and using the same calculations, we get

$$q_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) = \sum_{i=1}^{d_{\mathbf{u}}} \left(-\frac{(\widetilde{\mathbf{T}}_t^{-1}\mathbf{y})_i^2}{2} \left(\frac{1}{(\boldsymbol{\sigma}_i^e)^2} - \frac{1}{(\boldsymbol{\sigma}_i^0)^2} \right) - \ln \frac{\boldsymbol{\sigma}_i^e}{\boldsymbol{\sigma}_i^0} \right) \tag{45}$$

Since we observe the same distribution in both parameter settings, we have

$$q_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) = \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},e,t}(\mathbf{y}) - \ln p_{\mathbf{A},\mathbf{B},\mathbf{C},0,t}(\mathbf{y}) \tag{46}$$

$$= \ln p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) - \ln p_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},0,t}(\mathbf{y}) \tag{47}$$

$$= q_{\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}},e,t}(\mathbf{y}) \tag{48}$$

Substituting the expressions, we get

$$\sum_{i=1}^{d_{\mathbf{u}}} \left(((\mathbf{T}_t^{-1}\mathbf{y})_i^2 - (\widetilde{\mathbf{T}}_t^{-1}\mathbf{y})_i^2) \left(\frac{1}{(\boldsymbol{\sigma}_i^e)^2} - \frac{1}{(\boldsymbol{\sigma}_i^0)^2} \right) \right) = 0 \tag{49}$$

Define the variable $\mathbf{u} = \mathbf{T}_t^{-1}\mathbf{y}$ and the matrix $\mathbf{H}_t = \widetilde{\mathbf{T}}_t^{-1}\mathbf{T}_t$, then $\widetilde{\mathbf{T}}_t^{-1}\mathbf{y} = \mathbf{H}_t\mathbf{u}$. Therefore, the above expression simplifies to

$$\sum_{i=1}^{d_{\mathbf{u}}} \left(\mathbf{u}_i^2 - (\mathbf{H}_t\mathbf{u})_i^2) \right) \boldsymbol{\Delta}_{e,i} = 0 \tag{50}$$

Expanding out the inner term,

$$\sum_{i=1}^{d_{\mathbf{u}}} \left(\mathbf{u}_i^2 - (\sum_{j=1}^{d_{\mathbf{u}}} (\mathbf{H}_t)_{i,j} \mathbf{u}_j)^2 \right) \mathbf{\Delta}_{e,i} = 0 \tag{51}$$

Fix an arbitrary $i \leq d_{\mathbf{u}}$. Because this is a functional identity, we can differentiate with respect to $\mathbf{u}_i$. Note that the coefficient of $\mathbf{u}_i^2$ in (51) is

$$\mathbf{\Delta}_{e,i} - (\mathbf{H}_t)_{1,i}^2 \mathbf{\Delta}_{e,1} - (\mathbf{H}_t)_{2,i}^2 \mathbf{\Delta}_{e,2} - \ldots = \mathbf{\Delta}_{e,i} - \sum_{k \leq d_{\mathbf{u}}} (\mathbf{H}_t)_{k,i}^2 \mathbf{\Delta}_{e,k} \tag{52}$$

Also, the coefficient of $\mathbf{u}_i$ in the k-th term of (51) is

$$2\mathbf{\Delta}_{e,k} \sum_{j \leq d_{\mathbf{u}}, j \neq i} (\mathbf{H}_t)_{k,i} \mathbf{H}_{k,j} \mathbf{u}_j \tag{53}$$

which we obtained by expanding out the k-th term as

$$(\sum_{j \leq d_{\mathbf{u}}} (\mathbf{H}_t)_{k,j} \mathbf{u}_j)^2 \mathbf{\Delta}_{e,k} = \left(\sum_{j_1, j_2 \leq d_{\mathbf{u}}} (\mathbf{H}_t)_{k,j_1} (\mathbf{H}_t)_{k,j_2} \mathbf{u}_{j_1} \mathbf{u}_{j_2} \right) \mathbf{\Delta}_{e,k} \tag{54}$$

Putting them together, we can finally write the derivative of (51) with respect to $\mathbf{u}_i$ as

$$2\mathbf{u}_i (\mathbf{\Delta}_{e,i} - \sum_{k \leq d_{\mathbf{u}}} (\mathbf{H}_t)_{k,i}^2 \mathbf{\Delta}_{e,k}) - 2 \sum_{k \leq d_{\mathbf{u}}, j \leq d_{\mathbf{u}}, j \neq i} (\mathbf{H}_t)_{k,i} (\mathbf{H}_t)_{k,j} \mathbf{u}_j \mathbf{\Delta}_{e,k} = 0 \tag{55}$$

Now, this is yet another functional identity. So, we fix an arbitrary $j \neq i$ and again differentiate with respect to $\mathbf{u}_j$ to get

$$\sum_{k \leq d_{\mathbf{u}}} (\mathbf{H}_t)_{k,i} (\mathbf{H}_t)_{k,j} \mathbf{\Delta}_{e,k} = 0 \tag{56}$$

Define the vector $\boldsymbol{h} \in \mathbb{R}^{d_{\mathbf{u}}}$ with the k-th entry being $(\mathbf{H}_t)_{k,i} (\mathbf{H}_t)_{k,j}$ (recall that i, j have been fixed). Then, the above equation can be written succinctly as

$$\mathbf{\Delta} \cdot \boldsymbol{h} = 0 \tag{57}$$

Since $\mathbf{\Delta}$ has rank d_u, we must have $\boldsymbol{h} = 0$. That is, for every $k \leq d_{\mathbf{u}}$, we have $(\mathbf{H}_t)_{k,i} (\mathbf{H}_t)_{k,j} = 0$. Note that the choice of $i \neq j$ was arbitrary and could have been any other two indices. This implies that for all $k, i, j \leq n$ with $i \neq j$, we have

$$(\mathbf{H}_t)_{k,i} (\mathbf{H}_t)_{k,j} = 0 \tag{58}$$

Therefore, we conclude that each row of $\mathbf{H}_t$ has at most one nonzero entry. Moreover, $\mathbf{H}_t$ is full rank since $\mathbf{T}_t, \widetilde{\mathbf{T}}_t$ are invertible. Therefore, $\mathbf{H}_t$ must be a scaled permutation matrix, i.e. $\mathbf{H}_t = \mathbf{PD}$ where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix. This implies

$$\sum_{i=1}^{t} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B} = \mathbf{T}_t = \widetilde{\mathbf{T}}_t \mathbf{H}_t = \widetilde{\mathbf{T}}_t \mathbf{P}\mathbf{D} = (\sum_{i=1}^{t} \widetilde{\mathbf{C}}\widetilde{\mathbf{A}}^{i-1}\widetilde{\mathbf{B}})\mathbf{P}\mathbf{D} \tag{59}$$

Since $t \leq T$ was arbitrary, using this, we can conclude that we can identify the system's Markov parameters given by

$$\mathbf{G} = [\mathbf{I}, \mathbf{CB}, \mathbf{CAB}, \ldots, \mathbf{CA}^{T-1}\mathbf{B}] \tag{60}$$

up to permutations and diagonal scaling.

■

B.2. Proof of Lemma 12

In this section, we prove Thm. 12, which we restate for convenience

Lemma 23 *[Identifiability via the multi-environmental log-likelihood] Under Assums. 2, 6 and 9, the parameters that maximize the log-likelihood of a Gaussian LTI system relate to the ground truth via a linear transformation; or, equivalently, the corresponding transfer function is equivalent to the ground truth up to permutations and scalings.*

Proof Let $p_{\mathcal{N}}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ denote the density of the univariate Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}; \boldsymbol{\sigma}^2\right)$ evaluated at point x, so that

$$\ln p_{\mathcal{N}}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = -\frac{(x-\boldsymbol{\mu})^2}{2\boldsymbol{\sigma}^2} - \ln \boldsymbol{\sigma} - \ln \sqrt{2\pi}. \tag{61}$$

Also note that the control signal $\mathbf{u}_t$ can be expressed by Defn. 1 as

$$\begin{aligned}
\mathbf{u}_t &= \mathbf{B}^{-1}\left[\mathbf{C}^{-1}\mathbf{y}_{t+1} - \mathbf{AC}^{-1}\mathbf{y}_t\right] && (62)\\
&= \mathbf{B}^{-1}\begin{pmatrix}\mathbf{I} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & -\mathbf{AC}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{y}_{t+1}\\ \mathbf{y}_t\end{pmatrix} && (63)\\
&= \mathbf{T}^{-1}\begin{pmatrix}\mathbf{y}_{t+1}\\ \mathbf{y}_t\end{pmatrix}, && (64)
\end{aligned}$$

where we use $\mathbf{T}^{-1}$ with a slight abuse of notation to refer to the map from the observations to the control signal. (63) inherently has the same indeterminacies as the transfer function (cf. Lem. 20), irrespective of the symmetries of the distribution of $\mathbf{u}_t$. By assumption, $\mathbf{u}_t^e$ follows a normal distribution for each environment e; furthermore its components are idependent. Thus, the log-likelihood across all environments factorizes both over environments and dimensions, yielding:

$$\mathcal{L} = \sum_e \sum_t \ln p_{\mathcal{N}}\left(\mathbf{u}_t^e | \mathbf{y}_{t+1}^e, \mathbf{y}_t^e, \mathbf{A}, \mathbf{B}, \mathbf{C}\right) \tag{65}$$

Assuming zero mean $\mathbf{u}_t^e$ and exploiting that $\boldsymbol{\Sigma}^e$ is known, the multivariate Gaussian log-likelihood becomes a weighted least squares problem:

$$\mathcal{L} \propto \sum_e \sum_t (\mathbf{u}_t^e)^\top \boldsymbol{\Sigma}^e \mathbf{u}_t^e.$$

Also note that the LTI model considered is a first-order Markov chain; thus, conditioning on $\mathbf{y}_{t+1}^e, \mathbf{y}_t^e, \mathbf{A}, \mathbf{B}, \mathbf{C}$ deterministically determines $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ in the noiseless case, removing any other

conditional dependence. For the remainder of the proof, we assume that the control signal has zero mean in each environment. By the rotational symmetry of the Gaussian distribution, we would have a rotational indeterminacy. As we show next, the sufficient variability condition of Assum. 6 will break those symmetries. Since Thm. 7 holds for any multi-environmental setting which satisfies Assum. 6, let us illustrate with a simple example for clarity. Assume that the baseline environment (with index 0) is an isotropic Gaussian with a variance of 1. Then, let each environment e have $(\boldsymbol{\sigma}_e^e)^2 = 2$ (i.e., the e^{th} variance component is two, all the others are unchanged). This yields a full-rank matrix $\boldsymbol{\Delta}$, satisfying Assum. 6. Note that in this case, each environment can remove one degree of freedom (the corresponding Gaussian is invariant to rotations in the subspace not including the e^{th} component). By having as many environments as components, this means that we can remove the rotational indeterminacy, yielding an identifiability of $\mathbf{T}^{-1}$ up to scaling and permutations. ■

Appendix C. The Causal de Finetti connection

In this section, we detail the conditions required for the CdF theorem (Guo et al. (2022); cf. Thm. 25) to hold, followed by stating the CdF theorem. Then, we show how these conditions are fulfilled in HMMs.

C.1. The CdF conditions

Often, RVs are assumed to be independent and identically distributed (i.i.d.), but that assumption can sometimes be unrealistic. Exchangeability can be seen as relaxing the i.i.d. assumption and is defined for RV pairs as:

Definition 24 (Exchangeable pairs) *An infinite sequence of RV pairs $(\mathbf{x}_e, \mathbf{y}_e)_{e\in\mathbb{N}}$ is exchangeable if for any permutation $\pi : \mathbb{N} \to \mathbb{N}$ and for any finite E*

$$p\left(\mathbf{y}_1, \ldots, \mathbf{y}_E; \mathbf{x}_1, \ldots, \mathbf{x}_E\right) = p\left(\mathbf{y}_{\pi(1)}, \ldots, \mathbf{y}_{\pi(E)}; \mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(E)}\right) \tag{66}$$

Intuitively, exchangeability means that the arguments of the distribution (as with positional arguments in Python code) can be permuted. Alternatively, this means that the pairs are i.i.d. conditioned on a distributional form.

The CdF theorem requires exchangeability, and also two conditional independence statements to hold in the underlying causal graph. To introduce these two requirements, let $\boldsymbol{ND}$ denote a node's non-descendants (*including* parents), $\overline{\boldsymbol{ND}}$ the non-descendants *excluding* parents, $\boldsymbol{Pa}$ the parents, and $\mathbf{Z}$ an arbitrary node (and, by abuse of notation, the corresponding RV). We index the components of a vector variable by i and the exchangeable tuples by e (these tuples will correspond to trajectories in HMMs, i.e., time series for environment e)[4]. There are two required conditional independence statements; the **first** is:

$$Z_{i,[e]} \perp \overline{\boldsymbol{ND}}_{i,[e]}, | \boldsymbol{Pa}_{i,[e]} \text{ where } [e] = \{1; \ldots; e\} \tag{67}$$

That is, if we condition on its parents, a RV $\mathbf{Z}$ should be independent of its non-descendants (*excluding* its parents) in all the tuples we are considering.

4. Here we follow the notation of Guo et al. (2022); in the main text, we index by the environment e as a superscript

The **second** conditional independence statement is:

$$Z_{i,[e]} \perp \boldsymbol{ND}_{i,e+1} | \boldsymbol{Pa}_{i,[e]} \text{ where } [e] = \{1; \dots; e\}, \tag{68}$$

which requires that given its parents in a set of tuples, the RV is independent of all its non-descendants (including its parents) in any other tuples.

C.2. The CdF theorem

For completeness, we state the CdF theorem for exchangeable pairs (Defn. 24).

Theorem 25 (Causal de Finetti (Guo et al., 2022)) *Given an infinite sequence of RV pairs $(\mathbf{x}_e, \mathbf{y}_e)_{e\in\mathbb{N}}$, if $(\mathbf{x}_e, \mathbf{y}_e)_{e\in\mathbb{N}}$ is infinitely exchangeable (Defn. 24) and there exists a DAG, where the following two conditions hold $\forall i \in [e], e \in \mathbb{N}$ (explained in Appx. C.1; $[e] = \{1; \dots; e\}$):*

$$Z_{i,[e]} \perp \overline{\boldsymbol{ND}}_{i,[e]}, | \boldsymbol{Pa}_{i,[e]} \tag{69}$$

$$Z_{i,[e]} \perp \boldsymbol{ND}_{i,e+1} | \boldsymbol{Pa}_{i,[e]}, \tag{70}$$

then

$$p\left(\{\mathbf{y}_1^e, \dots, \mathbf{y}_T^e; \mathbf{x}_1^e, \dots, \mathbf{x}_T^e\}_{e=1}^{|E|}\right) = \int \prod_{e=1}^{|E|} \prod_t p\left(\mathbf{y}_t^e | \mathbf{x}_t^e; \psi\right) p\left(\mathbf{x}_{t+1}^e | \boldsymbol{Pa}_i^e; \theta\right) d\mu(\theta) d\nu(\psi), \tag{71}$$

where $\psi \perp \theta$ are the CdF parameters with corresponding measures μ, ν.

C.3. The CdF conditions for HMMs

Exchangeability means that conditioned on the distributional form, the sequences are i.i.d.. For HMMs (consider Fig. 2 as an example), we define the tuples as trajectories, and show that exchangeability holds for each trajectory in a given environment (environment means that the distributional form is now endowed with parameters, i.e., for Gaussians, this would mean that we set the mean and the covariance). We call a trajectory t for a given environment e and state the exchangeability condition as (cf. (Guo et al., 2022, Fig. (a)-(b))):

$$t = (\mathbf{x}_1^e, \mathbf{y}_1^e, \dots, \mathbf{y}_T^e, \mathbf{x}_T^e) \tag{72}$$

$$p(t_1, \dots, t_N) = p\left(t_{\pi(1)}, \dots t_{\pi(N)}\right), \tag{73}$$

where t denotes the trajectory and N the number of trajectories in the environment e, $\mathbf{x}_t$ refers to the (hidden) state and $\mathbf{y}_t$ to the observed variables. Note that the above holds since each t comes from the same distribution, and the trajectories are jointly independent (knowing anything about a trajectory does not provide further information about another trajectory; cf. also the conditional independence statements below).

Now we show that the conditional independence required for the CdF theorem hold in HMMs. We start with the first condition in (67), and conclude that it is automatically satisfied in HMMs, since the parents of a variable by definition form a Markov blanket, making any variable independent from its non-descendants (excluding its parents)—cf. (7) for the conditional independencies that hold by definition in any HMMs. Regarding (68), we use Fig. 2 for a visual proof. For example, is $\mathbf{x}_2^2 \perp \mathbf{x}_1^1 | \mathbf{x}_1^2, \theta_2$? Note that each path between $\mathbf{x}_1^1$ and $\mathbf{x}_2^2$ needs to go through one of the θ_i.

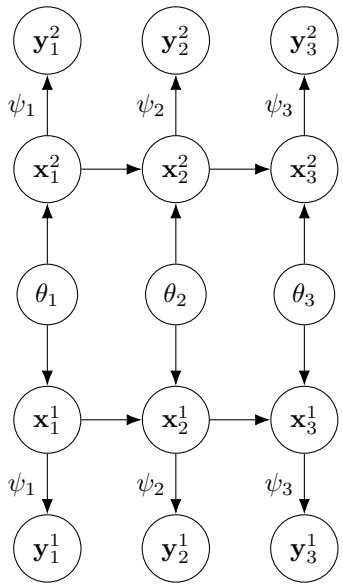


Figure 2: An example HMM for a single environment (i.e., a single set of parameters θ, ψ) and two trajectories (denoted via superscripts). θ determines the state transition probabilities $p\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)$, whereas ψ the conditional probabilities for the observations $p\left(\mathbf{y}_t|\mathbf{x}_t\right)$ (note that ψ is also the same for *both* trajectories)

1. The paths through θ_1 are blocked by conditioning on $\mathbf{x}_1^2$ (conditioning on the middle variable in the chain $\theta_1 \to \mathbf{x}_1^2 \to \mathbf{x}_2^2$ blocks the chain)
2. The paths through θ_2 are blocked by conditioning on θ_2 (conditioning on the middle variable in the collider $\mathbf{x}_2^1 \leftarrow \theta_2 \to \mathbf{x}_2^2$ blocks the collider)
3. The paths through θ_3 are blocked by **not** conditioning on $\mathbf{x}_3^1$ (not conditioning on the middle variable in the v-structure $\mathbf{x}_2^1 \to \mathbf{x}_3^1 \leftarrow \theta_3$ blocks the v-structure)

Appendix D. Practical implications

D.1. Choice of $(\boldsymbol{\sigma}_i^e)^2$

Our proof relies on that the environment variability matrix $\boldsymbol{\Delta} \in \mathbb{R}^{|E| \times d_{\mathbf{u}}}$ has column rank $d_{\mathbf{u}}$. Thus, only when the vector $\mathbf{h}$ is the zero vector will the matrix-vector product $\boldsymbol{\Delta} \cdot \mathbf{h}$ in (57) will be zero. When $(\boldsymbol{\sigma}_i^e)^2$ is not carefully chosen, the resulting $\boldsymbol{\Delta}$ might admits an (almost) zero matrix-vector product even if $\mathbf{h}$ is only to be *approximately* the zero vector. To see this, recall that a matrix's condition number w.r.t. the ℓ_2-norm is the ratio of the maximum and minimum singular values. The singular values intuitively express the scaling of a linear transformation in a specific direction; thus, if there is a non-zero component in $\mathbf{h}$ that is affected by a very small singular value, then the matrix-vector product can still be close to zero, even by violating the assumption that $\mathbf{h} = \mathbf{0}$. Thus, selecting $(\boldsymbol{\sigma}_i^e)^2$ such that it minimizes the condition number of $\boldsymbol{\Delta}$ can help avoid the above edge case. This requires $\boldsymbol{\Delta}$ to be orthogonal; thus, elucidating our maximum variability strategy (cf. results in the right-most column in Tab. 1). Note that this condition does not depend on the matrices of the LTI system. Additionally, as opposed to white-noise–based system identification (Ljung, 1998), our

proposed method only requires control signals with practically limited spectra—the Gaussian has infinite support, though most of the probability mass concentrates within three standard deviations.

D.2. Robustness to observation noise

We provide preliminary experiments on how observation noise affects identifiability, i.e., when

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \boldsymbol{\varepsilon}.$$

We use $\mathbf{C} \neq \mathbf{I}$, $\mathbf{B} \neq \mathbf{I}$ and $\boldsymbol{\mu}_{\mathbf{u}}^e \neq \mathbf{0}$. All other hyperparameters are the same as in § 4. The preliminary results in Tab. 2 suggest that our method is somewhat robust to the presence of observation noise, justifying our denoising argument in our theory. However, the setting of noisy observations needs to be more thoroughly investigated.

Table 2: Robustness of our method against observation noise. We use the minimal $(d_{\mathbf{u}} + 1)$ number of environments. Mean and standard deviation are reported across 3 runs. $d_{\mathbf{u}}$ is the dimensionality of $\mathbf{u}$ ($d_{\mathbf{x}} = d_{\mathbf{u}} = d_{\mathbf{y}}$), $|E|$ is the number of environments, $\boldsymbol{\sigma}_{\varepsilon}^2$ is the variance of the measurement noise, Mean Correlation Coefficient (MCC) measures identifiability in $[0; 1]$ (higher is better)

$d_{\mathbf{u}}$	$\lvert E\rvert$	MCC ↑				
		$\boldsymbol{\sigma}_{\varepsilon}^2 = 0$	$\boldsymbol{\sigma}_{\varepsilon}^2 = 1\text{e}{-4}$	$\boldsymbol{\sigma}_{\varepsilon}^2 = 1\text{e}{-2}$	$\boldsymbol{\sigma}_{\varepsilon}^2 = 1\text{e}{-1}$	$\boldsymbol{\sigma}_{\varepsilon}^2 = 1$
2	3	$0.627_{\pm 0.090}$	$0.679_{\pm 0.162}$	$0.643_{\pm 0.068}$	$0.590_{\pm 0.034}$	$0.625_{\pm 0.038}$
3	4	$0.886_{\pm 0.057}$	$0.823_{\pm 0.052}$	$0.718_{\pm 0.168}$	$0.637_{\pm 0.194}$	$0.796_{\pm 0.022}$