Nonparametric Factor Analysis and Beyond

Yujia Zheng[1] **Yang Liu**[2] **Jiaxiong Yao**[2] **Yingyao Hu**[†,3] **Kun Zhang**[†,1,4]
[1]Carnegie Mellon University
[2]International Monetary Fund
[3]Johns Hopkins University
[4] Mohamed bin Zayed University of Artificial Intelligence

Abstract

Nearly all identifiability results in unsupervised representation learning inspired by, e.g., independent component analysis, factor analysis, and causal representation learning, rely on assumptions of additive independent noise or noiseless regimes. In contrast, we study the more general case where noise can take arbitrary forms, depend on latent variables, and be non-invertibly entangled within a nonlinear function. We propose a general framework for identifying latent variables in the nonparametric noisy settings. We first show that, under suitable conditions, the generative model is identifiable up to certain submanifold indeterminacies even in the presence of non-negligible noise. Furthermore, under the structural or distributional variability conditions, we prove that latent variables of the general nonlinear models are identifiable up to trivial indeterminacies. Based on the proposed theoretical framework, we have also developed corresponding estimation methods and validated them in various synthetic and real-world settings. Interestingly, our estimate of the true GDP growth from alternative measurements suggests more insightful information on the economies than official reports. We expect our framework to provide new insight into how both researchers and practitioners deal with latent variables in real-world scenarios.

1 Introduction

Revealing the hidden process that generates observed data is fundamental to scientific discovery. A typical example is the so-called hidden Markov model, where a series of latent variables are observed with errors in multiple periods under conditional independence. Although machine learning can model intricate patterns in data, it frequently falls short in ensuring that its representations match the true underlying factors driving the observations (Locatello et al., 2019). Reliable identification of these latent factors is crucial for unbiased analysis across various fields, such as economics (Hu, 2008, 2017; Schennach, 2020; Hu, 2025) and psychology (Bollen, 2002; Marsh and Hau, 2007).

A typical set of approaches to identifying the hidden process underlying data generation have predominantly addressed linear relations between hidden and observed variables, providing strong theoretical backing; see, e.g., (Aigner et al., 1984; Comon, 1994; Bishop, 1998). Recent developments, for instance, nonlinear independent component analysis (ICA), have broadened this focus to capture more intricate, nonlinear relationships (Hyvärinen and Pajunen, 1999; Hyvärinen et al., 2024). These methods often introduce additional requirements, such as leveraging auxiliary variables (Hyvärinen and Morioka, 2016), utilizing time-series information (Hyvärinen and Morioka, 2017), imposing structural assumptions (Zheng et al., 2022), or specifying certain functional forms (Taleb and Jutten, 1999). Despite these advances, many models assume a noise-free environment, limiting their effectiveness in practical situations where data is inherently subject to random fluctuations.

Some works have focused on the latent variable models in noisy settings. For instance, classical factor analysis models (Reiersøl, 1950; Lawley and Maxwell, 1962; Bekker and ten Berge, 1997) can incorporate noise but remain subject to certain limitations. First, like several approaches in noisy ICA and other models (Ikeda and Toyama, 2000; Beckmann and Smith, 2004; Bon-

[†]Equal senior-authorship.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

homme and Robin, 2009; Khemakhem et al., 2020), factor analysis is constrained to handle noise that is specifically *additive* and *independent* of the latent variables. This restricts its flexibility in real-world scenarios where noise may interact with latent representation in more complex ways. Second, factor analysis relies on a fundamentally linear relationship between latent variables and observations or the model that can be reduced to a linear one, limiting its capacity to capture general nonlinearity in the generative processes. Furthermore, even when these assumptions hold, the model's identifiability is usually only guaranteed up to a linear subspace, leaving latent variables partially entangled and preventing a complete recovery of the true underlying factors. These limitations underscore the need for more general frameworks that can handle broader classes of noise and nonlinear relationships while providing stronger identifiability guarantees.

To address those concerns, we establish a general framework for identifying latent variables in nonparametric models with *nonlinear* generating processes based on the so-called Hu-Schennach Theorem, even when confronted with *non-negligible* noise. The generality of both the latent model and the noise allows us to tackle complex nonlinear transformations underlying the data, even when the generative process is *noninvertible* due to the general noise. We first show to what extent the nonparametric factor analysis model is identifiable. Moreover, unlike previous work in factor analysis, our focus extends beyond submanifold identification, demonstrating that all latent variables can be identified, thereby fully disentangling the underlying mixture of generative factors. Specifically, we show that, under standard conditions, such as structural or distributional variability, latent variables of nonparametric models are identifiable up to a permutation and component-wise invertible transformation. We also propose estimation methods to support this identification and validate our results on both synthetic and real-world datasets. Notably, we demonstrate that GDP growth estimates derived from alternative measurements, like Google search trends and nightlight intensity, offer deeper insights into economic conditions than traditional official reports.

2 Preliminary

We consider a general data-generating process as follows:

$$X = f(Z, \epsilon), \tag{1}$$

where $X = (X_1, X_2, \dots, X_m) \in \mathcal{X} \subseteq \mathbb{R}^m$ denotes observed variables, $Z = (Z_1, Z_2, \dots, Z_n) \in \mathcal{Z} \subseteq \mathbb{R}^n$ denotes latent variables, and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n) \in \mathcal{E} \subseteq \mathbb{R}^n$ denotes noise. Notably, we do *not* require independent noise and thus it is fully possible that $Z \not\perp\!\!\!\perp \epsilon$, which is different from most previous work in the literature. Moreover, the function f is generally not invertible, further extending the considered setting.

Technical Notations. We use capital letters to stand for a random variable and lower case letters to stand for the realization of a random variable. For example, $p_V(v)$ denotes the probability density function of random variable V with realization argument v, and $p_{V|U}(v|u)$ denotes the conditional density of V on U. The capital letter P denotes the distribution.

Throughout this work, for any matrix S, we use $S_{i,:}$ to indicate its i-th row and $S_{:,j}$ to indicate its j-th column. For any index set $\mathcal{I} \subset \{1, \dots, a\} \times \{1, \dots, b\}$, we define $\mathcal{I}_{i,:} := \{j \mid (i,j) \in \mathcal{I}\}$ and $\mathcal{I}_{:,j} := \{i \mid (i,j) \in \mathcal{I}\}$. Based on this, we define the support of a matrix $S \in \mathbb{R}^{a \times b}$ as $\operatorname{supp}(S) := \{(i,j) \mid S_{i,j} \neq 0\}$. Similarly, the support of a matrix-valued function $\mathbf{S}(\Theta) : \Theta \to \mathbb{R}^{a \times b}$ is defined as $\operatorname{supp}(\mathbf{S}(\Theta)) := \{(i,j) \mid \exists \theta \in \Theta, \mathbf{S}(\theta)_{i,j} \neq 0\}$. Furthermore, for any subset $\mathcal{S} \subseteq \{1, \dots, n\}$, we define its subspace $\mathbb{R}^n_{\mathcal{S}}$ as $\mathbb{R}^n_{\mathcal{S}} := \{s \in \mathbb{R}^n \mid s_i = 0, \forall i \notin \mathcal{S}\}$, where s_i is the i-th element of the vector s. All estimated quantities are denoted using the hat symbol, e.g., $\hat{Z}$ and $\hat{f}$. For ease of reference, we have included a summary of the notation in Appendix A.

3 Identifiability Theory

Suppose that the ideal data for estimating a model consists of a sample of (X, Z), but the researcher only observes X. Our objective is to identify the latent variable(s) Z under the most general conditions. We first show how to identify the latent manifold (Section 3.1), and then the identifiability of individual latent variables (Section 3.2).

3.1 Distribution Identifiability

We assume that a researcher observes the distribution of $\{X_1, X_2, \dots, X_m\}$ from a random sample. Putting the estimation of the population distribution $P_{X_1, X_2, \dots, X_m}$ from the random sample aside, we face a key identification challenge: How to determine the distribution $P_{X_1, X_2, \dots, X_m, Z}$ from the observed distribution $P_{X_1, X_2, \dots, X_m}$. We first introduce a nonparametric identification result for the hidden distribution.

Assumption 1. *The observed variables X can be split into three parts $\{X_A, X_B, X_C\}$, where variables in each part are conditionally independent of variables in other parts given Z.*

We may consider the observables $(X_1, X_2, \dots, X_m)$ as measurements of Z. Assumption 1 implies the conditional independence structure, which is commonly observed in many real-world scenarios. For instance,

symptoms such as fever, cough, and muscle aches may exhibit dependencies but are conditionally independent given the latent cause, such as influenza. Here we leverage (Hu and Schennach, 2008) to show the uniqueness of $p(X_1, X_2, \ldots, X_m, Z)$. We make the following assumption.

Assumption 2. *The joint distribution of $(X_1, X_2, \ldots, X_m, Z)$ admits a bounded density with respect to the product measure of some dominating measure defined on their supports. All marginal and conditional densities are also bounded.*

Assumption 2 requires the densities to be bounded because the decomposition of linear operators is well-established for bounded linear operators. For unbounded linear operators, the uniqueness of the decomposition is quite challenging, which previous works did not explore. Nevertheless, the support of the densities can be the whole real line, i.e., unbounded. Note that it is possible to transform an unbounded density over a bounded support to a bounded density over an unbounded support, so it can be extended to some cases where densities are unbounded.

Before introducing more assumptions, we define an integral operator corresponding to $p_{X_A|Z}$, which maps p_Z over support $\mathcal{Z}$ to p_{X_A} over support $\mathcal{X}_A$. Suppose that we know both p_Z and p_{X_A} are bounded and integrable. We define $\mathcal{L}^1_{bnd}(\mathcal{Z})$ as the set of bounded and integrable functions defined on $\mathcal{Z}$, i.e.,

$$\begin{aligned}&\mathcal{L}^1_{bnd}(\mathcal{Z})\\ =\ &\left\{g : \int_{\mathcal{Z}} |g(z)|\, dz < \infty\ ,\ \sup_{z\in\mathcal{Z}} |g(z)| < \infty\right\}.\end{aligned}$$

The linear operator can be defined as

$$\begin{aligned} L_{X_A|Z} \quad &: \quad \mathcal{L}^1_{bnd}(\mathcal{Z}) \to \mathcal{L}^1_{bnd}(\mathcal{X}_A) \qquad (2)\\ \left(L_{X_A|Z} h\right)(x) \quad &= \quad \int_{\mathcal{Z}} p_{X_A|Z}(x|Z) h(Z) dZ. \end{aligned}$$

In order to identify the unknown distributions, we need the observables to be informative so that the following assumptions hold.

Assumption 3. *The operators $L_{X_A|Z}$ and $L_{X_B|X_A}$ are injective.*[1]

Assumption 3 intuitively introduces sufficient variation in the densities, which is a mild and common condition in the nonparametric identification literature. It is also equivalent to the completeness of the density over a certain functional space (Mattner, 1993).

Assumption 4. *For all $\overline{z} \neq \widetilde{z}$ in $\mathcal{Z}$, the set $\left\{x_C : p_{X_C|Z}(x_C|\overline{z}) \neq p_{X_C|Z}(x_C|\widetilde{z})\right\}$ has positive probability.*

[1] $L_{X_B|X_A}$ is defined in the same way as $L_{X_A|Z}$ in Eq. (2).

Assumption 4 is a generally mild condition to ensure each possible value of the latent variable affects the distribution of observed variables. It is only violated when the conditional distribution is identical for two distinct values of the conditioning variable.

Assumption 5. *There exists a known functional M such that $M\left[p_{X_A|Z}(\cdot|Z)\right] = Z$ for all $Z \in \mathcal{Z}$.*

The functional M may be the mean, mode, medium, an arbitrary quantile of the probability measure $p_{X_A|Z}(\cdot|Z)$, or any other properties. The identification result may be summarized as follows:

Theorem 1. *(Hu and Schennach, 2008) Under assumptions 1, 2, 3, 4, and 5, the joint distribution $p_{X_1,X_2,\ldots,X_m}$ uniquely determines the joint distribution $p_{X_1,X_2,\ldots,X_m,Z}$, which satisfies*

$$p_{X_1,X_2,\ldots,X_m,Z} = p_{X_A|Z} p_{X_B|Z} p_{X_C|Z} p_Z. \qquad (3)$$

This identification result implies that if we have three sets of qualified measurements X_A, X_B and X_C, we are able to provide a consistent estimator of $p_{X_1,X_2,\ldots,X_m,Z}$, or $p_{Z\,|\,X_1,X_2,\ldots,X_m}$, from a sample of $(X_1, X_2, \ldots, X_m)$.

Notably, Assumption 5 plays a role in determining the order of values. Without this assumption, we can only identify the corresponding submanifold rather than the full distribution. However, as we will demonstrate later, for the purpose of identifying latent variables, recovering the submanifold is sufficient. Therefore, Assumption 5 is not essential for the broader framework.

3.2 Identifiability of Latent Variables

In the previous section, we know that the identifiability of distribution can be guaranteed in a nonparametric manner. However, identifying the distribution is often not sufficient for many practical applications. In many scenarios, we need to recover the individual latent components to gain deeper insights into the underlying processes. For instance, in understanding complex systems like economic indicators or biological mechanisms, it is crucial not only to know the latent distribution but also to disentangle the specific factors leading to the observations. This component-wise identification allows for a more granular understanding, enabling targeted interventions, improved interpretations, and more precise inference. Therefore, it becomes necessary to go beyond distributional identifiability and focus on recovering the individual latent components.

We first propose the following theorem for identifying the submanifold of the latent variables.

Theorem 2. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2, under*

Assumptions 1, 2, 3, and 4, there exists an invertible function h such that

$$p(x;\theta) = p(x;\hat{\theta}) \implies \hat{Z} = h(Z).$$

The proof is in Appendix B.1. Different from Theorem 1, we remove Assumption 5 to minimize the reliance on prior knowledge. As a trade-off, for now, we can only identify $\mathbf{Z}$ up to a submanifold instead of pinning down the whole distribution. But we will soon show that, under appropriate conditions, the submanifold identification leads to the desired component-wise identification.

We also have the following results to ensure that the dimension of the recovered unobserved variable cannot be further reduced under continuity, with their proofs in Appendix B.2 and B.3.

Lemma 1. *A one-to-one function $g : \mathbb{R}^n \to \mathbb{R}$ cannot be continuous. Therefore, the dimensionality of $\mathbb{R}^n$ cannot be reduced to that of $\mathbb{R}$ under the continuity.*

Lemma 2. *A one-to-one function $g : \mathbb{R}^n \to \mathbb{R}^k$ with $k < n$ cannot be continuous. Therefore, the dimensionality of $\mathbb{R}^n$ cannot be reduced to that of $\mathbb{R}^k$ under the continuity.*

These results have important implications for our identification strategy. They imply that the unobserved variable Z must retain its intrinsic dimensionality when being recovered from observed data under continuity assumptions. In other words, we cannot hope to represent a higher-dimensional unobserved variable using a lower-dimensional observed counterpart without losing continuity or injectivity. This reinforces the necessity of our approach in maintaining the dimensionality of Z during estimation to ensure accurate and unique recovery of its realizations, which is especially essential for cases with multivariate latent variables.

Based on these, we are now ready to move beyond the submanifold identifiability to the component identifiability of the latent variables, under appropriate conditions on the connective structure between Z and X, or the changeability of the latent distribution. For simplicity, we denote the support of the Jacobian J_f as $\mathcal{F}$, i.e., $\mathcal{F} = \text{supp}(J_f)$, where J_f is the derivative of f with respect to Z. Additionally, let $\mathcal{T}$ represent the set of matrices with the same support as $\mathbf{T}$ in the equation $J_{\hat{f}} = \mathbf{T} J_f$, where $\mathbf{T}$ is a matrix-valued function.

Theorem 3. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2. In addition to the assumptions in Theorem 2, suppose $|\hat{\mathcal{F}}| \leq |\mathcal{F}|$ and the following assumptions hold:*

i. The density p_Z is positive and smooth.

ii. For each $i \in \{1, \dots, n\}$, there exists a set of points $\{z^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ and a matrix $\mathrm{T} \in \mathcal{T}$ s.t. $\text{span}\{J_f(z^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^n_{\mathcal{F}_{i,:}}$ and $\left[J_f(z^{(\ell)})\mathrm{T}\right]_{i,:} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}$.

iii. (Structural Variability) For each $k \in \{1, \dots, n\}$, there exists $\mathcal{C}_k$ s.t. $\bigcap_{i \in \mathcal{C}_k} \mathcal{F}_{i,:} = \{k\}$.

Then there exists a component-wise invertible function h and a permutation π such that, $\forall i \in \{1, \dots, n\}$,

$$p(x;\theta) = p(x;\hat{\theta}) \implies \hat{z}_i = h_i(\pi(z_i)).$$

The proof is in Appendix B.4. The structural variability (Assumption iii) has been introduced in (Zheng et al., 2022). However, the identifiability results in (Zheng et al., 2022) are limited to deterministic transformations, thus requiring the generative process to be a diffeomorphism without any noise. In Theorem 3, we prove that, even in the presence of non-negligible noise, we can still identify the latent variables up to the same component-wise indeterminacy as that in the previous results.

Intuitively, Assumption ii avoids some unlikely cases where the samples are all from a very small population that spans only a degenerate submanifold. Therefore, it is always almost satisfied in the considered asymptotic case. Assumption iii requires sufficient structural diversity on the connective structure between latent and observed variables. For example, in a biomedical context, consider latent variables representing different underlying health conditions or genetic factors, with observed variables such as blood pressure, cholesterol levels, and glucose levels. It is unlikely that each health condition would impact exactly the same set of clinical measurements. A latent condition related to cardiovascular health might influence blood pressure and heart rate, while another condition related to metabolic health might affect glucose and cholesterol levels. Even if some overlap exists, the pattern of which latent variables affect which observed variables differs, ensuring distinct dependency structures.

Moreover, since the structural variability assumption only requires a subset of observed variables to meet the condition, it is generally satisfied when the number of observed variables exceeds the number of latent variables, as shown in (Zheng and Zhang, 2023). When the number of observed variables is greater than that of the latent variables, even if the current structure does not initially meet the condition, additional measurements can be introduced to achieve the required variability. Therefore, this provides a practical way to manually ensure that the assumption is met, which is particularly valuable given that real-world generative

processes are often unknown, and most identifiability conditions in the literature are not directly testable.

It might be worth noting that, the structural sparsity assumption implicitly requires that the latent variables are independent of one another. This aligns with recent research in nonlinear ICA, which assumes independence among latent variables along with additional conditions to achieve identifiability. Following the spirit of the seminal foundations laid by previous work leveraging auxiliary information (Hyvärinen et al., 2019; Wang et al., 2020; Lachapelle et al., 2022), we introduce distributional changes that further guarantee component-wise identifiability of latent variables. Specifically, we assume that the change stems from an auxiliary variable U, which could be the domain index or time steps. In line with the approach of achieving identifiability through sparsity and most prior work on component-wise identifiability (e.g., nonlinear ICA), we assume that the latent variables are conditionally independent given U, i.e., $p(Z) = \prod_{i=1}^{n} p(Z_i|U)$. The identifiability is as follows.

Theorem 4. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2. In addition to the assumptions in Theorem 2, suppose the following assumptions hold:*

i. The density p_Z is positive and smooth.

ii. (Distributional Variability) There exist $2n+1$ values of U, i.e., $U^{(i)}$ with $i \in \{0, 1, \ldots, 2n\}$, s.t. the $2n$ vectors $\mathbf{w}(Z, U^{(i)}) - \mathbf{w}(Z, U^{(0)})$ with $i \in \{1, \ldots, 2n\}$ are linearly independent, where vector $\mathbf{w}(Z, U)$ is defined as follows:

$$\mathbf{w}(Z, U^{(i)}) = \left(\mathbf{v}(Z, U^{(i)}), \mathbf{v}'(Z, U^{(i)})\right),$$

where

$$\mathbf{v}(Z, U^{(i)}) = \left(\frac{\partial \log p(z_1|U^{(i)})}{\partial z_1}, \cdots, \frac{\partial \log p(z_n|U^{(i)})}{\partial z_n}\right),$$
$$\mathbf{v}'(Z, U^{(i)}) = \left(\frac{\partial^2 \log p(z_1|U^{(i)})}{(\partial z_1)^2}, \cdots, \frac{\partial^2 \log p(z_n|U^{(i)})}{(\partial z_n)^2}\right).$$

Then there exists a component-wise invertible function h and a permutation π such that, $\forall i \in \{1, \ldots, n\}$,

$$p(x; \theta) = p(x; \hat{\theta}) \implies \hat{z}_i = h_i(\pi(z_i)).$$

The proof is in Appendix B.5. As discussed, the distributional variability (Assumption ii in Thm. 4) has been widely leveraged in the literature of identifiable latent variable models. Intuitively, it indicates that the auxiliary variable U should have a sufficiently diverse impact on latent variables, which is usually satisfied unless the changing mechanism is almost invariant. For instance, the assumption could imply that the latent variables should evolve over time rather than remain constant. This could manifest as gradual changes in economic indicators, shifts in user behavior across different time periods, or evolving trends in datasets collected over time. Such temporal variation ensures that the underlying structure of the latent variables is exposed, facilitating their identification through changes in their distribution.

Therefore, we conclude that under conditions of fundamentally different forms of diversity—whether structural or distributional—all latent variables can be identified component-wise, enabling the complete disentanglement of the latent generative factors. Different from existing theories, our framework is based on one of the most general settings, where we consider *nonparametric*, *noninvertible* generating function, in the presence of *general noise*.

4 Estimation

In this section, we propose two methods for estimating the unobserved variable using observational data. The first method utilizes the KL divergence between two constructed distributions, employing a kernel-based density estimator, and is efficient for univariate unobservables. The second method employs a regularized variational autoencoder (VAE), designed to handle multivariate cases in the general scenarios.

4.1 Divergence-based Estimator

The discussion above implies that we can measure the dissimilarity between a general joint distribution $p_{X_1, X_2, \ldots, X_m, Z}$ and a distribution satisfying conditional independence $p_{ci} = p_{X_1|Z} p_{X_2|Z} \cdots p_{X_m|Z} p_Z$ in order to search for latent draws Z_i. One of the choices is the Kullback–Leibler divergence

$$D_{KL}\left(p(x)||p_{ci}(x)\right) = \int p(x) \ln\left(\frac{p(x)}{p_{ci}(x)}\right) dx.$$

It is worth noting that our theory requires as few as three conditionally independent groups, and the estimation methods can be easily modified to incorporate this if the grouping is known as a prior. We build a divergence-based estimator, called Generative Element Extraction Networks (GEEN), to generate the latent realizations of Z_i satisfying the conditional independence. Let $\vec{V}$ stand for the vector of draws of variable V in the sample, i.e., $\vec{Z} = (Z^{(1)}, Z^{(2)}, ..., Z^{(N)})^T$ and $\vec{X}_j = (X_j^{(1)}, X_j^{(2)}, ..., X_j^{(N)})^T$. We generate $\vec{\hat{Z}}$ as follows:

$$\vec{\hat{Z}} = G(\vec{X}_1, \vec{X}_2, ..., \vec{X}_m).$$

with $\vec{\hat{Z}} = (\hat{Z}^{(1)}, \hat{Z}^{(2)}, ..., \hat{Z}^{(N)})^T$. The neural network G is trained to minimize the divergence

$$\min_G D\left(\hat{p}\,,\, \hat{p}_{ci}\right) \quad s.t. \int x\hat{p}_{X_1|\hat{Z}}(x|z)dx = z$$

with

$$\hat{p} = \hat{p}_{X_1,X_2,...,X_m,\hat{Z}},$$
$$\hat{p}_{ci} = \hat{p}_{X_1|\hat{Z}}\hat{p}_{X_2|\hat{Z}}...\hat{p}_{X_m|\hat{Z}}\hat{p}_{\hat{Z}},$$

where $\hat{p}$ are empirical distribution functions based on sample $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_m, \vec{\hat{Z}})$.

Notice that G enters the loss function through $\vec{\hat{Z}} = (\hat{Z}^{(1)}, \hat{Z}^{(2)}, ..., \hat{Z}^{(N)})^T$ in density estimators. To be specific, we can have a kernel density estimator

$$\begin{aligned}&\hat{p}(x_1, \ldots, x_k, \hat{z})\\ =&\frac{1}{m}\sum_{i=1}^{m} K_{h^*}\left(\hat{z} - \hat{Z}^{(i)}\right)\prod_{j=1}^{k} K_{h_j}\left(x_j - X_j^{(i)}\right),\end{aligned}$$

where

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right),$$

and the conditional density estimator as

$$\hat{p}_{X_j|\hat{Z}}(x|\hat{z}) = \frac{\sum_{i=1}^{m} K_{h_j}\left(x - X_j^{(i)}\right) K_{h^*}\left(\hat{z} - \hat{Z}^{(i)}\right)}{\sum_{i=1}^{m} K_{h^*}\left(\hat{z} - \hat{Z}^{(i)}\right)},$$

where h stands for bandwidths, N is the total sampled observations, m is the number of points in each observation and k is the number of features.

4.2 Regularized Autoencoder

Since kernel density estimation suffers from the curse of dimensionality, in which the number of computations required increases exponentially with the number of dimensions, we propose a regularized autoencoder-based estimator to deal with multivariate cases. Unlike traditional losses, we need to incorporate the conditional independence constraints. The i-th observed variable X_i is generated as $X_i = f_i(Z, \epsilon_i)$. The log-likelihood can be transformed as follows:

$$\begin{aligned}&\log \hat{p}(X|\hat{Z}) = \sum_i \log \hat{p}(X_i|\hat{Z})\\ =&\sum_i \log\left(\frac{\hat{p}(\hat{\epsilon}_i)}{\left|\frac{\partial \hat{f}_i}{\partial \hat{\epsilon}_i}\right|}\right) = \sum_i\left(\log \hat{p}(\hat{\epsilon}^i) - \log\left|\frac{\partial \hat{f}^i}{\partial \hat{\epsilon}^i}\right|\right).\end{aligned}$$

Thus, the loss of our regularized autoencoder is defined as:

$$\mathcal{L}_{\text{RAE}} = -\log \hat{p}(X|\hat{Z}) + D_{KL}\left([\hat{p}(\hat{Z}), \hat{p}(\hat{\epsilon})]\,\|\,\mathcal{N}(0, \mathbb{I})\right),$$

where we use KL divergence to enforce the independence among components in $\hat{X}$ and $\hat{\epsilon}$, and $\mathbb{I}$ is an identity matrix.

5 Experiments

In this section, we conduct experiments on both synthetic and real-world datasets to verify our claims. Additional details are included in Appendix C.

5.1 Simulations

Basis validation. We first conduct experiments on the basic setting to evaluate the identification from observations. The samples are generated as follows:

$$X_j^{(i)} = f_j(Z^i) + \epsilon_j^{(i)} \tag{4}$$

for $j = 1, 2, ..., k$ and $i = 1, 2, ..., N$. Without loss of generality, we normalize $f_1(x) = x$ and $\mathbb{E}[\epsilon_1|Z] = 0$. We pick distributions for $(\epsilon_1, ..., \epsilon_k, X^*)$ and functions $(f_2, ..., f_k)$ to generate a sample $(X_1, ..., X_k, Z)$. In this setting, we focus on directly validating the proposed theory, and thus we start with a single latent variable. Thus, we use the divergence-based estimator GEEN (Section 4.1) and train G using the observed sample $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_k)$ to generate $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_k, \hat{Z})$. That is $\vec{\hat{Z}} = G(\vec{X}_1, \vec{X}_2, ..., \vec{X}_k)$. For the baseline case, we consider the following generating process:

$$\begin{aligned}k = 4, \quad &\epsilon_1 \sim \mathcal{N}(0, 1),\\ f_1(z) = z, \quad &\epsilon_2 \sim Beta(2, 2) - \frac{1}{2},\\ f_2(z) = \frac{1}{1+e^z}, \quad &\epsilon_3 \sim Laplace(0, 1),\\ f_3(z) = z^2, \quad &\epsilon_4 \sim Bernoulli\left(\frac{1}{2}\right),\\ X_4 = \Phi(Z/3)\cdot(-1)^{I(\epsilon_4>0.5)}, \quad &Z \sim \mathcal{N}(0, 4).\end{aligned}$$

We sample 8000 points as training points from the above distributions for Z, ϵ_1, ϵ_2, ϵ_3 and ϵ_4. Then we sample another 1000 points for validation points and 1000 points for test points. We draw 500 points from the training points with replacement 8000 times to build our training set and 1000 times from the validation/test points to build our validation/test set. In the second experiment, we let the error terms correlate with Z while keeping the rest of the setup the same as the baseline. Specifically, we use:

$$\epsilon_1 = \mathcal{N}(0, \frac{1}{4}z^2), \quad \epsilon_3 = Laplace(0, \frac{1}{2}|z|).$$

In the third experiment, we double the variance of the error terms while keeping the rest setup the same as the baseline:

$$\epsilon_1 \sim \mathcal{N}(0, 4), \epsilon_2 \sim Beta(2, 4) - \frac{1}{3}, \epsilon_3 \sim Laplace(0, 2).$$

Table 1 demonstrates the min, median and max correlations of $\vec{Z}$ and $\vec{\hat{Z}}$ in the test sample for the three

Table 1: Basis Validation for Continuous Data

Simulation Name	corr($\vec{Z}$, $\vec{\hat{Z}}$)			corr($\vec{Z}$, $\vec{X_1}$)
	min	median	max	
Baseline	0.97	0.98	0.98	0.89
Linear Error	0.93	0.96	0.97	0.89
Double Error	0.80	0.89	0.91	0.70

Table 2: Comparison of corr($\vec{Z}$, $\vec{\hat{Z}}$) between our estimator and k-means for discrete data.

Simulation Name	k-means	GEEN
Baseline	0.98	0.99
Linear Error	0.96	0.97
Double Error	0.97	0.98

experiments after running each one 25 times. It shows that the estimation is robust with randomly picked initial values of the parameters and provides a better measurement of Z than X_1. The correlation between Z and the generated Z is well above 0.9 for the baseline and the linear error case and remains strong when the variance is doubled in the third experiment.

In addition to generating continuous Z, we will also demonstrate how our method performs with discrete Z. We have the same set-up as the continuous examples except that now we sample Z from the binomial distribution. Similar to the continuous case, we also have three different settings for discrete data, i.e., baseline, linear error, and double error. Details of the data generating process are included in Appendix C.

When Z are sampled from discrete random variables, the task to identify Z is similar to an unsupervised clustering task. Therefore, we also compare our method with k-means. In order to facilitate direct comparison, when running the k-means algorithm, we set the number of clusters equal to 11 and randomly pick one point from each cluster (Cluster k is the set of points with $Z = k$) as initial points. With this setup, k-means is actually put in an advantageous position since Z is completely unkown to our estimator, but k-means is provided with limited information of the clusters (e.g. the number of clusters and initial point from each cluster). As shown in Table 2, in all three cases, our estimator has better performance than k-means when measuring the correlation between Z and $\hat{X}_*$, especially for linear error and double error cases.

Generalized validation. In the previous experiments, we have carefully validated our theoretical claims in various settings. Now we would like to explore the estimation of latent variables in the general settings, where there are multiple latent variables. Thus, we conduct experiments on the regularized-autoencoder-based estimator (Section 4.2). For each latent variable $Z_i \sim \mathcal{N}(0, 4)$, we have three observed

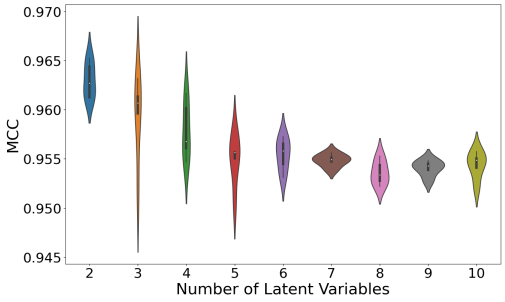

Figure 1: Results w.r.t. different numbers of latent variables.

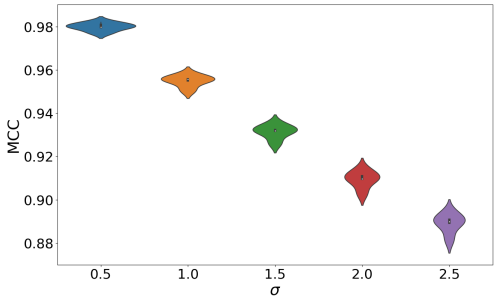

Figure 2: Results w.r.t. different standard deviations (σ) of the noise.

variables generated from Z_i by a nonlinear transformation together with a noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, satisfying the structural sparsity condition. Following previous work (Hyvärinen et al., 2024), we use mean correlation coefficient (MCC) between the true latent variables and the estimated ones as the evaluation metric. Loosely speaking, a higher MCC indicates a more disentangled recovered latent representation, which quantifies the identification quality of multiple latent variables.

We begin by conducting experiments with a noise variance of one, i.e., $\epsilon_i \sim \mathcal{N}(0, 1)$, while varying the number of latent variables from 2 to 10. The results, presented in Figure 1, show that our estimator consistently achieves MCC values close to one across datasets of different dimensions, demonstrating its effectiveness in general multivariate settings.

We further evaluate the estimator under varying noise levels by adjusting the standard deviation σ in $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The results, shown in Figure 2, reveal a slight performance decline as the noise level increases. However, even with a substantial noise variance, the method continues to achieve great identification results, highlighting its robustness and effectiveness.

5.2 Refining Official GDP Measurements

One of the important applications of our methodology is to reduce measurement errors. In this case, true

values are unobservables Z. X_1 is a direct measure of Z with the expected measurement error δ as zero. X_j $(j \neq 1)$ are indirect/direct measures of Z with unknown function forms of Z. Their error terms can be flexible and do not necessarily have zero means.

In this section, we conduct real-world experiments to study the implications of our method in practice. We apply the divergence-based estimator (GEEN) to refine GDP data using official GDP data (X_1) and alternative measures of economic activity, including satellite-recorded nighttime light (Hu and Yao, 2022) and Google Search Volume (Woloszko, 2021) as X_2 and X_3. In this experiment, true GDP (Z) are completely unknown. We demonstrate how our method can help reduce measurement errors from official GDP data.

Our sample consists of all the developing countries that have quarterly GDP data. We focus on developing countries because nighttime light data are more appropriate for tracking economic activity in those countries (Hu and Yao, 2022; Beyer et al., 2022). To account for time trends common to all countries, we remove time effects from official GDP growth rates with a fixed effect model when training and later add back the time effects to reconstruct our generated true GDP growth rates when comparing our model's performance with official data. We separate our sample into training and validation subsets, and run training 100 times and select the best model with the lowest loss in the validation sample to minimize the impact of initialization. We do not have the testing dataset, since in this case true values Z are completely unknown and the model just learns how to generate Z that can minimize the distance between the two probability densities in equation (3). Therefore, conventional testing method is not applicable here. Instead, we compare our generated GDP growth rates with official data from the macroeconomic viewpoint, which is crucial to reveal systematic differences between official data and true underlying GDP growth data.

In Figure 3, the left axis is GDP growth rate in percentage points (ppts) and the right axis marks the difference between the official GDP and our generated underlying GDP growth rates (Official - GEEN as shown in the plot). Figure 3 shows that refined GDP data reveal important patterns in official GDP data and are useful in a number of aspects. First, most countries' official GDP growth data align well with our refined estimates. For example, both Chile and South Africa have differences within 0.15 percentage points despite volatile economic growth. It suggests that GEEN would not contradict official GDP data when they are realatively accurate and could possibly improve upon them.

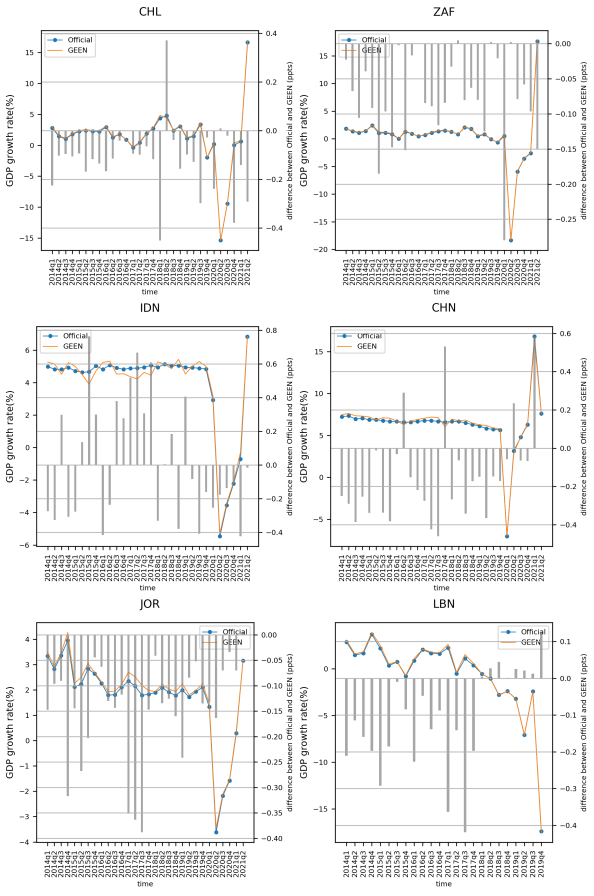

Figure 3: Country examples of official and refined GDP growth.

Second, some countries, such as China and Indonesia, have excessively smooth official GDP data compared to our refined estimates. Such excess smoothness might mask underlying dynamics and volatility of economic activity (for countries like Indonesia and China, an adjustment of 0.5 percentage points in GDP is considered significant). Estimates of underlying economic growth could therefore enrich policymakers' understanding of the state of macroeconomy, including output gap and inflationary pressures, and inform efficient policy making.

Third, the official GDP growth data of some economies systematically diverge from our refined estimates. For instance, when Lebanon's economy contracted after 2017, the official data consistently overstated its performance, whereas Jordan's official data tended to understate economic growth. A likely explanation is the presence of informal sectors not captured by official statistics. Identifying these discrepancies is a crucial first step in investigating their underlying causes, whether they relate to the statistical agency's capacity, the recording of informal economic activity, or factors within the political economy.

6 Conclusion

We have established a comprehensive framework for one of the most general settings in latent variable identification. Specifically, we prove the identifiability of latent variables in nonparametric models with nonlinear, noninvertible generating processes, even when confronted with general non-negligible noise. We show that, under standard conditions such as structural or distributional variability, latent variables in these nonlinear models can be identified up to minor ambiguities, despite the presence of complex noise. Building on the theoretical foundation, we developed estimation methods and validated them through extensive experiments on both synthetic and real-world data. The results indicate a strong correlation between the estimated and true values across various scenarios. Additionally, our analysis of real-world economic data reveals that our method can uncover more detailed insights than those provided by official reports. While our study might be limited by a lack of downstream applications across a broader range of fields, we believe our framework has the potential to transform the way researchers address latent variables in practice.

Acknowledgement

We appreciate the anonymous reviewers for their constructive feedback. We would also like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program.

References

Dennis J Aigner, Cheng Hsiao, Arie Kapteyn, and Tom Wansbeek. Latent variable models in econometrics. *Handbook of econometrics*, 2:1321–1393, 1984.

Christian F Beckmann and Stephen M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152, 2004.

Paul A Bekker and Jos MF ten Berge. Generic global indentification in factor analysis. *Linear Algebra and its Applications*, 264:255–263, 1997.

Robert Beyer, Yingyao Hu, and Jiaxiong Yao. Measuring quarterly economic growth from outer space. 2022.

Christopher M Bishop. Latent variable models. In *Learning in graphical models*, pages 371–403. Springer, 1998.

Kenneth A Bollen. Latent variables in psychology and the social sciences. *Annual review of psychology*, 53 (1):605–634, 2002.

Stéphane Bonhomme and Jean-Marc Robin. Consistent noisy independent component analysis. *Journal of Econometrics*, 149(1):12–25, 2009.

Karol Borsuk. Drei sätze über die n-dimensionale euklidische sphäre. *Fundamenta Mathematicae*, 20(1): 177–190, 1933.

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

John B Conway. *A course in functional analysis*, volume 96. Springer, 2019.

Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008.

Yingyao Hu. The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of econometrics*, 200(2):154–168, 2017.

Yingyao Hu. The econometrics of unobservables – latent variable and measurement error models and their applications, 2025. Online manuscript available at http://www.econ2.jhu.edu/people/hu/.

Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

Yingyao Hu and Jiaxiong Yao. Illuminating economic growth. *Journal of Econometrics*, 228(2):359–378, 2022.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and

structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

Shiro Ikeda and Keisuke Toyama. Independent component analysis for noisy data—meg data analysis. *Neural Networks*, 13(10):1063–1074, 2000.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. *Conference on Causal Learning and Reasoning*, 2022.

David N Lawley and Adam E Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12 (3):209–229, 1962.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

L. Lyusternik and L. Shnirel'man. Topological methods in variational problems. *Issledovaniya po Matematike i Mekhanike Omskogo Gosudarstvennogo Universiteta (OMGU)*, 1930.

Herbert W Marsh and Kit-Tai Hau. Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary educational psychology*, 32(1): 151–170, 2007.

Lutz Mattner. Some incomplete but boundedly complete location families. *The Annals of Statistics*, pages 2158–2162, 1993.

Olav Reiersøl. Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society*, pages 375–389, 1950.

Susanne M Schennach. Mismeasured and unobserved variables. In *Handbook of Econometrics*, volume 7, pages 487–565. Elsevier, 2020.

Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 5th edition, 2016.

Anisse Taleb and Christian Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

Tian-Zuo Wang, Xi-Zhu Wu, Sheng-Jun Huang, and Zhi-Hua Zhou. Cost-effectively identifying causal effects when only response variable is observable. In *International Conference on Machine Learning*, pages 10060–10069. PMLR, 2020.

Nicolas Woloszko. Tracking gdp using google trends and machine learning: A new oecd model. *Central Banking*, 12:12, 2021.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Yujia Zheng[1], Yang Liu[2], Jiaxiong Yao[2], Yingyao Hu[†,3], Kun Zhang[†,1,4]

Supplement to "Nonparametric Factor Analysis and Beyond"

Table of Contents

A Summary of Notation

This appendix summarizes the notation used throughout the paper for easy reference.

Variables

We consider the following variables:

- *Observed variables*: $X = (X_1, X_2, \ldots, X_m)$, where $X \in \mathcal{X} \subseteq \mathbb{R}^m$.
- *Latent variables*: $Z = (Z_1, Z_2, \ldots, Z_n)$, with $Z \in \mathcal{Z} \subseteq \mathbb{R}^n$.
- *Noise variables*: $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, where $\epsilon \in \mathcal{E} \subseteq \mathbb{R}^n$.

Data-Generating Process

The data are generated according to the function:

$$X = f(Z, \epsilon).$$

Note that Z and ϵ may not be independent (it is possible that $Z \not\perp\!\!\!\perp \epsilon$), and thus the function f is generally non-invertible.

Probability Notation

- Random variables are denoted by capital letters (e.g., V); realizations are denoted by lowercase letters (e.g., v).
- The probability density function (pdf) of V is denoted by $p_V(v)$.
- The conditional density of V given U is denoted by $p_{V|U}(v|u)$.
- Probability distributions are denoted by P.

Matrix Notation

For any matrix S:

- $S_{i,:}$ denotes the i-th row of S.
- $S_{:,j}$ denotes the j-th column of S.

Given an index set $\mathcal{I} \subseteq \{1, \ldots, a\} \times \{1, \ldots, b\}$, we define:

$$\mathcal{I}_{i,:} \coloneqq \{j \mid (i,j) \in \mathcal{I}\},$$
$$\mathcal{I}_{:,j} \coloneqq \{i \mid (i,j) \in \mathcal{I}\}.$$

Supports and Subspaces

- The support of a matrix $S \in \mathbb{R}^{a \times b}$ is defined as:

$$\operatorname{supp}(S) \coloneqq \{(i,j) \mid S_{i,j} \neq 0\}.$$

- The support of a matrix-valued function $\mathbf{S}(\Theta) : \Theta \to \mathbb{R}^{a \times b}$ is defined as:

$$\operatorname{supp}(\mathbf{S}(\Theta)) \coloneqq \{(i,j) \mid \exists \theta \in \Theta,\ \mathbf{S}(\theta)_{i,j} \neq 0\}.$$

- For any subset $\mathcal{S} \subseteq \{1, \ldots, n\}$, we define the subspace:

$$\mathbb{R}^n_{\mathcal{S}} \coloneqq \{s \in \mathbb{R}^n \mid s_i = 0 \text{ for all } i \notin \mathcal{S}\}.$$

Jacobians

- The Jacobian of f with respect to Z is denoted by J_f.
- The support of J_f is $\mathcal{F} = \text{supp}(J_f)$.

Sets of Matrices

We define $\mathcal{T}$ as the set of matrices that share the same support as $\mathbf{T}$ in the equation:

$$J_{\hat{f}} = J_f \mathbf{T},$$

where $\mathbf{T}$ is a matrix-valued function.

Estimated Quantities

Estimated quantities are denoted with a hat symbol, for example:

- $\hat{Z}$ denotes an estimate of Z.
- $\hat{f}$ denotes an estimate of f.

Function Spaces

The space of bounded and integrable functions on $\mathcal{Z}$ is defined as:

$$\mathcal{L}^1_{\text{bnd}}(\mathcal{Z}) \coloneqq \left\{ g : \int_{\mathcal{Z}} |g(z)| dz < \infty,\ \sup_{z \in \mathcal{Z}} |g(z)| < \infty \right\}.$$

Linear Operators

The integral operator corresponding to $p_{X_1|Z}$ is defined by:

$$L_{X_1|Z} : \mathcal{L}^1_{\text{bnd}}(\mathcal{Z}) \to \mathcal{L}^1_{\text{bnd}}(\mathcal{X}_1),$$
$$(L_{X_1|Z} h)(x) = \int_{\mathcal{Z}} p_{X_1|Z}(x|Z)\, h(Z)\, dZ.$$

Vector Functions

For vectors $\mathbf{v}(Z, U^{(i)})$ and $\mathbf{v}'(Z, U^{(i)})$, we define:

$$\mathbf{v}(Z, U^{(i)}) = \left(\frac{\partial \log p(z_1 \mid U^{(i)})}{\partial z_1}, \ldots, \frac{\partial \log p(z_n \mid U^{(i)})}{\partial z_n} \right),$$
$$\mathbf{v}'(Z, U^{(i)}) = \left(\frac{\partial^2 \log p(z_1 \mid U^{(i)})}{\partial z_1^2}, \ldots, \frac{\partial^2 \log p(z_n \mid U^{(i)})}{\partial z_n^2} \right).$$

Then, the vector $\mathbf{w}(Z, U^{(i)})$ is given by:

$$\mathbf{w}(Z, U^{(i)}) = \left(\mathbf{v}(Z, U^{(i)}),\ \mathbf{v}'(Z, U^{(i)}) \right).$$

B Proofs

B.1 Proof of Theorem 2

Theorem 2. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2, under Assumptions 1, 2, 3, and 4, there exists an invertible function h such that*

$$p(x; \theta) = p(x; \hat{\theta}) \implies \hat{Z} = h(Z).$$

Proof. According to Assumption 1, the observed variables X can be partitioned into three subsets $\{X_A, X_B, X_C\}$, where variables in each subset are conditionally independent of those in the other subsets given Z. Let us start with

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C X_A Z|X_B}(x_C, x_A, z|x_B)\, dz. \tag{5}$$

Applying the chain rule of conditional probability, we decompose the joint density in the integral:

$$p_{X_C X_A Z|X_B}(x_C, x_A, z|x_B) = p_{X_C|X_A Z X_B}(x_C|x_A, z, x_B) p_{X_A Z|X_B}(x_A, z|x_B). \tag{6}$$

Substituting this into the previous equation, we obtain:

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C|X_A Z X_B}(x_C|x_A, z, x_B) p_{X_A Z|X_B}(x_A, z|x_B)\, dz. \tag{7}$$

By the conditional independence structure, we have $X_C \perp X_B \mid (X_A, Z)$, which simplifies the first conditional density:

$$p_{X_C|X_A Z X_B}(x_C|x_A, z, x_B) = p_{X_C|X_A Z}(x_C|x_A, z). \tag{8}$$

Thus, we substitute this into the previous expression:

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C|X_A Z}(x_C|x_A, z) p_{X_A Z|X_B}(x_A, z|x_B)\, dz. \tag{9}$$

Next, we apply the decomposition of $p_{X_A Z|X_B}$. Since $X_A \perp X_B \mid Z$, we have:

$$p_{X_A Z|X_B}(x_A, z|x_B) = p_{X_A|Z X_B}(x_A|z, x_B) p_{Z|X_B}(z|x_B). \tag{10}$$

Substituting this into the integral:

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C|X_A Z}(x_C|x_A, z) p_{X_A|Z X_B}(x_A|z, x_B) p_{Z|X_B}(z|x_B)\, dz. \tag{11}$$

Under Assumption 1, $X_A \perp X_B \mid Z$ further simplifies the second conditional probability:

$$p_{X_A|Z,X_B}(x_A|z, x_B) = p_{X_A|Z}(x_A|z). \tag{12}$$

Thus, substituting this into the previous equation, we obtain:

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C|X_A Z}(x_C|x_A, z) p_{X_A|Z}(x_A|z) p_{Z|X_B}(z|x_B)\, dz. \tag{13}$$

Finally, Assumption 1 states that $X_C \perp X_A \mid Z$, which implies:

$$p_{X_C|X_A Z}(x_C|x_A, z) = p_{X_C|Z}(x_C|z). \tag{14}$$

Substituting this, we obtain the form:

$$p_{X_C X_A|X_B}(x_C, x_A|x_B) = \int p_{X_C|Z}(x_C|z) p_{X_A|Z}(x_A|z) p_{Z|X_B}(z|x_B)\, dz. \tag{15}$$

Let S and V be random variables with supports $\mathcal{S}$ and $\mathcal{V}$, respectively. A kernel operator $K_{V|S}$ is defined as a mapping from a function f' in a function space $\mathcal{F}(\mathcal{S})$ onto a function $K_{S|V} f'$ in $\mathcal{F}(\mathcal{V})$, given by:

$$(K_{V|S} f')(v) = \int p_{V|S}(v|s) f'(s)\, ds. \tag{16}$$

Similarly, let T be a random variable with support $\mathcal{T}$. A kernel operator $K_{T;V|S}$ is defined as a mapping from a function f' in a function space $\mathcal{F}(\mathcal{S})$ onto a function $K_{T;V|S} f'$ in $\mathcal{F}(\mathcal{V})$, given by:

$$(K_{T;V|S} f')(v) = \int p_{T,V|S}(t, v|s) f'(s)\, ds. \tag{17}$$

Moreover, the scaling operator $\Lambda_{V|S}$ maps the function $f'(s)$ to another function $(\Lambda_{V|S}f')(s)$ defined by the pointwise multiplication as follows:

$$(\Lambda_{V|S}f')(s) = p_{V|S}(v|s)f'(s). \tag{18}$$

Starting from the kernel operator $K_{X_C;X_A|X_B}$, defined as:

$$\left[K_{X_C;X_A|X_B}f'\right](x_A) = \int p_{X_C X_A|X_B}(x_C, x_A|x_B)f'(x_B)\,dx_B, \tag{19}$$

we substitute the decomposition of $p_{X_C X_A|X_B}$ as obtained previously:

$$\left[K_{X_C;X_A|X_B}f'\right](x_A) = \int\int p_{X_A|Z}(x_A|z)p_{X_C|Z}(x_C|z)\,p_{Z|X_B}(z|x_B)f'(x_B)\,dx_B\,dz. \tag{20}$$

Rearranging the terms, we factorize the integral as:

$$\left[K_{X_C;X_A|X_B}f'\right](x_A) = \int p_{X_A|Z}(x_A|z)\,p_{X_C|Z}(x_C|z)\,\left[K_{Z|X_B}f'\right](z)\,dz, \tag{21}$$

where the operator $K_{Z|X_B}$ is defined as:

$$\left[K_{Z|X_B}f'\right](z) = \int p_{Z|X_B}(z|x_B)f'(x_B)\,dx_B. \tag{22}$$

Substituting $\Lambda_{X_C;Z}$ into the integral, we rewrite the equation as:

$$\left[K_{X_C;X_A|X_B}f'\right](x_A) = \int p_{X_A|Z}(x_A|z)\,\left[\Lambda_{X_C;Z}K_{Z|X_B}f'\right](z)\,dz. \tag{23}$$

Finally, we apply the kernel operator $K_{X_A|Z}$, defined as:

$$\left[K_{X_A|Z}f\right](x_A) = \int p_{X_A|Z}(x_A|z)f(z)\,dz, \tag{24}$$

to yield the operator equivalence:

$$\left[K_{X_C;X_A|X_B}f'\right](x_A) = \left[K_{X_A|Z}\Lambda_{X_C;Z}K_{Z|X_B}f'\right](x_A). \tag{25}$$

This demonstrates the hierarchical decomposition of the operator $K_{X_C;X_A|X_B}$ into a composition of the kernel operators $K_{X_A|Z}$, $K_{Z|X_B}$, and the scaling operator $\Lambda_{X_C;Z}$, reflecting the conditional independence structure of the observed variables.

From Equation (25), we derive the operator equivalence:

$$K_{X_C;X_A|X_B} = K_{X_A|Z}\Lambda_{X_C;Z}K_{Z|X_B}. \tag{26}$$

This equivalence holds over the space of functions $\mathcal{G}(\mathcal{Z})$, given the factorization properties of the conditional densities established earlier.

Now, integrating over X_C, we use the fact that:

$$\int K_{X_C;X_A|X_B}f'(x_C)\,dx_C = K_{X_A|X_B}f', \tag{27}$$

and for the scaling operator:

$$\int \Lambda_{X_C;Z}f''(z)\,dx_C = f''(z), \tag{28}$$

which together imply:

$$K_{X_A|X_B} = K_{X_A|Z}K_{Z|X_B}. \tag{29}$$

For any two functions $f_1, f_2 \in L^2(\mathbb{Z})$ satisfy $K_{X_A|Z}f_1(x_A) = K_{X_A|Z}f_2(x_A)$ for all $x_A \in X_A$. Then:

$$K_{X_A|Z}(f_1 - f_2)(x_A) = \int p_{X_A|Z}(x_A|z)(f_1(z) - f_2(z))dz = 0, \forall x_A. \tag{30}$$

Assumption 3 implies that if $\alpha_1(z)$ (here $\alpha_1 = f_1 - f_2$) satisfies:

$$\int p_{X_A|Z}(x_A|z)\alpha_1(z)dz = 0 \text{ for all } x_A \in X_A, \tag{31}$$

then $\alpha_1(z) = 0$ almost everywhere in Z. Therefore, $f_1(z) - f_2(z) = 0$ almost everywhere in Z, which means $f_1 = f_2$ in $L^2(Z)$.

Thus, $K_{X_A|Z}$ is injective. Using this property, we deduce that $K_{Z|X_B}$ can be expressed as:

$$K_{Z|X_B} = K^{-1}_{X_A|Z} K_{X_A|X_B}. \tag{32}$$

The well-definedness of $K^{-1}_{X_A|Z}$ over a sufficiently large domain ensures that this operator equivalence holds consistently. Substituting the expression for $K_{Z|X_B}$ from Equation (32) into Equation (26), we arrive at:

$$K_{X_C;X_A|X_B} = K_{X_A|Z}\Lambda_{X_C;Z}K^{-1}_{X_A|Z}K_{X_A|X_B}. \tag{33}$$

The injectivity of $K_{X_A|X_B}$ can be established by examining its adjoint operator $K^\dagger_{X_A|X_B}$. Under Assumption 3, injectivity of $K_{X_B|X_A}$ implies injectivity of $K^\dagger_{X_A|X_B}$, which is the adjoint operator of $K_{X_A|X_B}$, since for any $g(\cdot)/f_{X_A}(\cdot) \in \mathcal{F}(\mathcal{X}_A)$, the condition $g \in \mathcal{F}(\mathcal{X}_A)$ holds.

We then view $K_{X_A|X_B}$ as a mapping of the closure of the range $\mathcal{R}(K^\dagger_{X_A|X_B})$ into $\mathcal{F}(\mathcal{X}_A)$. By the closed range theorem, the closure $\overline{\mathcal{R}(K^\dagger_{X_A|X_B})}$ is the orthogonal complement of the null space of $K_{X_A|X_B}$, denoted $\mathcal{N}(K_{X_A|X_B})$, and $\overline{\mathcal{R}(K_{X_A|X_B})}$ is the orthogonal complement of $\mathcal{N}(K^\dagger_{X_A|X_B})$. Therefore, $K^{-1}_{X_A|X_B}$ exists.

Because $K^\dagger_{X_A|X_B}$ is injective, we have $\mathcal{N}(K^\dagger_{X_A|X_B}) = \{0\}$. Consequently, $\overline{\mathcal{R}(K_{X_A|X_B})} = \mathcal{F}(\mathcal{X}_A)$, and the inverse $K^{-1}_{X_A|X_B}$ is well-defined and densely defined over $\mathcal{F}(\mathcal{X}_A)$.

This result allows us to write:

$$K_{X_C;X_A|X_B}K^{-1}_{X_A|X_B} = K_{X_A|Z}\Lambda_{X_C;Z}K^{-1}_{X_A|Z}.$$

The operator $K_{X_C;X_A|X_B}K^{-1}_{X_A|X_B}$ admits a spectral decomposition, where the eigenvalues are given by the diagonal elements of $\Lambda_{X_C;Z}$, i.e., $\{p_{X_C|Z}(x_C|z)\}$, and the eigenfunctions are given by the kernel of $K_{X_A|Z}$, i.e., $\{p_{X_A|Z}(x_A|z)\}$.

Finally, the identification of $p_{X_C|Z}(x_C|z)$ and $p_{X_A|Z}(\cdot|z)$ is guaranteed by the uniqueness of the spectral decomposition, which is ensured by the injectivity of $K_{X_A|X_B}$ and $K_{X_A|Z}$.

Since $K_{X_A|X_B}$ is injective, its inverse $K^{-1}_{X_A|X_B}$ exists. Substituting into Equation (33), we obtain

$$K_{X_C;X_A|X_B}\, K^{-1}_{X_A|X_B} \;=\; K_{X_A|Z}\, \Lambda_{X_C;Z}\, K^{-1}_{X_A|Z}. \tag{34}$$

Define

$$T \;\coloneqq\; K_{X_C;X_A|X_B}\, K^{-1}_{X_A|X_B}.$$

By Assumption 2, the conditional densities are bounded, so T is a bounded operator. Moreover, the structure of T implies it admits a spectral decomposition in which the eigenvalues of T are precisely the entries of $\Lambda_{X_C;Z}$, i.e., $\{p_{X_C|Z}(x_C \mid z)\}$, and the corresponding eigenfunctions are encoded in the columns of $K_{X_A|Z}$, i.e., $\{p_{X_A|Z}(x_A \mid z)\}$.

Hence, by the uniqueness of the spectral measure associated with such an operator (see e.g., (Conway, 2019, Ch. VII)), the decomposition into eigenvalues and eigenfunctions can only be realized in one way, up to standard

indeterminacies. The first indeterminacy is the scaling. Specifically, we could replace $K_{X_A|Z}$ by $\alpha K_{X_A|Z}$ and $K^{-1}_{X_A|Z}$ by $(1/\alpha) K^{-1}_{X_A|Z}$ for any nonzero α, leaving T unchanged. But because $p_{X_C|Z}$ is a conditional density satisfying

$$\int p_{X_C|Z}(x_c|z)\, dx_c = 1, \tag{35}$$

this normalizing condition forces $\alpha = 1$. Hence, no further scaling is possible.

Another indeterminacy is due to the degeneracy of eigenvalues. The diagonal operator $\Lambda_{X_C;Z}$ has eigenvalues governed by $p_{X_C|Z}$. Clearly, without additional constraints, we could have distinct Z values leading to the same eigenvalue. However, Assumption 4 avoids this by ensuring the set $\{x : p(x_C \mid z) \neq p(x_C \mid z')\}$ has positive probability for all $z, z' \in \mathcal{Z}$ with $z \neq z'$.

Even if the eigenvalues are distinct, one can permute the labeling via a bijection $h : \mathcal{Z} \to \mathcal{Z}$. Instead of indexing by Z, we could reindex by $\tilde{Z} = h(Z)$, leaving Λ unchanged but altering the labeling of $\Lambda_{X_C;Z}$. Consequently, the latent variable Z is identified up to an invertible mapping h. □

B.2 Proof of Lemma 1

Lemma 1. *A one-to-one function $g : \mathbb{R}^n \to \mathbb{R}$ cannot be continuous. Therefore, the dimensionality of $\mathbb{R}^n$ cannot be reduced to that of $\mathbb{R}$ under the continuity.*

Proof. This is a special case of the Borsuk-Ulam theorem (Lyusternik and Shnirel'man, 1930; Borsuk, 1933). We prove this by a contradiction. Suppose g is continuous. $v_i \in \mathbb{R}^n$ for $i = 1, 2, 3$ be distinct and their convex combinations be in the domain of g. Furthermore, we require that

$$v_3 \neq (1-\lambda)v_1 + \lambda v_2$$

for any $\lambda \in (0, 1)$.

Given that g is one-to-one, we have

$$g(v_1) - g(v_2) \neq 0 \tag{36}$$

Consider a function $g : [0, 1] \to \mathbb{R}$ as follows:

$$\begin{aligned} t(\lambda) &= g(v_a(\lambda)) - g(v_b(\lambda)) \\ v_a(\lambda) &= (1-\lambda)v_1 + \lambda v_2 \\ v_b(\lambda) &= [1 - \lambda(1-\lambda)][\lambda v_1 + (1-\lambda)v_2] + \lambda(1-\lambda)v_3 \end{aligned}$$

Because g is continuous, t is a continuous function with

$$\begin{aligned} t(0) &= g(v_1) - g(v_2) \neq 0 \\ t(1) &= g(v_2) - g(v_1) \neq 0 \\ t(1) &= -t(0) \end{aligned}$$

Therefore, there must exist a $\lambda_0 \in (0, 1)$ such that

$$t(\lambda_0) = 0$$

which means

$$g(v_a(\lambda_0)) = g(v_b(\lambda_0)).$$

Because $v_a(\lambda_0) \neq v_b(\lambda_0)$, this is contradictory to the assumption that g is one-to-one. Therefore, g cannot be continuous. □

B.3 Proof of Lemma 2

Lemma 2. *A one-to-one function $g : \mathbb{R}^n \to \mathbb{R}^k$ with $k < n$ cannot be continuous. Therefore, the dimensionality of $\mathbb{R}^n$ cannot be reduced to that of $\mathbb{R}^k$ under the continuity.*

Proof. The Borsuk–Ulam theorem (Lyusternik and Shnirel'man, 1930; Borsuk, 1933) states that if $g : S^k \to \mathbb{R}^k$ is continuous then there exists an $x \in S^k$ such that $g(-x) = g(x)$, where

$$S^k = \{x \in \mathbb{R}^{k+1} : ||x|| = 1\}.$$

Given that $k \leq n-1$, we show by contradiction that g cannot be continuous and one-to-one, in particular, over S^k, a subset of the domain $\mathbb{R}^n$. Suppose g is continuous. The Borsuk–Ulam theorem implies that there exists an $x \in S^k$ such that $g(-x) = g(x)$. That is contradictory to the assumption that g is one-to-one. Therefore, g cannot be continuous. □

B.4 Proof of Theorem 3

Theorem 3. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2. In addition to the assumptions in Theorem 2, suppose $|\hat{\mathcal{F}}| \leq |\mathcal{F}|$ and the following assumptions hold:*

i. The density p_Z is positive and smooth.

ii. For each $i \in \{1, \ldots, n\}$, there exists a set of points $\{z^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ and a matrix $\mathrm{T} \in \mathcal{T}$ s.t. $\operatorname{span}\{J_f(z^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^n_{\mathcal{F}_{i,:}}$ and $\left[J_f(z^{(\ell)})\mathrm{T}\right]_{i,:} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}$.

iii. (Structural Variability) For each $k \in \{1, \ldots, n\}$, there exists $\mathcal{C}_k$ s.t. $\bigcap_{i \in \mathcal{C}_k} \mathcal{F}_{i,:} = \{k\}$.

Then there exists a component-wise invertible function h and a permutation π such that, $\forall i \in \{1, \ldots, n\}$,

$$p(x; \theta) = p(x; \hat{\theta}) \implies \hat{z}_i = h_i(\pi(z_i)).$$

Proof. By Theorem 2, there exists an invertible function t such that $\hat{Z} = t(Z)$. Applying the chain rule to this transformation, we obtain the relationship between the Jacobians of $\hat{f}$ and f:

$$J_{\hat{f}} = J_f J_t. \tag{37}$$

Our goal is to show that t is a composition of a permutation and component-wise transformations, which is equivalent to proving that J_t is a generalized permutation matrix.

Let us denote J_t by $\mathbf{T}$. According to our assumption, for each index i, the set of vectors $\{J_f(z^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ spans the space $\mathbb{R}^n_{\mathcal{F}_{i,:}}$. This means any vector in $\mathbb{R}^n_{\mathcal{F}_{i,:}}$ can be expressed as a linear combination of these vectors. Specifically, for any standard basis vector e_{j_0} (where $j_0 \in \mathcal{F}_{i,:}$), there exist coefficients α_ℓ such that:

$$e_{j_0} = \sum_{\ell \in \mathcal{F}_{i,:}} \alpha_\ell J_f(z^{(\ell)})_{i,:}. \tag{38}$$

Multiplying both sides by T, we get:

$$e_{j_0}\mathrm{T} = \sum_{\ell \in \mathcal{F}_{i,:}} \alpha_\ell J_f(z^{(\ell)})_{i,:}\mathrm{T}. \tag{39}$$

Since $J_{\hat{f}} = J_f\mathbf{T}$ and based on our assumptions, each term $J_f(z^{(\ell)})_{i,:}\mathrm{T}$ lies in $\mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}$. Therefore, $e_{j_0}\mathrm{T}$ also belongs to $\mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}$, implying:

$$\mathrm{T}_{j_0,:} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}. \tag{40}$$

This holds for all $j \in \mathcal{F}_{i,:}$, so we have:

$$\forall j \in \mathcal{F}_{i,:}, \quad \mathrm{T}_{j,:} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}. \tag{41}$$

This establishes a connection between the supports:

$$\forall (i,j) \in \mathcal{F}, \quad \{i\} \times \operatorname{supp}(\mathbf{T}_{j,:}) \subset \hat{\mathcal{F}}. \tag{42}$$

A similar approach has been utilized in prior works such as (Strang, 2016; Lachapelle et al., 2022; Zheng et al., 2022). Since $J_f(z^{(\ell)})$ and $J_{\hat{f}}(\hat{z}^{(\ell)})$ both have full column rank n, the matrix $\mathbf{T}(z^{(\ell)})$ must be invertible, meaning its determinant is non-zero. Using the Leibniz formula for the determinant:

$$\det(\mathbf{T}(z^{(\ell)})) = \sum_{\sigma \in \mathcal{S}_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \mathbf{T}(z^{(\ell)})_{i,\sigma(i)} \neq 0, \tag{43}$$

where $\mathcal{S}_n$ is the set of all permutations of $\{1, \ldots, n\}$. Therefore, there exists at least one permutation σ such that:

$$\forall i \in \{1, \ldots, n\}, \quad \mathbf{T}(z^{(\ell)})_{i,\sigma(i)} \neq 0. \tag{44}$$

This implies that $\sigma(j) \in \operatorname{supp}(\mathbf{T}_{j,:})$ for all j. Combining this with Eq. (42), we obtain:

$$\forall (i,j) \in \mathcal{F}, \quad (i, \sigma(j)) \in \hat{\mathcal{F}}. \tag{45}$$

Define the permuted set:

$$\sigma(\mathcal{F}) = \{(i, \sigma(j)) \mid (i,j) \in \mathcal{F}\}. \tag{46}$$

Thus, we have:

$$\sigma(\mathcal{F}) \subset \hat{\mathcal{F}}. \tag{47}$$

Due to sparsity regularization on the estimated Jacobian, we know:

$$|\hat{\mathcal{F}}| \leq |\mathcal{F}| = |\sigma(\mathcal{F})|. \tag{48}$$

Combining this with Eq. (47), it follows that:

$$\sigma(\mathcal{F}) = \hat{\mathcal{F}}. \tag{49}$$

Suppose, for the sake of contradiction, that $\mathbf{T}(z)$ is not a composition of a diagonal matrix and a permutation matrix, i.e., there exist $j_1 \neq j_2$ such that:

$$\operatorname{supp}(\mathbf{T}_{j_1,:}) \cap \operatorname{supp}(\mathbf{T}_{j_2,:}) \neq \emptyset. \tag{50}$$

Let j_3 be an element in this intersection, so $\sigma(j_3) \in \operatorname{supp}(\mathbf{T}_{j_1,:}) \cap \operatorname{supp}(\mathbf{T}_{j_2,:})$. Without loss of generality, assume $j_3 \neq j_1$. According to Assumption iii, there exists a set $\mathcal{C}_{j_1}$ containing j_1 such that:

$$\bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{F}_{i,:} = \{j_1\}. \tag{51}$$

Since $j_3 \neq j_1$, it must be that:

$$j_3 \notin \bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{F}_{i,:}, \tag{52}$$

implying there exists some $i_3 \in \mathcal{C}_{j_1}$ such that:

$$j_3 \notin \mathcal{F}_{i_3,:}. \tag{53}$$

However, since $j_1 \in \mathcal{F}_{i_3,:}$, we have $(i_3, j_1) \in \mathcal{F}$. Using Eq. (42), we find:

$$(i_3, \sigma(j_3)) \in \hat{\mathcal{F}}. \tag{54}$$

But from Eq. (49), this means $(i_3, j_3) \in \mathcal{F}$, which contradicts Eq. (53). This contradiction implies our assumption is false, and therefore $\mathbf{T}(z)$ must be a composition of a permutation matrix and a diagonal matrix.

Together with the equation $J_{\hat{f}} = J_f \mathbf{T}$, we achieve the desired result that t is composed of a permutation and component-wise invertible functions. □

B.5 Proof of Theorem 4

Theorem 4. *Consider two models $\theta = (f, p_Z, p_\epsilon)$ and $\hat{\theta} = (\hat{f}, p_{\hat{Z}}, p_{\hat{\epsilon}})$ following the process in Section 2. In addition to the assumptions in Theorem 2, suppose the following assumptions hold:*

i. The density p_Z is positive and smooth.

ii. (Distributional Variability) There exist $2n+1$ values of U, i.e., $U^{(i)}$ with $i \in \{0, 1, \ldots, 2n\}$, s.t. the $2n$ vectors $\mathbf{w}(Z, U^{(i)}) - \mathbf{w}(Z, U^{(0)})$ with $i \in \{1, \ldots, 2n\}$ are linearly independent, where vector $\mathbf{w}(Z, U)$ is defined as follows:

$$\mathbf{w}(Z, U^{(i)}) = \left(\mathbf{v}(Z, U^{(i)}), \mathbf{v}'(Z, U^{(i)})\right),$$

where

$$\mathbf{v}(Z, U^{(i)}) = \left(\frac{\partial \log p(z_1|U^{(i)})}{\partial z_1}, \cdots, \frac{\partial \log p(z_n|U^{(i)})}{\partial z_n}\right),$$
$$\mathbf{v}'(Z, U^{(i)}) = \left(\frac{\partial^2 \log p(z_1|U^{(i)})}{(\partial z_1)^2}, \cdots, \frac{\partial^2 \log p(z_n|U^{(i)})}{(\partial z_n)^2}\right).$$

Then there exists a component-wise invertible function h and a permutation π such that, $\forall i \in \{1, \ldots, n\}$,

$$p(x; \theta) = p(x; \hat{\theta}) \implies \hat{z}_i = h_i(\pi(z_i)).$$

Proof. By Theorem 2, there exists an invertible function t such that $\hat{Z} = t(Z)$. Applying the change of variables formula for conditional densities, we have:

$$p_{\hat{Z}|U}(\hat{z}|U) = p_{Z|U}(z|U) \left|\det\left(J_{t^{-1}}(\hat{z})\right)\right|. \tag{55}$$

Taking the logarithm of both sides:

$$\log p_{\hat{Z}|U}(\hat{z}|U) = \log p_{Z|U}(z|U) + \log \left|\det\left(J_{t^{-1}}(\hat{z})\right)\right|. \tag{56}$$

Assuming that the conditional density $p_{Z|U}(z|U)$ factorizes over components, i.e.,

$$p_{Z|U}(z|U) = \prod_{i=1}^{n} p_{Z_i|U}(z_i|U), \tag{57}$$

and similarly for $p_{\hat{Z}|U}(\hat{z}|U)$. Substituting Eq. (57) into Eq. (56), we obtain:

$$\sum_{i=1}^{n} \log p_{\hat{Z}_i|U}(\hat{z}_i|U) = \sum_{i=1}^{n} \log p_{Z_i|U}(z_i|U) + \log \left|\det\left(J_{t^{-1}}(\hat{z})\right)\right|. \tag{58}$$

Next, following a common technique in the literature (Hyvärinen et al., 2024), we take the second derivatives of both sides with respect to $\hat{Z}_k$ and $\hat{Z}_v$, where $k \neq v$. Note that for $i \neq k$, we have $\partial \log p_{\hat{Z}_i|U}(\hat{z}_i|U)/\partial \hat{Z}_k = 0$. Therefore, the left-hand side simplifies to:

$$\frac{\partial^2}{\partial \hat{Z}_k \partial \hat{Z}_v} \sum_{i=1}^{n} \log p_{\hat{Z}_i|U}(\hat{z}_i|U) = 0. \tag{59}$$

For the right-hand side, we define:

$$h'_{i,(k)} := \frac{\partial Z_i}{\partial \hat{Z}_k}, \tag{60}$$

$$h''_{i,(k,v)} := \frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_v}, \tag{61}$$

$$\eta'_i(z_i, U) := \frac{\partial \log p_{Z_i|U}(z_i|U)}{\partial Z_i}, \tag{62}$$

$$\eta''_i(z_i, U) := \frac{\partial^2 \log p_{Z_i|U}(z_i|U)}{(\partial Z_i)^2}. \tag{63}$$

Then, the second derivative of the right-hand side is:

$$\sum_{i=1}^{n}\left(\eta_i''(z_i,U)\cdot h'_{i,(k)}h'_{i,(v)}+\eta_i'(z_i,U)\cdot h''_{i,(k,v)}\right)+\frac{\partial^2}{\partial\hat{z}_k\partial\hat{z}_v}\log\left|\det\left(J_{t^{-1}}(\hat{z})\right)\right|. \tag{64}$$

Setting the left-hand side and right-hand side equal and simplifying, we obtain:

$$\sum_{i=1}^{n}\left(\eta_i''(z_i,U)\cdot h'_{i,(k)}h'_{i,(v)}+\eta_i'(z_i,U)\cdot h''_{i,(k,v)}\right)+\frac{\partial^2}{\partial\hat{z}_k\partial\hat{z}_v}\log\left|\det\left(J_{t^{-1}}(\hat{z})\right)\right|=0. \tag{65}$$

Consider $2n+1$ different values of U, denoted by $U^{(i)}$ for $i\in\{0,1,\ldots,2n\}$. Evaluating Eq. (65) at these values, we obtain $2n+1$ equations. Subtracting the equation corresponding to $U^{(0)}$ from each of the other equations, we get $2n$ equations:

$$\sum_{i=1}^{n}\left(\left[\eta_i''(z_i,U^{(j)})-\eta_i''(z_i,U^{(0)})\right]h'_{i,(k)}h'_{i,(v)}+\left[\eta_i'(z_i,U^{(j)})-\eta_i'(z_i,U^{(0)})\right]h''_{i,(k,v)}\right)=0, \tag{66}$$

for $j\in\{0,1,\ldots,2n\}$.

Under Assumption ii, the $2n$ vectors formed by the differences $\mathbf{w}(Z,U^{(j)})-\mathbf{w}(Z,U^{(0)})$ are linearly independent. This implies that the only solution to the linear system in Eq. (66) is:

$$h'_{i,(k)}h'_{i,(v)}=0\quad\text{and}\quad h''_{i,(k,v)}=0,\quad\text{for all } i \text{ and } k\neq v. \tag{67}$$

This means that for each i, there is at most one index r_i such that $h'_{i,(r_i)}\neq 0$, and all mixed second derivatives $h''_{i,(k,v)}$ with $k\neq v$ are zero. Since t^{-1} is invertible, each row of the Jacobian $J_{t^{-1}}(\hat{Z})$ must have at least one non-zero entry. Therefore, there exists a permutation π such that each Z_i depends only on $\hat{Z}_{\pi(i)}$, i.e.,

$$Z_i=h_i^{-1}(\hat{Z}_{\pi(i)}). \tag{68}$$

Equivalently, we have the following equation:

$$\hat{Z}_{\pi(i)}=h_i(Z_i), \tag{69}$$

where h_i is a univariate invertible function.

Thus, $\hat{z}$ is related to z through a component-wise invertible transformation composed with a permutation. This completes the proof. □

C Supplementary Experiments

C.1 Supplementary details of the settings

Generating process for discrete data. For the baseline case, we use

$$\begin{aligned}
k=4,\quad & \epsilon_1\sim\mathcal{N}(0,1),\\
f_1(z)=z,\quad & \epsilon_2\sim Beta(2,2)-\frac{1}{2},\\
f_2(z)=\frac{1}{1+e^z},\quad & \epsilon_3\sim Laplace(0,1),\\
f_3(z)=z^2,\quad & \epsilon_4\sim Bernoulli\left(\frac{1}{2}\right),\\
X_4=\Phi(Z/3)\cdot(-1)^{I(\epsilon_4>0.5)},\ & Z\sim Binomial(10,0.5).
\end{aligned}$$

For the linear error case, we use:

$$\epsilon_1=\mathcal{N}(0,\frac{1}{4}Z^2),\epsilon_3=Laplace(0,\frac{1}{2}|Z|).$$

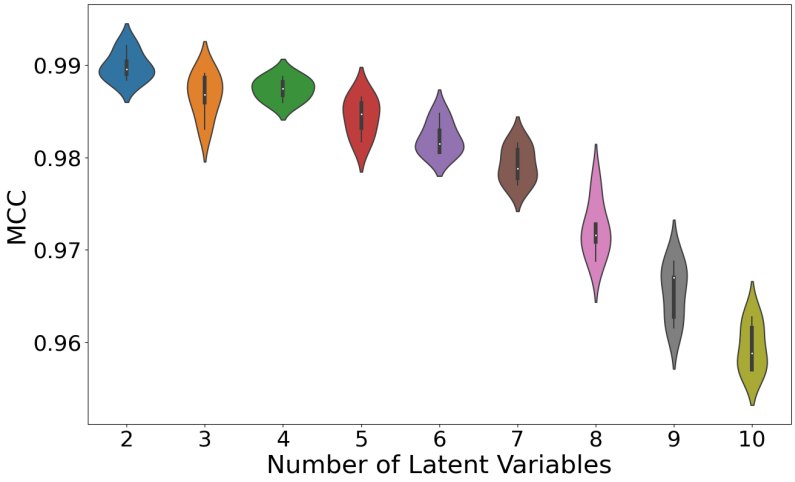

Figure 4: Results w.r.t. different numbers of latent variables for model satisfying distributional variability.

For the double error case, we use:

$$\epsilon_1 = \mathcal{N}(0, 4), \epsilon_2 = Beta(2, 4), \epsilon_3 = Laplace(0, 2).$$

Additional details for basis validation. We use a 6-layer with 10 hidden nodes fully connected neural network. The window size w and normalization term λ are tuned as hyper-parameters. In the loss function defined in Eq, (4.1), it requires more than one data point to estimate the kernel density function. As a result, unlike other use cases that one training point is enough to calculate its corresponding loss, we need to sample M (> 1) points as one observation to calculate its loss. For example, to build the training sample we sample with replacement M points from the entire training data points and repeat N times, and we end up with N observations in our training sample. We use kernel functions to approximate their density functions. The kernel function $K(\cdot)$ can simply be the standard normal density function. For the bandwidth, we adopt the so-called Silverman's rule, i.e., $h_j = w\sigma_j N^{-1/5}$ where σ_j is the standard error of X_j, and w is the window size that is determined by hyper parameters tuning. Similarly, we may take $h^* = w\hat{\sigma}N^{-1/5}$, where $\hat{\sigma}$ is the standard error of $\hat{Z}$. Theoretically, if a distribution is normal, the best choice for w used in the kernel function is 1, so to tune w we choose the range from 0.5 to 4. To tune λ, we arbitrarily choose the range from 0 to 1.

Additional details for generalized validation. For all datasets used in generalized validation, the training set consists of $10,000$ samples, and the test set consists of $2,000$ samples. For datasets satisfying structural variability, each observed variable is generated through a transformation of its dependent latent variables based on the structural condition. For datasets satisfying distributional variability, latent variables are sampled from $2n + 1$ Gaussian distributions. All experiments are repeated over 5 runs with different random seeds and are performed on 12 CPUs.

For the results w.r.t. different standard deviations of the noise (Figure. 2), we fix the number of latent variables as 5 and vary the standard deviations of the noise across $\{0.5, 1, 1.5, 2, 2.5\}$.

C.2 Supplementary experimental results

Results with distributional variability. In addition to the results with structural variability, we also validate our theory of the identifiability with distributional variability (Theorem 4). For n latent variables, we sample them from $2n + 1$ Gaussian distributions, with means uniformly drawn from $[-5, 5]$ and variances from $[0.5, 2]$. The number of latent variables ranges from 2 to 10, and the observed variables are set as three times the corresponding latent variables.

The results are shown in Figure 4. We can observe that, across all settings, our model achieves high MCC consistently. This confirms the component-wise identifiability under the condition of distributional variability.

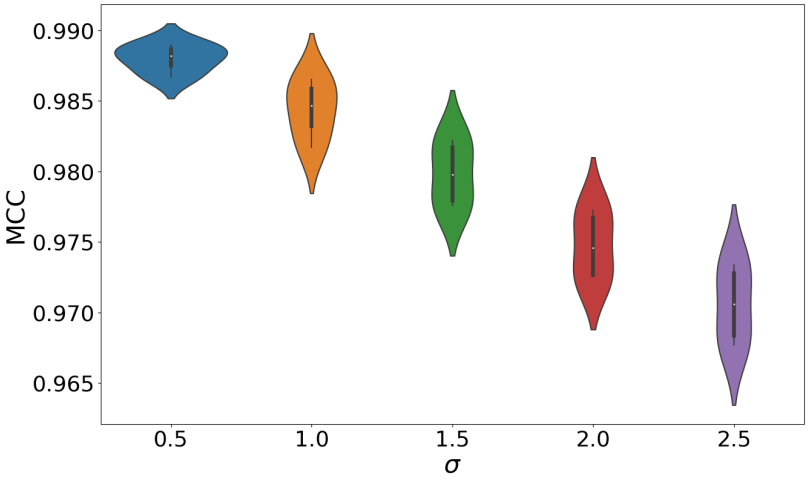

Figure 5: Results w.r.t. different standard deviations (σ) of the noise for model satisfying distributional variability. We set the number of latent variables n as 5.

In addition, we also evaluate the model with different noise levels, i.e., different standard deviations of the noise. Specifically, we fix the number of latent variables as 5 and vary the standard deviation across $\{0.5, 1, 1.5, 2, 2.5\}$. From the results (Figure 5), we observe that the quality of identification stays robust across different noise levels. This further supports the proposed identifiability theory in complicated noisy settings.