Learning Measurement Models via Subspace Identification and Clustering

Shuyu Dong, Arthur Tenenhaus, Laurent Le Brusquet, Mohammed Nabil El Korso
L2S, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
{shuyu.dong, arthur.tenenhaus, laurent.lebrusquet, mohammed.nabil.el-korso}@centralesupelec.fr

***Abstract*—Identifying latent variables and their causal relationships with the observed variables is an important yet challenging problem. Existing works may face empirical challenges such as testing-order dependency, error propagation, and selection of an appropriate significance level. The complex nature of the problem also poses difficulties in scaling up to large data. In this work, we tackle these challenges in the problem known as measurement model learning. We propose a computationally efficient two-step procedure involving low-rank subspace identification and clustering. The first step is to retrieve from the observed covariance matrix a low-rank matrix that reflects the measurement model structure. The second step conducts clustering of the observed variables given the found low-rank matrix using what we call selective flipping. Our theoretical analysis shows the optimality of the proposed method in the large sample limit. Experimental results demonstrate that the proposed method achieves improvements in both learning accuracy and computational efficiency.**

***Index Terms*—measurement models, latent variable, causal learning, clustering.**

I. Introduction

Latent variable models [1]–[4] play a fundamental role in understanding causal relationships when some unobservable or unmeasurable variables have causal impacts upon observed data. Latent variable models extend structural causal models (e.g., [5]) by incorporating hidden or latent variables that influence observed variables, and they make an important part of trending topics such as causal representation learning and reasoning [6].

Learning measurement models [2], i.e., discovering latent variables and their connectivity with the observed variables is an essential part of latent variable causal modeling. Existing works include constraint-based methods using, e.g., conditional independence constraints [7], [8], Tetrad conditions [9]–[11], or rank deficiency tests [3], and scored-based methods by optimizing, e.g., the BIC criterion [12] that subsume measurement model learning. Constraint-based methods may face empirical challenges such as testing-order dependency and error propagation, and selection of an appropriate significance level; while more recent score-based methods are limited to small-scale data due to the highly complex nature in jointly learning the measurement model and the latent causal structure, which involves optimizing both discrete (binary adjacency matrices) and continuous variables (edge weights) [12].

In this work, we tackle these challenges by leveraging rank deficiency properties but in a different way as those used in constraint-based methods. We propose a method involving low-rank subspace identification and clustering. Under the 1-factor measurement assumption [2], we show that the measurement model can be recovered from observational data, without any knowledge on the latent variables such as their number or the causal structure among them, via a simple and computationally efficient two-step procedure.

The first step is to extract a low-rank matrix from the observed covariance matrix in view of its low-rank plus diagonal structure. Specifically, both regular singular value decomposition (SVD) and heteroskedastic PCA [13] are considered for recovering the low-rank matrix that captures the interactions between the latent variables and the observed ones. We show that the low-rank matrix reveals the clustering pattern of the measurement model. Subsequently, the second step leverages the factor matrices of the found low-rank matrix to conduct clustering of the observed variables. For this purpose, a technique that we call selective flipping is proposed. We provide theoretical analysis showing the optimality of the proposed method in the large sample limit. Experimental results show improvements in both learning accuracy and computational efficiency by the proposed method.

©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The official version is available at https://ieeexplore.ieee.org/document/11073322.

II. Background

For an integer $m \geq 1$, let $[\![m]\!]$ denote the set of integers $\{1, \ldots, m\}$. For any matrix A, the i-row and j-th column of A are denoted as $A_{i,:}$ and $A_{:,j}$ respectively. For any square matrix A, let $\mathrm{Diag}(A)$ be the diagonal matrix whose diagonal is identical to that of A. Given a set of numbers $\{\omega_i\}_{i=1,\ldots,m}$, $\mathrm{Diag}(\omega_1, \ldots, \omega_m)$ denotes the diagonal matrix whose diagonal entries are ω_i's in the same order.

A. Problem Setting

Definition 1. A linear latent variable causal model (LVCM) with a directed acyclic graph (DAG) G has the observed

variables $\mathbf{X} = (X_1, \ldots, X_m)$ and latent variables $\mathbf{L} = (L_1, \ldots, L_r)$ that follow the generating process:

$$\mathbf{L} = C^{\mathrm{T}}\mathbf{L} + N_L \quad \text{and} \quad \mathbf{X} = B\mathbf{L} + N_X,$$

where N_L and N_X are jointly independent Gaussian noise terms. The DAG G is such that $L_k \to L_\ell$ is an edge if $C_{k\ell} \neq 0$ and $L_k \to X_i$ is an edge if $B_{ik} \neq 0$, with real-valued edge weights (in B and C).

The goal of this work is to identify the latent variables and their relationships with the observed variables in a LVCM. This means learning the graphical structure of B

$$G(B) := \{(i, \ell) \in [\![m]\!] \times [\![r]\!] : B_{i\ell} \neq 0\}$$

from only the observational data of $\mathbf{X}$. Note in particular that no knowledge about the LVs such as the number r or the causal structure $G(C)$ is given.

Let Σ_X and Σ_L be the population covariance matrices of the observed variables $\mathbf{X}$ and the LVs $\mathbf{L}$ respectively. Let Ω_X and Ω_L be the diagonal covariance matrices of the noise terms N_X and N_L. Then Σ_L satisfies $\Sigma_L = (I-C)^{-\mathrm{T}}\Omega_L(I-C)^{-1}$. By $\Sigma_X = B\Sigma_L B^{\mathrm{T}} + \Omega_X$, it follows that

$$\Sigma_X = B(I-C)^{-\mathrm{T}}\Omega_L(I-C)^{-1}B^{\mathrm{T}} + \Omega_X. \tag{1}$$

Remark 1 (Equivalent LVCMs [12, Lemma 1])**.** *For any parameters B, C, Ω_X, Ω_L, and Σ_X following* (1)*, there exist $\tilde{B}$ and $\tilde{C}$ (obtained through diagonal scaling) satisfying $G(\tilde{B}) = G(B)$ and $G(\tilde{C}) = G(C)$ such that*

$$\Sigma_X = \tilde{B}(I-\tilde{C})^{-\mathrm{T}}(I-\tilde{C})^{-1}\tilde{B}^{\mathrm{T}} + \Omega_X.$$

As a consequence, one can set $\Omega_L = I_r$ without loss of generality, which is the setting adopted in the rest of the paper.

The following structural assumption is used when considering LVCMs with which each observed variable has a single latent variable as parent node.

Assumption 1 (1-factor measurement [2])**.** For $\mathbf{X}$ following a LVCM (Definition 1), each observed variable has a single latent parent, and each latent variable has at least three observed variables as children.

Definition 2. Under Assumption 1, the graph $G(B)$ defines a *label function* $\phi : [\![m]\!] \to [\![r]\!]$ such that for any $i \in [\![m]\!]$: $\phi(i) = \ell$, where ℓ is the unique index satisfying $B_{i,\ell} \neq 0$. The *cluster* of the ℓ-th latent variable L_ℓ refers to its observed children, i.e., the subset of $\mathbf{X}$ indexed as follows:

$$\mathcal{C}(\ell) := \phi^{-1}(\ell) = \{i \in [\![m]\!] : \phi(i) = \ell\}. \tag{2}$$

The accuracy of an estimated clustering result with respect to the ground-truth can be measured by the Normalized Mutual Information (NMI) [14] and the Adjusted Rand Index (ARI) [15], both ranging from 0 to 1 (with 1 indicating perfect recovery of the sought clustering).

III. Proposed Method

A. Clustering via Subspace Identification

Under the common assumption that the latent variables are much fewer in number than the observed variables [2], equation (1) corresponds to a *low-rank plus diagonal decomposition* (e.g., [16]–[18]) of Σ_X, in the form of $\Sigma_X = \Sigma_\star + \Omega_X$, where Ω_X is the diagonal matrix of the noise variances and $\Sigma_\star$ is the low-rank component written as

$$\Sigma_\star = B(I-C)^{-\mathrm{T}}(I-C)^{-1}B^{\mathrm{T}}. \tag{3}$$

The low-rank component $\Sigma_\star$ is responsible for the observed covariances that are solely due to the cause-effect interactions between the latent variables and the observed ones.

Recovering $\Sigma_\star$ is possible through the SVD of Σ_X when the noise terms of the LVCM are homoskedastic, i.e., all diagonal entries in Ω_X are equal. This is because Ω_X being a multiple of the identity matrix, its impact on the sum $\Sigma_X = \Sigma_\star + \Omega_X$ can be canceled out from the singular values of Σ_X, which boils down to subtracting the noise variance from the singular values, as detailed in line 3 of Algorithm 1.

Algorithm 1 Subspace identification (homoskedastic case)

1: **Input:** covariance matrix Σ, rank parameter k
2: Compute the SVD of Σ written as $\Sigma = USU^{\mathrm{T}}$ where $S := \mathrm{Diag}(s_1, \ldots, s_m)$ for $s_1 \geq \cdots \geq s_m$.
3: Update singular values for $i \in [\![m]\!]$: $s_i \leftarrow s_i - \bar{\omega}$ where $\bar{\omega}$ denotes the median of $\{s_{k+1}, \ldots, s_m\}$.
4: **Output** the k leading singular vectors and singular values:

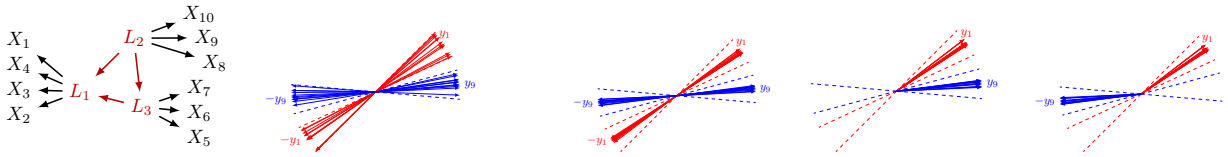
$$U_{(k)} = [U_{:,1}, \ldots, U_{:,k}],\ \tilde{S}_{(k)} = \mathrm{Diag}(s_1, \ldots, s_k)$$

We study the output $(U_{(r)}, \tilde{S}_{(r)})$ of Algorithm 1 given Σ_X (with the rank parameter $k = r$), which consists of the r leading singular vectors of Σ_X and corresponding singular values adjusted by noise variance-subtraction. The number r of latent variables is an unknown parameter of the problem, which will be addressed in Section III-B.

The following theorem shows how the subspace information of Σ_X by Algorithm 1 helps revealing the clustering pattern of the measurement model. Proofs to the results including Theorem 1 are given in Appendix A.

Theorem 1. *Suppose a LVCM (Definition 1) satisfies Assumption 1 and has homoskedastic noise terms, i.e., all diagonal entries in Ω_X are equal. Let $(U_{(r)}, \tilde{S}_{(r)})$ be the output of Algorithm 1 given the population covariance matrix Σ_X and parameter $k = r$ as input. Then $\tilde{Y} := U_{(r)}\tilde{S}_{(r)}^{\frac{1}{2}}$ satisfies: (i) for any LV $L_\ell \in \mathbf{L}$, the rows of $\tilde{Y}$ corresponding to its cluster $\mathcal{C}(\ell)$* (2)*, namely $\{\tilde{Y}_{i,:} : i \in \mathcal{C}(\ell)\}$, are confined in a unique one-dimensional (row) subspace; (ii) any pair of rows of $\tilde{Y}$ from two different clusters live in different subspaces.*

Remark 2. *Theorem 1 can be extended to the case of heteroskedastic noises if one has $\tilde{Y}$ that still satisfies $\tilde{Y}\tilde{Y}^{\mathrm{T}} = \Sigma_\star$. In such case, the r leading singular vectors are generally*



(a) A LVCM (b) (c) (d)

Fig. 1. (a) A LVCM satisfying Assumption 1. (b–d) Illustration of Theorem 2 and Algorithm 2: (b) Rows of $\tilde{Y}$ with parameter $k \neq r$; (c) Rows of $\tilde{Y}$ ($k = r$) after normalization and before selective flipping; (d) Possible outcomes of rows of $\tilde{Y}$ ($k = r$) after selective flipping.

no longer the sought subspace due to the non-homogeneous noise variances. However, HeteroPCA [13] provides a way to estimate the subspace information of $\Sigma_\star$, which we will also denote as $(U_{(r)}, \tilde{S}_{(r)})$ in Algorithm 2 to be introduced next.

Theorem 1 shows that the clustering pattern of $G(B)$ identifies with the subspace memberships of $\tilde{Y}$'s rows.

B. Algorithm

The result in the previous section shows that by inferring the subspace information of the low-rank component $\Sigma_\star$ (3), one can discover the measurement model structure of the LVCM provided that the parameter k equals the rank of $\Sigma_\star$. The difficulty is due to this unknown rank. Indeed, by Theorem 1, the rank chosen for matrix $\tilde{Y} = U_{(r)}\tilde{S}_{(r)}^{\frac{1}{2}}$ through truncated SVD should be the number of latent variables r, which is unknown. With only finite samples of $\mathbf{X}$, it is not obvious to infer accurately this optimal rank just from the spectrum of the sample covariance matrix $\widehat{\Sigma}$, since the noise terms N_X can add non-negligible perturbations to the SVD of $\widehat{\Sigma}$.

1) Selective flipping: To address this difficulty, we propose a two-step method, detailed in Algorithm 2, which can be used in a simple model selection procedure for choosing the rank k. The selection criterion is based on the sum-of-norms (SON) function that measures the accuracy of centroid-based clustering [19]–[21]. More precisely, given the outputs of Algorithm 2, namely $\tilde{Y} := [y_1, \ldots, y_m]^{\mathrm{T}}$ and the clustering's label function ϕ, define the SON objective function as

$$\Phi(\tilde{Y}, \phi) = \frac{1}{2}\sum_{i=1}^{m} \|y_i - c_\phi(y_i)\|_2^2, \tag{4}$$

where $c_\phi(y_i) := \frac{1}{|\mathcal{C}(\phi(i))|}\sum_{j\in\mathcal{C}(\phi(i))} y_j$, with $\mathcal{C}(\cdot)$ defined in (2) (Definition 2), is the centroid of the cluster containing y_i. Then we have the following result.

Theorem 2. *Under Assumption 1, and given N samples from a LVCM (Definition 1, with r latent variables), let $(\phi_{(k)}, \tilde{Y}_{(k)})$ be the output of Algorithm 2 using rank parameter k. Then we have, in the large sample limit (N tends to ∞): (i) for $k = r$, the SON value $\Phi(\tilde{Y}_{(r)}, \phi_{(r)})$ converges to zero; and (ii) $\Phi(\tilde{Y}_{(r)}, \phi_{(r)}) < \Phi(\tilde{Y}_{(k)}, \phi_{(k)})$ for all $k \neq r$.*

The key ingredient for the above result is the selective flipping operation in Algorithm 2, which maps the rows of $\tilde{Y}$ of each true cluster from different places of their subspace (due to positive and negative weights in B) onto one unit-norm vector, as illustrated in Fig. 1 (c-d). Since k-means is a centroid-based method, the rows of $\tilde{Y}_{(k)}$ (when $k = r$) after selective flipping allow k-means to recover the true clusters, which will be reflected by a close-to-zero SON value.

Algorithm 2 Proposed method

1: **Input:** Observational data $\mathcal{X}$, rank parameter k, small enough tolerance $\epsilon \in (0, 1)$
2: **Subspace identification:** get $(U_{(k)}, \tilde{S}_{(k)})$ using Algorithm 1 or alternatively HeteroPCA [13, Algorithm 1]
3: **Compute:** $\tilde{\Sigma} = U_{(k)}\tilde{S}_{(k)}{U_{(k)}}^{\mathrm{T}}$ and $\tilde{Y} = U_{(k)}\tilde{S}_{(k)}^{\frac{1}{2}}$
4: **Normalization:** for $D = \mathrm{Diag}(\tilde{\Sigma})$, update by

$$\tilde{\Sigma} \leftarrow D^{-\frac{1}{2}}\tilde{\Sigma}D^{-\frac{1}{2}}, \quad \tilde{Y} \leftarrow D^{-\frac{1}{2}}\tilde{Y}$$

5: **for** $i = 1, \ldots, m-1$ **do** *#Selective Flipping*
6: **for** $j = i+1, \ldots, m$ **do**
7: **if** $\tilde{\Sigma}_{i,j} < -(1-\epsilon)$ **then**
8: $\tilde{Y}_{j,:} \leftarrow -\tilde{Y}_{j,:}$
9: $\tilde{\Sigma}_{j,:} \leftarrow -\tilde{\Sigma}_{j,:}$ and $\tilde{\Sigma}_{:,j} \leftarrow -\tilde{\Sigma}_{:,j}$
10: **end if**
11: **end for**
12: **end for**
13: Label function $\phi \leftarrow$ Perform k-means over the rows of $\tilde{Y}$
14: **Output:** ϕ and $\tilde{Y}$

2) Model Selection: Based on Theorem 2, we consider a simple model selection procedure as follows:
(a) Determine a range of rank parameters for selection, $[k_{\min}, k_{\max}]$, depending on the singular values $\{s_i\}_{i=1,\ldots,m}$ of the sample covariance matrix $\widehat{\Sigma}$. In our experiments (with large m), we set $k_{\min} = 3$ and $k_{\max}$ as the smallest k such that the sum of the k leading s_i's attain 99% of the total sum.
(b) For all k in $[k_{\min}, k_{\max}]$, run Algorithm 2 with k; get the SON value $\Phi(\tilde{Y}_{(k)}, \phi_{(k)})$. Then finally select the output attaining the minimal SON value.

We label Algorithm 2 with the model selection procedure as LVlearn. The variant with HeteroPCA in Algorithm 2 (line 2) is labeled as LVlearn-NV (for Non-equal Variances).

IV. Experiments

We conduct experiments to evaluate the proposed method[1] in the learning of measurement models and to compare it with FOFC [11] and ClustVarLV (CLV) [22].

For the ground-truth 1-factor LVCMs, we consider $m = 100$ observed variables and $r = 10$ LVs, with edge weights in

[1]Implementation available at https://github.com/shuyu-d/lvlearning-exp.

B generated from the uniform distribution Unif($[-2,-0.5] \cup [0.5,2]$), and the graph $G(C)$ is drawn from the ER2 DAGs [23]. The noise terms in N_X follow Gaussian distributions with (i) variances equal to one in the homoskedastic case, and (ii) variances sampled uniformly from $[1,3]$ in the heteroskedastic case. The data of $\mathbf{X}$ is then generated following Definition 1 with sample size N ranging from 100, 200 and up to 3200. We evaluate estimated structures of B using the clustering scores ARI and NMI. Each learning task is run with 5 random trials. In light of Theorem 2's proof, we tune the tolerance parameter ϵ of Algorithm 2 through grid search with 100 samples, and then fix it to the chosen value (0.07).

On the above data, Fig. 2 shows evidence that LVlearn retrieves near-optimal rank choice with small enough sample sizes ($N = 100$ and 200), and starts to recover the optimal rank for 500 samples and more. The recovery scores in NMI of all chosen solutions range from 0.9 to close to 1.

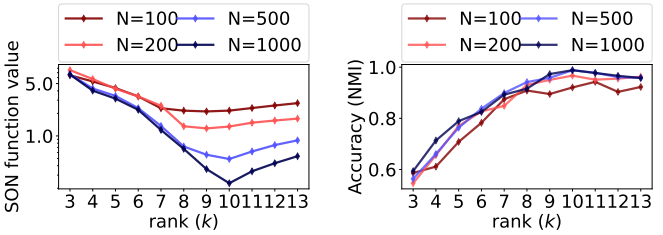
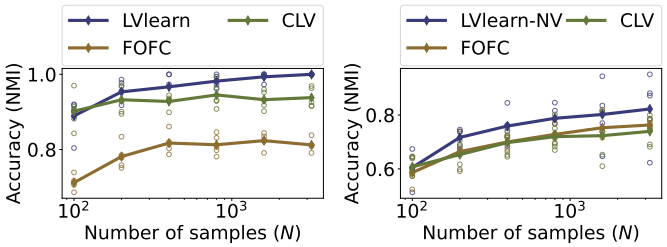


Fig. 2. Model selection with Algorithm 2 for different sample sizes from a LVCM with $r = 10$ latent variables and $m = 100$ observed ones. Left: mean SON value of different models (k); Right: mean accuracy of the models.

TABLE I
RESULTS ON DATA FROM LVCMS WITH $m = 100$ (OBSERVED) AND $r = 10$ (LATENT) WITH HOMOSKEDASTIC NOISES.

N	Algorithm	ARI	NMI	time (sec)
100	LVlearn	0.73 ± 0.09	0.89 ± 0.05	0.501
100	CLV*	0.76 ± 0.11	0.90 ± 0.05	1.428
100	FOFC	0.39 ± 0.01	0.71 ± 0.02	16.401
400	LVlearn	0.91 ± 0.10	0.97 ± 0.03	0.575
400	CLV*	0.81 ± 0.04	0.93 ± 0.02	2.123
400	FOFC	0.57 ± 0.06	0.82 ± 0.03	12.990
1600	LVlearn	0.98 ± 0.03	0.99 ± 0.01	0.632
1600	CLV*	0.81 ± 0.06	0.93 ± 0.02	4.285
1600	FOFC	0.58 ± 0.02	0.82 ± 0.02	15.205
3200	LVlearn	1.00 ± 0.00	1.00 ± 0.00	0.654
3200	CLV*	0.83 ± 0.07	0.94 ± 0.03	6.720
3200	FOFC	0.56 ± 0.05	0.81 ± 0.02	12.783

On data with homoskedastic noises (Table I and Fig. 3 (a)), LVlearn outperforms the baselines in accuracy and computation time in all settings (except for 100 samples, where comparable accuracy as CLV* is observed). Specifically, CLV* indicates CLV-$k$means [22] using the same k parameter from our LVlearn. The running time of CLV* does not include the time for selecting k while ours does. Fig. 3 (b) shows that LVlearn-NV outperforms the baselines on heteroskedastic noised data.



(a) Homoskedastic noises (b) Heteroskedastic noises

Fig. 3. Learning accuracy by different number of samples from LVCMs with $r = 10$ LVs and $m = 100$ observed variables.

V. CONCLUSION

We proposed a new method for learning measurement models involving low-rank subspace identification and clustering techniques. Under the 1-factor measurement assumption [2], we showed that, in the large sample limit, the proposed method is guaranteed to discover graphical structure of the measurement model. The proposed method showed improved computational efficiency while attaining better learning accuracy in the experiments.

APPENDIX A

Proof of Theorem 1. For $\tilde{Y} = U_{(r)}\tilde{S}_{(r)}^{\frac{1}{2}}$, we have $\tilde{Y}\tilde{Y}^{\mathrm{T}} = \Sigma_\star$ since the output of Algorithm 1 gives the exact SVD factor matrices of $\Sigma_\star$. Note that Ω_X is a multiple of the identity matrix, the last $m-r$ singular values of Σ_X equal the constant noise variance, and the subtraction operation (Algorithm 1, line 3) produces the singular values of $\Sigma_\star$. Therefore $\tilde{Y}\tilde{Y}^{\mathrm{T}} = ZZ^{\mathrm{T}}$ for $Z := B(I-C)^{-\mathrm{T}}$, where all rows in Z are strictly nonzero (as $(I-C)^{-\mathrm{T}}$ is invertible given that $G(C)$ is a DAG). Thus the angle between $\tilde{Y}_{i,:}$ and $\tilde{Y}_{j,:}$ equals that between $\tilde{Z}_{i,:}$ and $\tilde{Z}_{j,:}$, for any pair $i, j \in [\![m]\!]$. We proceed as follows.

If X_i and X_j are in one cluster $\mathcal{C}(\ell)$, i.e., $B_{i\ell} \neq 0$ and $B_{j\ell} \neq 0$ for an $\ell \in [\![r]\!]$, then their rows in $\tilde{Y}$, namely $\tilde{Y}_{i,:}$ and $\tilde{Y}_{j,:}$, belong to the same 1-dimensional row subspace, because the angle between them equals the angle between $Z_{i,:} = B_{i\ell}((I-C)^{-\mathrm{T}})_{\ell,:}$ and $Z_{j,:} = B_{j\ell}((I-C)^{-\mathrm{T}})_{\ell,:}$, which has cosine value in $\{-1, 1\}$.

Otherwise, if X_i and X_j are in different clusters, then their rows in $\tilde{Y}$ are separated by a strictly nonzero angle depending only on the causal matrix C, as it equals that between $B_{i\ell}((I-C)^{-\mathrm{T}})_{\ell,:}$ and $B_{j\ell'}((I-C)^{-\mathrm{T}})_{\ell',:}$ for $\ell \neq \ell'$, which are linearly independent ($(I-C)^{-\mathrm{T}}$ being invertible). □

Proof of Theorem 2. (i) When $k = r$, the matrices output by Algorithm 1 (or HeteroPCA [13] under heteroskedastic noises), $U_{(r)}$ and $\tilde{S}_{(r)}$, converge to the SVD factor matrices of $\Sigma_\star$ in the large sample limit. Therefore (via Theorem 1, i), the normalization and selective flipping operations (Algorithm 2, lines 4-12) map the rows of $\tilde{Y}$ of each single cluster onto one single vector (in $\mathbb{R}^r$) with unit-norm, since the flipping operates only on rows with opposite directions (with small tolerance ϵ on the angle). Rows from different clusters are always in different subspaces separated by constant angles

(Theorem 1, ii), which entails a constant range in $(0, 1)$ for the tolerance ϵ to be valid. Hence the centroid of each cluster of rows in the output $\tilde{Y}_{(r)}$ coincide with all those rows. This is an extremely favorable case where k-means recovers exactly the clusters [19]. Hence $\Phi(\tilde{Y}_{(r)}, \phi_{(r)})$ converges to zero.

(ii) Otherwise, when $k \neq r$, the mutual angles between the rows of $\tilde{Y}_{(k)}$ are no longer equal to those with $Z := B(I - C)^{-\mathrm{T}}$. This induces perturbations to the rows of $\tilde{Y}_{(k)}$ around their respective centroids, leading to SON values bounded away from zero. □

REFERENCES

[1] K. A. Bollen, *Structural equations with latent variables*. John Wiley & Sons, 1989, vol. 210.

[2] R. Silva, R. Scheines, C. Glymour, and P. Spirtes, "Learning measurement models for unobserved variables," in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 2003, pp. 543–550.

[3] F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang, "Generalized independent noise condition for estimating latent variable causal graphs," *Advances in neural information processing systems*, vol. 33, pp. 14 891–14 902, 2020.

[4] B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang, "Latent hierarchical causal structure discovery with rank constraints," *Advances in neural information processing systems*, vol. 35, pp. 5549–5561, 2022.

[5] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[6] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[7] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 499–506.

[8] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, pp. 294–321, 2012.

[9] K. A. Bollen and K.-f. Ting, "Confirmatory tetrad analysis," *Sociological methodology*, pp. 147–175, 1993.

[10] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[11] E. Kummerfeld and J. Ramsey, "Causal clustering for 1-factor measurement models," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1655–1664.

[12] I. Ng, X. Dong, H. Dai, B. Huang, P. Spirtes, and K. Zhang, "Score-based causal discovery of latent variable causal models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 37 531–37 552. [Online]. Available: https://proceedings.mlr.press/v235/ng24a.html

[13] A. R. Zhang, T. T. Cai, and Y. Wu, "Heteroskedastic pca: Algorithm, optimality, and applications," *The Annals of Statistics*, vol. 50, no. 1, pp. 53–80, 2022.

[14] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

[15] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[16] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1395–1416, 2012.

[17] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *The Annals of Statistics*, vol. 40, no. 4, pp. 1935–1967, 2012. [Online]. Available: http://www.jstor.org/stable/41806519

[18] B. Frot, P. Nandy, and M. H. Maathuis, "Robust causal structure learning with some hidden variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 81, no. 3, pp. 459–487, 2019.

[19] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[20] E. C. Chi and K. Lange, "Splitting methods for convex clustering," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 994–1013, 2015.

[21] A. Panahi, D. Dubhashi, F. D. Johansson, and C. Bhattacharyya, "Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery," in *International conference on machine learning*. PMLR, 2017, pp. 2769–2777.

[22] E. Vigneau, M. Chen, and E. M. Qannari, "Clustvarlv: an r package for the clustering of variables around latent variables." *R Journal*, vol. 7, no. 2, 2015.

[23] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.