SUPPLEMENTARY MATERIAL

A PROOFS

A.1 PROOF OF THEOREM 1

Theorem 1. *Consider a random matrix Ω following a G-Wishart distribution with graph $G = (\boldsymbol{V}, \boldsymbol{E})$ as well as parameters ν and Ψ, i.e., $\Omega \sim \mathcal{W}_G(\nu, \Psi)$. Let $\Sigma = \Omega^{-1}$ and $\widetilde{\Sigma}$ be the normalized matrix of Σ, i.e., $\widetilde{\Sigma}_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$. Then, for large ν, we have*

$$\mathrm{Var}\left[\widetilde{\Sigma}_{ij}\right] \approx \frac{(1 - (\mathrm{E}\left[\widetilde{\Sigma}_{ij}\right])^2)^2}{\nu} ,$$

for off-diagonal elements $\widetilde{\Sigma}_{ij}$ whenever $(i, j) \in \boldsymbol{E}$.

Proof. If Ω follows a G-Wishart distribution, i.e., $\Omega \sim \mathcal{W}_G(\nu, \Psi)$, and $\Sigma = \Omega^{-1}$, then we say that Σ follows a hyper inverse-Wishart distribution [21], denoted by $\Sigma \sim \mathcal{HIW}_G(\nu, \Psi)$.

Lemma 1 (see [22]). *For a graph $G = (\boldsymbol{V}, \boldsymbol{E})$, assume $\Sigma \sim \mathcal{HIW}_G(\nu, \Psi)$. Then, for any $B \subseteq \boldsymbol{V}$, we have*

$$\Sigma_{BB} \sim \mathcal{HIW}_{G_B}(\nu, \Psi_{BB}) ,$$

where G_B is the subgraph only involving variables in B.

Lemma 2 (see [9]). *If Σ follows an inverse-Wishart distribution, i.e., $\Sigma \sim \mathcal{IW}(\nu, \Psi)$, and $\widetilde{\Sigma} = (\widetilde{\Sigma}_{ij})$ with $\widetilde{\Sigma}_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, then for each off-diagonal element $\widetilde{\Sigma}_{ij}$ and large ν, we have*

$$\mathrm{Var}\left[\widetilde{\Sigma}_{ij}\right] \approx \frac{(1 - (\mathrm{E}\left[\widetilde{\Sigma}_{ij}\right])^2)^2}{\nu} .$$

Suppose $\Sigma \sim \mathcal{HIW}_G(\nu, \Psi)$ with a graph $G = (\boldsymbol{V}, \boldsymbol{E})$. According to Lemma 1, for any subset $B = \{i, j\} \subseteq \boldsymbol{V}$, we have

$$\begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \sim \mathcal{HIW}_{G_B}\left(\nu, \begin{bmatrix} \Psi_{ii} & \Psi_{ij} \\ \Psi_{ji} & \Psi_{jj} \end{bmatrix}\right) . \quad (8)$$

If there exists an edge between node i and node j in graph G, i.e., $(i, j) \in \boldsymbol{E}$, the subgraph G_B is a fully connected graph. Then, the hyper inverse-Wishart distribution in (8) reduces to an inverse-Wishart distribution, i.e.,

$$\begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \sim \mathcal{IW}\left(\nu, \begin{bmatrix} \Psi_{ii} & \Psi_{ij} \\ \Psi_{ji} & \Psi_{jj} \end{bmatrix}\right) ,$$

when $(i, j) \in \boldsymbol{E}$.

Let $\widetilde{\Sigma}_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, then according to Lemma 2, for large ν, we have

$$\mathrm{Var}\left[\widetilde{\Sigma}_{ij}\right] \approx \frac{(1 - (\mathrm{E}\left[\widetilde{\Sigma}_{ij}\right])^2)^2}{\nu} ,$$

whenever $(i, j) \in \boldsymbol{E}$ in graph G. □

A.2 PROOF OF THEOREM 2

Theorem 2 (Consistency of the CFPC algorithm). *Let $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ be independent observations drawn from a Gaussian copula factor model. If 1) the measurement model per factor is known and pure; and 2) the distribution over factors is faithful to a DAG $\mathcal{G}$, then*

$$\lim_{n\to\infty} P\big(\hat{\mathcal{M}}_n(\mathcal{G}) = \mathcal{M}(\mathcal{G})\big) = 1 ,$$

where $\hat{\mathcal{M}}_n(\mathcal{G})$ is the output of the CFPC algorithm and $\mathcal{M}(\mathcal{G})$ is the Markov equivalent class of the true underlying DAG $\mathcal{G}$.

Proof. If $S = \Lambda C \Lambda^T + D$ is the response vector's covariance matrix, then its correlation matrix is $\widetilde{S} = V^{-\frac{1}{2}} S V^{-\frac{1}{2}} = V^{-\frac{1}{2}} \Lambda C \Lambda^T V^{-\frac{1}{2}} + V^{-\frac{1}{2}} D V^{-\frac{1}{2}} = \widetilde{\Lambda} C \widetilde{\Lambda}^T + \widetilde{D}$, where V is a diagonal matrix containing the diagonal entries of S. We make use of Theorem 1 from [18] to show the consistency of $\widetilde{S}$. Our factor-analytic prior puts positive probability density almost everywhere on the set of correlation matrices that have a k-factor decomposition. Then, by applying Theorem 1 in [18], we obtain the consistency of the posterior distribution on the response vector's correlation matrix:

$$\lim_{n\to\infty} \Pi\big(\widetilde{S} \in \mathcal{V}(\widetilde{S}_0) | \boldsymbol{Z}_n \in \mathcal{D}(\boldsymbol{Y}_n)\big) = 1 \; a.s. \; \forall \; \mathcal{V}(\widetilde{S}_0),$$

where $\mathcal{D}(\boldsymbol{Y}_n)$ is the space restricted by observed data, and $\mathcal{V}(\widetilde{S}_0)$ is a neighborhood of the true parameter $\widetilde{S}_0$.

From this point on, to simplify notation, we will omit adding the tilde to refer to the rescaled matrices $\widetilde{\Sigma}$, $\widetilde{S}$, $\widetilde{\Lambda}$, and $\widetilde{D}$, since scaling the covariance matrix to a correlation matrix does not change C. Thus, S from now on refers to the correlation matrix of the response vector.

The Gibbs sampler underlying the CFPC algorithm has the posterior of Σ as its stationary distribution. Σ contains S, the correlation matrix of the response random vector, in the upper left block and C in the lower right block. Here C is the correlation matrix of factors, which implicitly depends on the *Gaussian Copula Factor Model* from Definition 1 of the main paper via the formula $S = \Lambda C \Lambda^T + D$. In order to render this decomposition identifiable, we need to put constraints on C, Λ, D. Otherwise, we can always replace Λ with ΛU and C with $U^{-1} C U^{-1}$, where U is any $k \times k$ invertible matrix, to obtain the equivalent decomposition $S = (\Lambda U)(U^{-1} C U^{-T})(U^T \Lambda^T) + D$. However, we have assumed that Λ follows a particular sparsity structure in which there is only a single non-zero entry for each row. This assumption restricts the space of equivalent solutions, since any ΛU has to follow the same sparsity structure as Λ. More explicitly, ΛU maintains the

same sparsity pattern if and only if U is a diagonal matrix (Lemma 3).

By decomposing S, we get a class of solutions for C, i.e., $U^{-1}CU^{-1}$, where U can be any invertible diagonal matrix. However, we can show that all the members in this class encode the same set of conditional independencies (Lemma 4). Thus, all solutions in this class imply the same causal structure, which means that we can use any of these solutions as the input to the PC algorithm.

In order to get a unique solution for C, we impose two identifying conditions: 1) we restrict C to be a correlation matrix; 2) we force the first non-zero entry in each column of Λ to be positive. These conditions are sufficient for identifying C uniquely (Lemma 5).

Now, given the consistency of S and the unique smooth map from S to C, we obtain the consistency of the posterior mean of the parameter C. Finally, given the correct correlation matrix, the PC algorithm will output the correct Markov equivalent class [27] with high probability, that is

$$\lim_{n\to\infty} P\big(\hat{\mathcal{M}}_n(\mathcal{G}) = \mathcal{M}(\mathcal{G})\big) = 1\,.$$

□

Lemma 3. *If $\Lambda = (\lambda_{ij})$ is a $p \times k$ factor loading matrix with only a single non-zero entry for each row, then ΛU will have the same sparsity pattern if and only if $U = (u_{ij})$ is diagonal.*

Proof. ($\Rightarrow$) We prove the direct statement by contradiction. We assume that U has an off-diagonal entry that is not equal to zero. We arbitrarily choose that entry to be $u_{rs}, r, s \in \{1, 2, \ldots, k\}, r \neq s$. Due to the particular sparsity pattern we have chosen for Λ, there exists $q \in \{1, 2, \ldots, p\}$ such that $\lambda_{qr} \neq 0$ and $\lambda_{qs} = 0$, i.e., the unique factor corresponding to the response Z_q is η_r. However, we have $(\Lambda U)_{qs} = \lambda_{qr} u_{rs} \neq 0$, which means (ΛU) has a different sparsity pattern. We have reached a contradiction, therefore U is diagonal.

($\Leftarrow$) If U is diagonal, i.e., $U = \operatorname{diag}(u_1, u_2, \ldots, u_k)$, then $(\Lambda U)_{ij} = \lambda_{ij} u_j$. This means that $(\Lambda U)_{ij} = 0 \iff \lambda_{ij} u_j = 0 \iff \lambda_{ij} = 0$, so the sparsity pattern is preserved. □

Lemma 4. *Consider a random vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ that follows a multivariate normal distribution with population correlation matrix C. Then, for any invertible diagonal matrix $U = \operatorname{diag}(u_1, u_2, \ldots, u_k)$, the matrix $\widetilde{C} = UCU$ encodes the same set of conditional independencies among $\boldsymbol{\eta}$ as C.*

Proof. Let $i, j \in \{1, \ldots, k\}$, and $Q \subseteq \{1, \ldots, k\}/\{i, j\}$. In the Gaussian case, η_i is independent of η_j given $\boldsymbol{\eta}_Q$ if and only if the partial correlation between η_i and η_j given $\boldsymbol{\eta}_Q$, denoted by $\rho^C_{ij|Q}$, vanishes, i.e.,

$$\eta_i \perp\!\!\!\perp \eta_j | \boldsymbol{\eta}_Q \iff \rho^C_{ij|Q} = 0\,. \tag{9}$$

The partial correlation $\rho^C_{ij|Q}$ is uniquely determined by the correlation matrix C, that is,

$$\rho^C_{ij|Q} = -\frac{A_{ij}}{\sqrt{A_{ii}A_{jj}}}\,, \tag{10}$$

where $A = (C_{(i,j,Q)})^{-1}$ is the inverse of the principal submatrix of C over $\{i, j, Q\}$. Similarly, for the matrix $\widetilde{C}$, we have

$$\rho^{\widetilde{C}}_{ij|Q} = -\frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}\,, \tag{11}$$

where

$$\begin{aligned} B &= (\widetilde{C}_{(i,j,Q)})^{-1} \\ &= ((UCU)_{(i,j,Q)})^{-1} \\ &= (U_{(i,j,Q)} C_{(i,j,Q)} U_{(i,j,Q)})^{-1} \text{ (since } U \text{ is diagonal)} \\ &= (U_{(i,j,Q)})^{-1} (C_{(i,j,Q)})^{-1} (U_{(i,j,Q)})^{-1} \\ &= (U_{(i,j,Q)})^{-1} A (U_{(i,j,Q)})^{-1}\,. \end{aligned}$$

Since all the diagonal elements of U are non-zero, we have

$$B_{ij} = u_i^{-1} A_{ij} u_j^{-1} = 0 \iff A_{ij} = 0\,. \tag{12}$$

From Equation (10), (11), and (12), we have

$$\rho^{\widetilde{C}}_{ij|Q} = 0 \iff \rho^C_{ij|Q} = 0\,. \tag{13}$$

Therefore, according to Equation (9) and (13), $\forall i, j$, and Q, we have

$$\eta_i \perp\!\!\!\perp \eta_j | \boldsymbol{\eta}_Q \iff \rho^C_{ij|Q} = 0 \iff \rho^{\widetilde{C}}_{ij|Q} = 0\,,$$

which concludes our proof. □

Lemma 5. *Given the factor structure defined in Section 3 of the main paper, we can uniquely recover C from $S = \Lambda C \Lambda^T + D$ if we constrain it to be a correlation matrix and we force the first element in each column of Λ to be positive. If a factor η has a single response Z (reduction to a Gaussian copula model), we set $Z = \eta$.*

Proof. Here we assume that the model has the stated factor structure, i.e., that there is some Λ, C, and D such that $S = \Lambda C \Lambda^T + D$. We then show that our chosen restrictions are sufficient for identification using an argument

similar to that in [2]. The difference is that we only require C to be identified, while Λ and D may potentially still be non-identifiable in some situations.

The decomposition $S = \Lambda C \Lambda^T + D$ constitutes a system of $\frac{p(p+1)}{2}$ equations:

$$\begin{aligned} s_{ii} &= \lambda_{if(i)}^2 + d_{ii} \\ s_{ij} &= c_{f(i)f(j)} \lambda_{if(i)} \lambda_{jf(j)} \ , \ i < j \ , \end{aligned} \tag{14}$$

where $S = (s_{ij}), \Lambda = (\lambda_{ij}), C = (c_{ij}), D = (d_{ij})$, and $f : \{1, 2, \ldots, p\} \to \{1, 2, \ldots, k\}$ is the map from a response variable to its corresponding factor. Looking at the equation system in (14), we notice that each factor correlation term $c_{qr}, q \neq r$, appears only in the equations corresponding to response variables indexed by i and j such that $f(i) = q$ and $f(j) = r$ or vice versa. This suggests that we can restrict our analysis to submodels that include only two factors by considering the submatrices of S, Λ, C, D that only involve those two factors. To be more precise, the idea is to look only at the equations corresponding to the submatrix $S_{f^{-1}(q)f^{-1}(r)}$, where f^{-1} is the preimage of $\{1, 2, \ldots, k\}$ under f. Indeed, we will show that we can identify each individual correlation term corresponding to pairs of factors only by looking at these submatrices. Any information concerning the correlation term provided by the other equations is then redundant.

Let us then consider an arbitrary pair of factors in our model and the corresponding submatrices of Λ, C, D, and S. (The case of a single factor is trivial and not interesting for causal discovery.) In order to simplify notation, we will also use Λ, C, D, and S to refer to these submatrices. We also re-index the two factors involved to η_1 and η_2 for simplicity. In order to recover the correlation between a pair of factors from S, we have to analyze three separate cases to cover all the bases (see Figure 6 for examples concerning each case):

1. The two factors are not correlated, i.e., $c_{12} = 0$. (There are no restrictions on the number of response variables that the factors can have.)

2. The two factors are correlated, i.e., $c_{12} \neq 0$, and each has a single response, which implies that $Z_1 = \eta_1$ and $Z_2 = \eta_2$.

3. The two factors are correlated, i.e., $c_{12} \neq 0$, but at least one of them has at least two responses.

Case 1: If the two factors are not correlated (see example in the left panel of Figure 6), this fact will be reflected in the matrix S. More specifically, the off-diagonal blocks in S, which correspond to the covariance between the responses of one factor and the responses of the other factor, will be set to zero. If we notice this zero pattern in S, we can immediately determine that $c_{12} = 0$.

Case 2: If the two factors are correlated and each factor has a single associated response (see middle panel of Figure 6), the model reduces to a Gaussian Copula model, hence $d_{11} = d_{22} = 0$. Then, we directly get $c_{12} = s_{12}$ since we have put the constraints $Z = \eta$ if η has a single indicator Z.

Case 3: If at least one of the factors (w.l.o.g., η_1) is allowed to have more than one response (see the example in the right panel of Figure 6), we arbitrarily choose two of these responses. We also require one response variable corresponding to the other factor (η_2). We use $\lambda_{i1}, \lambda_{j1}$, and λ_{l2} to denote the loadings of these response variables, where $i, j, l \in \{1, 2, \ldots, p\}$. From (14) we have:

$$\begin{aligned} s_{ij} &= \lambda_{i1}\lambda_{j1} \\ s_{il} &= c_{12}\lambda_{i1}\lambda_{l2} \\ s_{jl} &= c_{12}\lambda_{j1}\lambda_{l2}. \end{aligned}$$

Since we are in the case in which $c_{12} \neq 0$, which automatically implies that $s_{jl} \neq 0$, we can divide the last two equations to obtain $\frac{s_{il}}{s_{jl}} = \frac{\lambda_{i1}}{\lambda_{j1}}$. We then multiply the result with the first equation to get $\frac{s_{ij}s_{il}}{s_{jl}} = \lambda_{i1}^2$. Without loss of generality, we can say that λ_{i1} is the first entry in the first column of Λ, which means that $\lambda_{i1} > 0$. This means that we have uniquely recovered λ_{i1} and λ_{j1}.

We can also assume without loss of generality that λ_{l2} is the first entry in the second column of Λ, so $\lambda_{l2} > 0$. If η_2 has at least two responses, we use a similar argument to the one before to uniquely recover λ_{l2}. We can then use the above equations to get c_{12}. If η_2 has only one response, then $d_{ll} = 0$, which means that $s_{ll} = \lambda_{l2}^2$, so again λ_{l2} is uniquely recoverable and we can obtain c_{12} from the equations above.

Thus, we have shown that we can correctly determine c_{qr} only from $S_{f^{-1}(q)f^{-1}(r)}$ in all three cases. By applying this approach to all pairs of factors, we can uniquely recover all pairwise correlations. This means that, given our constraints, we can identify a unique C from the decomposition of S.

□

B GENERALIZATION OF THE PC-MIMBUILD ALGORITHM

Given a pure and correct measurement model involving at least 2 indicators per factor, Spirtes et al. [27] proposed to test independence and conditional independence

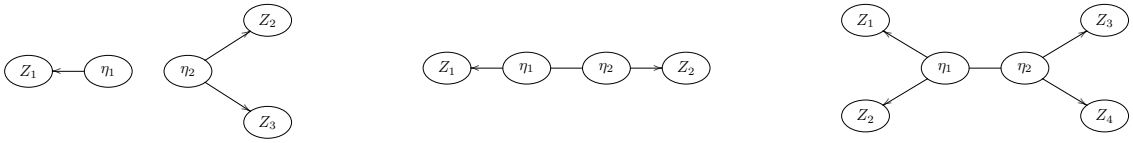

Figure 6: Left panel: *Case 1* ($c_{12} = 0$); Middle panel: *Case 2* ($c_{12} \neq 0$ and only one response per factor); Right panel: *Case 3* ($c_{12} \neq 0$ and at least one factor has multiple responses).

among the factors, by taking advantage of the following proposition (see also Theorem 19 of [24]):

Proposition 1 (Conditional Independence Test 1, CIT 1). *Let $\mathcal{G}$ be a pure linear latent variable model. Let η_1, η_2 be two factors in $\mathcal{G}$, and $\boldsymbol{Q}$ a set of factors in $\mathcal{G}$. Let Z_1 be an indicator of η_1, Z_2 be an indicator of η_2, and $\boldsymbol{Z_Q}$ be a set of indicators of $\boldsymbol{Q}$ containing at least two indicators per factor. Then η_1 is d-separated from η_2 given $\boldsymbol{Q}$ in $\mathcal{G}$ if and only if the rank of the correlation matrix of $\{Z_1, Z_2\} \cup \boldsymbol{Z_Q}$ is less than or equal to $|\boldsymbol{Q}|$ with probability 1 with respect to the Lebesgue measure over the linear coefficients and error variances of $\mathcal{G}$.*

One way to test if the rank of a covariance matrix in Gaussian models is at most q is to fit a factor analysis model with q latents and assess its significance [24]. The PC-MIMBuild algorithm arises when applying 'CIT 1' to test conditional independence among latent factors in the PC algorithm.

When a factor only has a single indicator, we propose to test conditional independence by making use of the following proposition:

Proposition 2 (Conditional Independence Test 2, CIT 2). *Let η_1, η_2 be two factors in $\mathcal{G}$, and $\boldsymbol{Q}$ a set of factors in $\mathcal{G}$. Let Z_1 be one of the indicators of η_1, Z_2 be one of the indicators of η_2, and $\boldsymbol{Z_Q}$ be all the indicators of $\boldsymbol{Q}$. Then η_1 is d-separated from η_2 given $\boldsymbol{Q}$ in $\mathcal{G}$ if and only if Z_1 is independent of Z_2 given $\boldsymbol{Z_Q}$ for all Z_1 and Z_2.*

This test can proceed via partial correlations for Gaussian data. By using 'CIT 1' or 'CIT 2', we generalize the PC-MIMBuild algorithm to the case where a factor has either a single or multiple indicators. Also, we extend the PC-MIMBuild algorithm to mixed continuous and discrete cases by learning the correlation matrix of response variables via the Gibbs sampler by [13] and taking it as input to the original PC-MIMBuild. The pseudocode of the extended PC-MIMBuild algorithm is summarized in Algorithm 2.

Algorithm 2 PC-MIMBuild algorithm

1: **Input:** Measurement models and indicator data $\boldsymbol{Y}$.
2: **Output:** Markov equivalent class $\mathcal{M}$ over latent factors.
3: Get correlation matrix of response variables via Gibbs sampler by [13] given $\boldsymbol{Y}$;
4: **if** Unconditional independence, i.e., $|\boldsymbol{Q}| = 0$, or all factors in $\boldsymbol{Q}$ have a single indicator. **then**
5: The PC algorithm with CIT 2;
6: **else**
7: The PC algorithm with CIT 1;
8: **end if**
9: Return $\mathcal{M}$.

C PSEUDOCODE OF THE GREEDY STEP-WISE PC ALGORITHM

The pseudocode of the greedy step-wise PC algorithm is summarized in Algorithm 3.

Algorithm 3 Greedy step-wise PC algorithm

1: **Input:** Measurement models (represented by the sparsity pattern of Λ) and indicator data $\boldsymbol{Y}$.
2: **Output:** Markov equivalent class $\mathcal{M}$ over latent factors.
3: **for** $i \in \{1, \ldots, k\}$ **do**
4: Let $\boldsymbol{Q} = \{Y_j : \lambda_{ji} \neq 0\}$, which is the set of indicators of the i-th factor;
5: **if** $|\boldsymbol{Q}| = 1$ **then**
6: Take the indicator data as the factor score, i.e., $\eta_i = Q$;
7: **else if** $|\boldsymbol{Q}| = 2$ **then**
8: Take the average of two indicators as the factor score, i.e., $\eta_i = (Q_1 + Q_2)/2$;
9: **else**
10: Fit the measurement model of the i-th factor to its indicator data $\boldsymbol{Q}$;
11: Obtain the factor score η_i from the fitted model;
12: **end if**
13: **end for**
14: Take pseudo data $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ as input to the 'Copula PC' algorithm to get $\mathcal{M}$.