We thank all the reviewers for the valuable insights and feedback. Below please see our response to the questions.

Rev #1: **(1)** *Brief description of SAEM:* Thank you for the suggestion. We will add a description before presenting the algorithm: "SAEM uses a stochastic approximation procedure to estimate the conditional expectation of the complete log-likelihood. More specifically, given the learned parameters in the current iteration, the values of latent variables are first sampled under the posteriori density. Then these sampled data are used to update the value of the conditional expectation of the complete log-likelihood with stochastic approximation." **(2)** p_l *and threshold:* We use cross-validation to choose the optimal time lag: we compare the averaged Q values over 10 iterations after convergence to choose p_l (Q values have small variations across iterations due to stochastic approximation). Regarding the threshold, 0.1 and 0.05 gave almost identical performance, and we reported the results with 0.1. We will include the results with Wald test to examine significance of edges, as in [24]. **(3)** Despite much time devoted to it, it does not seem feasible to have a concise proof for general cases in the near future, because it is hard to characterize identifiability in a Bayesian model.

Rev #2: **(Questions 23, 4, 8)** Causal direction flipping is not an assumption. We note that our model can handle more general cases, even the causal direction flips (line 143). For instance, in the brain network, different directions may be activated across subjects or states. It is hard to handle with traditional methods. **(24, 14)** We borrowed the framework of SAEM. The Gibbs sampling procedure in the E step and the derivations in the M step are our new contributions. We will make it clear. **(25, 12, 13)** Condition $l_s > 2q - 1$ is used to establish the identifiability of $P(X|z_k = 1)$ and $p(Z)$. As our model is a specific case of that by Vandermeulen & Scott (2015), we adapted their results. Following your suggestion, we will provide a complete proof in the supplements. **(26)** The current implementation is feasible for small- or median-scale systems (e.g., the 11-variable cellular network). As future work, we hope likelihood-free frameworks for parameter estimation with, e.g., adversarial learning, can improve the scalability. **(27, 18, 19)** In our implementation, skeletons and directions are generated in one step. Specifically, the graph G (including directions) is generated as follows: $G = triu(ones(m), 1). * binornd(1, p, m, m), G = PGP^T$, where $binornd$ is a random number generator for binomial distributions, P is a permutation matrix, and $G_{i,j} = 1$ means there is an edge from node i to node j. The proposed SSCM does cover the case of non-zero variance, but currently the identifiability proof is only shown in a specific case. In our simulations under non-zero variance settings, we never observed that the procedure converged to wrong solutions, suggesting that the non-zero-variance case is also identifiable. Following your suggestion, we will also include simulations with zero variances. **(28, 20, 21, 22, 10)** Non-Gaussianity can be checked by "normality test." For the fMRI and cellular data, the null hypothesis was rejected at significance level 0.01. Regarding causal structure variation, for fMRI data, it is well-known that neural connectivities may change across different external stimuli or intrinsic states. Hippocampus is activated in resting state, working on different tasks, such as consolidation of episodic, autobiographical, or declarative memory, depending on unmeasured intrinsic states. In different recording days, hippocampus may focus on different tasks, leading to nonconstant causal mechanisms [2,11,22,25]. For cellular data, causal structure may be different across conditions/interventions. **(0)** They are different. Our work focuses on propositional data and uses functional causal models to represent causal relationships. A well-known graphical model that uses propositional representation is the Bayesian network. Jensen et al.'s work focuses on relational data, using a relational schema to specify types of entities, relationships, and attributes. **(1, 3)** The difference is that our method can capture structure differences across groups/subjects and can provide both personalized and shared graphs, which are essential for many tasks. **(2)** "Omitted factors" refers to unobserved variables which affect the causal effects between observed variables. **(5)** In light of previous identifiability result, we are able to extend the current method to allow confounders, by making use of the identifiability of over-complete ICA. **(6)** p_l is the maximum time lag. **(9)** "Disjoint" means that the cycles do not interact with each other (i.e., no variable is involved in more than one cycle). Under this assumption, we can uniquely identify the causal graph. **(15)** The fMRI data we used contain 6 main regions in hippocampus. The Gibbs procedure is considered as convergent if the correlation between successive samples is smaller than a threshold. **(16)** If we know some edges are not possible, we can fix corresponding entries of A or B to 0. **(17)** If one randomly initializes all values, the F1 score is around 0.06 less, compared to the initialization given in lines 259-262, so the performance depends on initialization but not heavily. This will be made explicit.

Rev #3: **(1)** Yes. It is straightforward to extend it to more general forms. We will revise it, following your suggestion. **(2)** Thanks for the valuable comments. Yes. It should be $n \to \infty$. We will revise it in line 138: "...is identifiable, as $n \to \infty$, under the following conditions". **(3)** To be clearer, we can alternatively view SSCM as two separate steps: i) learning the causal structure of each subject and ii) clustering with the learned structure. Moreover, from Eqs. 13 and 14, we can see that the estimation of $P(X^s|z_k)$ depends only on causal adjacency matrices A and B. Thus, SSCM performs clustering based on only structure. Following your suggestion, we directly applied K-means on the data and the clustering accuracy is 0.87, lower than that by SSCM. In contrast, K-means performs clustering based on the data distribution, not the causal influences. The dataset we used is obtained under different interventions, and we know that intervention may break some causal influences, so the structure is expected to differ across conditions, making it sensible to use structure information for clustering. We will include the result of K-means and the variability for estimated graphs in each group. **Others:** Regarding the number of groups, a naive way is to use cross validation, and using the Silhouette score with close clusters merged may be a better solution. This will be briefly discussed.