A. Proof of Proposition 2

Proof. For $X_u \in X1$,

$$P(X_u|\backslash X_u) = \frac{P(X_u, X_{\backslash N(u)} \cap X2|X1 \cup X_{N(u)}\backslash X_u)}{P(X_{\backslash N(u)} \cap X2|X1 \cup X_{N(u)}\backslash X_u)}$$

Since $P(X_u|\backslash X_u) = P(X_u|X1 \cup X_{N(u)}\backslash X_u)$ by the Markovian property of PMN, we have $X_u \perp\!\!\!\perp X_{\backslash N(u)} \cap X2|X1 \cup X_{N(u)}\backslash X_u$.

$X_v \notin X_{N(u)}$ means $X_v \in X_{\backslash N(u)} \cap X2$. Using the weak union rule for conditional independence (see e.g., (Koller & Friedman, 2009), 2.1.4.3), we obtain $X_u \perp\!\!\!\perp X_v|\backslash\{X_u, X_v\}$.

For $X_u \in X2$, the proof is the same. □

B. Proof of Theorem 3

Proof. We define that $\mathbf{B}(i)$ is the set of passages contains X_i. Here we only show the proof that Eq. (1) holds for GPR. Let's denote ϕ_B as short for $\phi_B(X_B)$.

$$\begin{aligned}
&P(X_i|X1 \cup X_{N(i)}\backslash X_i) \\
=&\frac{\frac{1}{Z}\int_{X_{\backslash N(i)} \cap X2} P(X1)P(X_2)\prod_{B\in\mathbf{B}(G)}\phi_B}{\frac{1}{Z}\int_{X_i}\int_{X_{\backslash N(i)} \cap X2} P(X1)P(X_2)\prod_{B\in\mathbf{B}(G)}\phi_B} \\
=&\left(\frac{P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B}{\int_{X_i} P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B}\right)\cdot\left(\frac{\int_{X_{\backslash N(i)} \cap X2} P(X2)\prod_{B\in\backslash\mathbf{B}(i)}\phi_B}{\int_{X_{\backslash N(i)} \cap X2} P(X2)\prod_{B\in\backslash\mathbf{B}(i)}\phi_B}\right) \\
=&\frac{P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B}{\int_{X_i} P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B} \\
=&\frac{P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B}{\int_{X_i} P(X1)\prod_{B\in\mathbf{B}(i)}\phi_B}\cdot\frac{\frac{1}{Z}P(X2)\prod_{B\in\backslash\mathbf{B}(i)}\phi_B}{\frac{1}{Z}P(X2)\prod_{B\in\backslash\mathbf{B}(i)}\phi_B} \\
=&P(X_i|\backslash X_i),
\end{aligned}$$

from which, we obtain the desired equality. Note that we used the fact that $X_{\mathbf{B}(i)} \cap (X_{\backslash N(i)} \cap X2) = \emptyset$ from the second to the third and fourth line. □

C. Proof of Theorem 4

Proof. This proof is constructive. Let's clarify some notations used in this proof. Lower-case bold letter $\boldsymbol{a}$ is a vector-realization of a set of random variables A. $P(\boldsymbol{a}_K, \boldsymbol{c})$ means the probability of a realization where elements appearing on positions indexed by subgraph K are allowed to take random values, while other elements are fixed to value $c \in \mathrm{dom}(X)$. Note K might be $\emptyset$. We denote $P1(X)$ as the equivalency of marginal $P(X1)$.

First we define the following potential function:

$$\phi_S(X_S = \boldsymbol{x}_S) = \prod_{Z\subseteq S} \Delta_Z(X_Z = \boldsymbol{x}_Z)^{(-1)^{|S|-|Z|}},$$

where S is a subset of G, and

$$\Delta_Z(\boldsymbol{x}_Z) = \begin{cases} \frac{P(\boldsymbol{x}_Z,\boldsymbol{c})}{P1(\boldsymbol{x}_Z,\boldsymbol{c})P2(\boldsymbol{x}_Z,\boldsymbol{c})}, & \exists B \in \mathbf{B}(G), B \subseteq Z, \\ 1 & \text{otherwise}, \end{cases} \tag{6}$$

First we show by construction, the multiplication of all potential functions over all subgraph structures, i.e., $\prod_{S\subseteq G}\phi_S$ will actually give us the **PR**.

Due to the *inclusion-exclusion* principle (see, e.g.Koller & Friedman (2009), 4.4.2.1), it can be shown that

$$\prod_{S\subseteq G}\phi_S(X_S=\boldsymbol{x}_S)=\Delta_G(\boldsymbol{x}).$$

If the graph G contains any passage, then by definition $\Delta_G(\boldsymbol{x})=\frac{P(\boldsymbol{x})}{P1(\boldsymbol{x})P2(\boldsymbol{x})}$, which is exactly the PR. However, if G does not include any passage, meaning X_1 is completely independent of X_2, then $\Delta_G(\boldsymbol{x})=1$ by definition, which is the exact value that a PR would take in such case.

Second, we show this construction under PMN condition is actually a **GPR**. Specifically, we show if S is not a passage, then $\phi_S(X_S=\boldsymbol{x}_S)=1$, i.e. its potential function is nullified.

Obviously, for a "one-sided S", $X_S\cap X1=\emptyset$ or $X_S\cap X2=\emptyset$, by definition, $\phi_S=1$.

Otherwise, if S are "two-sided" but itself is not a passage, we should be able to find two nodes, indexed by $X_u\in X1\cap X_S$ and $X_v\in X2\cap X_S$, that are not connected by an edge. We may write the potential function for a subgraph S as

$$\phi_S(X_S=\boldsymbol{x}_S)=\prod_{W\subseteq S\setminus\{u,v\}}\left(\frac{\Delta_w(\boldsymbol{x}_W)\Delta_{W\cup\{u,v\}}(\boldsymbol{x}_{W\cup\{u,v\}})}{\Delta_{W\cup\{u\}}(\boldsymbol{x}_{W\cup\{u\}})\Delta_{W\cup\{v\}}(\boldsymbol{x}_{W\cup\{v\}})}\right)^*,$$

where $*$ means we do not care the exact power which can be either -1 or 1, and

$$\begin{aligned}&\frac{\Delta_W(\boldsymbol{x}_W)\Delta_{W\cup\{u,v\}}(\boldsymbol{x}_W)}{\Delta_{W\cup\{u\}}(\boldsymbol{x}_{W\cup\{u\}})\Delta_{W\cup\{v\}}(\boldsymbol{x}_{W\cup\{v\}})}=\\&\frac{P_WP_{W\cup\{u,v\}}}{P_{W\cup\{u\}}P_{W\cup\{v\}}}\cdot\frac{P2_{W\cup\{v\}}P2_WP1_{W\cup\{u\}}P1_W}{P1_WP2_WP1_{W\cup\{u\}}P2_{W\cup\{v\}}},\end{aligned}\tag{7}$$

where we have simplified the notation $P(\boldsymbol{x}_A,\boldsymbol{c})$ as P_A. The second factor in (7) is apparently 1. For the first factor in (7), we may divide both the numerator and denominator by $P_W\cdot P_W$. Then it yields $\frac{P(x_u,x_v|\boldsymbol{x}_W,\boldsymbol{c})}{P(x_u|\boldsymbol{x}_W,\boldsymbol{c})P(x_v|\boldsymbol{x}_W,\boldsymbol{c})}$ which equals to one if and only if $X_u\perp\!\!\!\perp X_v|\backslash\{X_u,X_v\}$. This is guaranteed by PMN condition and Proposition 2. □

D. Proof of Theorem 5

Since the PR is a density ratio between the joint density $p(\boldsymbol{x}_1,\boldsymbol{x}_2)$ and the product of two marginals $p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$, and the objective (5) is derived from the same sparsity inducing KLIEP criteria as it was discussed in Liu et al. (2015; 2016). The proof of Theorem 5 follows the primal-dual witness procedure (Wainwright, 2009).

First, the Assumptions 1, 2 and 3 we have made in Section 5 is essentially the same as those were imposed in Section 3.2 in Liu et al. (2016) (The Hessian of the negative log-likelihood is the sample Fisher information matrix). Then the proof follows the steps established in Section 4, Liu et al. (2016). However, the only thing we need to verify is that $\max_t\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\|$ is upper-bounded with high probability as $n\to\infty$. We formally state this in the following lemma:

Lemma 1. *If* $\lambda_n\geq\frac{24(2-\alpha)}{\alpha}\cdot\sqrt{\frac{c\log(m^2+m)/2}{n}}$, *then*

$$P\left(\max_t\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\|\geq\frac{\alpha\lambda_n}{4(2-\alpha)}\right)\leq 3\exp\left(-c''n\right),$$

where c *and* c'' *are some constants.*

Proof. For conveniences, let's denote the *approximated* PR model $\exp\left(\sum_{u\leq v}\boldsymbol{\theta}_{u,v}^\top\boldsymbol{\psi}(\boldsymbol{x}_{u,v})\right)/\hat{N}(\boldsymbol{\theta})$ as $\hat{g}(\boldsymbol{x};\boldsymbol{\theta})$. Since $\hat{g}(\boldsymbol{x};\boldsymbol{\theta})=\frac{N(\boldsymbol{\theta})}{\hat{N}(\boldsymbol{\theta})}g(\boldsymbol{x};\boldsymbol{\theta})$, and $\frac{\hat{N}(\boldsymbol{\theta})}{N(\boldsymbol{\theta})}=\frac{1}{\binom{n}{2}}\sum_{j\neq k}g(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta})$ is always bounded by $[C_{\min},C_{\max}]$, we can see $\hat{g}(\boldsymbol{x};\boldsymbol{\theta})$ is also bounded. For simplicity, we write

$$0<C'_{\min}\leq\hat{g}(\boldsymbol{x};\boldsymbol{\theta})\leq C'_{\max}<\infty.$$

We have

$$\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*) = -\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{f}_t(\boldsymbol{x}^{(i)})\right] + \left[\frac{1}{\binom{n}{2}}\sum_{j\leq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]})\right].$$

First we show that $\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\|$ can be upper-bounded as:

$$\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\| \leq \underbrace{\left\|-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{f}_t(\boldsymbol{x}^{(i)}) + \mathbb{E}_p[\boldsymbol{f}_t(\boldsymbol{x})]\right\|}_{a_n} + \underbrace{\left\|\frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]}) - \frac{1}{\binom{n}{2}}\sum_{j\neq k}g(\boldsymbol{x}^{[i,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]})\right\|}_{b_n}$$
$$+ \underbrace{\left\|\frac{1}{\binom{n}{2}}\sum_{j\neq k}g(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]}) - \mathbb{E}_{p1,p2}\left[g(\boldsymbol{x};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x})\right]\right\|}_{\|\boldsymbol{w}_n\|},$$

We now need Hoeffding inequality (Hoeffding, 1963) for bounded-norm vector random variables which has appeared in previous literatures such as (Steinwart & Christmann, 2008): For a set of bounded zero-mean vector-valued random variable $\{\boldsymbol{y}_i\}_{i=1}^n$, $\|\boldsymbol{y}\| \leq c$, we have

$$P(\left\|\sum_{i=1}^{n}\boldsymbol{y}_i\right\| \geq n\epsilon) \leq \exp\left(\frac{-n\epsilon^2}{2c^2}\right),$$

for all $\epsilon \geq \frac{2c}{\sqrt{n}}$. Now it is easy to see

$$P(a_n \geq \epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{C'^2_{\boldsymbol{f}_t,\max}}\right) \tag{8}$$

as long as

$$\epsilon \geq \frac{C'_{\boldsymbol{f}_t,\max}}{2\sqrt{n}}. \tag{9}$$

As to b_n, it can be upper-bounded by

$$\begin{aligned}
b_n &= \left\|\frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{(i)}) - \frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]})\right\| \\
&= \left\|\frac{\hat{N}(\boldsymbol{\theta}^*)}{N(\boldsymbol{\theta}^*)}\frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]}) - \frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]})\right\| \\
&\leq \left\|\frac{1}{\binom{n}{2}}\sum_{j\neq k}\hat{g}(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*)\boldsymbol{f}_t(\boldsymbol{x}^{[j,k]})\right\| \cdot \left\|\frac{\hat{N}(\boldsymbol{\theta}^*)}{N(\boldsymbol{\theta}^*)} - 1\right\| \\
&\leq C'_{\max}C'_{\boldsymbol{f}_t,\max}\left|\frac{1}{\binom{n}{2}}\sum_{j\neq k}g(\boldsymbol{x}^{[j,k]};\boldsymbol{\theta}^*) - 1\right|,
\end{aligned}$$

and due to Hoeffding inequality of the U-statistics (see (Hoeffding, 1963), 5b) we may obtain:

$$P(b_n > \epsilon) < 2\exp\left(-\frac{2n\epsilon^2}{C^2_{\max}C'^2_{\max}C'^2_{\boldsymbol{f}_t,\max}}\right). \tag{10}$$

As to $\boldsymbol{w}_n$, we first bound its i-th element $w_{i,n}$ using Hoeffding inequality for U-statistics,

$$P(|w_{i,n}| \geq \epsilon) \leq 2\exp\left(-\frac{2nb\epsilon^2}{C_{\max}^2 C_{\boldsymbol{f}_t,\max}^2}\right),$$

thus by using the union bound, we have

$$P(\|\boldsymbol{w}_n\|_\infty \geq \epsilon) \leq 2b\exp\left(-\frac{2nb\epsilon^2}{C_{\max}^2 C_{\boldsymbol{f}_t,\max}^2}\right),$$

and since $\|\boldsymbol{w}_n\| \leq \sqrt{b}\|\boldsymbol{w}_n\|_\infty$, we have

$$P(\|\boldsymbol{w}_n\| \geq \epsilon) \leq P(\sqrt{b}\|\boldsymbol{w}_n\|_\infty \geq \epsilon) \leq 2b\exp\left(-\frac{2n\epsilon^2}{C_{\max}^2 C_{\boldsymbol{f}_t,\max}^2}\right). \tag{11}$$

Therefore, combining (8), (10) and (11):

$$P(\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\| \geq 3\epsilon) \leq P(a_n + b_n + c_n \geq 3\epsilon) \leq c''\exp\left(-\frac{n\epsilon^2}{c'}\right),$$

where c' is a constant defined as $c' = \max\left(\frac{1}{2}C_{\max}^2 C_{\max}^2 C_{\boldsymbol{f}_t,\max}'^2, \frac{1}{2}C_{\max}^2 C_{\boldsymbol{f}_t,\max}^2, \frac{1}{2}C_{\boldsymbol{f}_t,\max}'^2\right)$, and $c'' = 2b + 3$, given $\epsilon \geq \frac{2C'_{\boldsymbol{f}_t,\max}}{\sqrt{n}}$. Applying the union-bound for all $t \in S \cup S^c$,

$$P(\max_{t\in S\cup S^c}\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\| \geq 3\epsilon) \leq \frac{c''(m^2+m)}{2}\exp\left(-\frac{n\epsilon^2}{c'}\right),$$

$$P\left(\max_{t\in S\cup S^c}\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\| \geq \frac{\alpha\lambda_n}{4(2-\alpha)}\right) \leq \frac{c''(m^2+m)}{2}\exp\left(-\left(\frac{\alpha\lambda_n}{12(2-\alpha)}\right)^2\frac{n}{c'}\right),$$

and when $\lambda_n \geq \frac{24(2-\alpha)}{\alpha}\sqrt{\frac{c'\log(m^2+m)/2}{n}}$,

$$P\left(\max_{t\in S\cup S^c}\|\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\theta}^*)\| \geq \frac{\alpha\lambda_n}{4(2-\alpha)}\right) \leq c''\exp\left(-c'''n\right),$$

where c''' is a constant. Assume that $\log\frac{m^2+m}{2} > 1$ and we set λ_n as

$$\lambda_n \geq \frac{24(2-\alpha)}{\alpha}\sqrt{\frac{(c' + C_{\boldsymbol{f}_t,\max}^2)\log(m^2+m)/2}{n}},$$

then (9), the condition of using vector Hoeffding-inequality is satisfied. □

Given Lemma 1, we may obtain other technical results, such as the estimation error bound, using the same proof as it was demonstrated in Section 4, Liu et al. (2016).

E. Experimental Settings

We measure the performance of three methods using True Postive Rate (TPR) and True Negative Rate (TNR) that are used in Zhao et al. (2014). The TPR and TFR are defined as:

$$\text{TPR} = \frac{\sum_{t'\in S}\delta(\hat{\boldsymbol{\theta}}_{t'} \neq \mathbf{0})}{\sum_{t'\in S}\delta(\boldsymbol{\theta}^*_{t'} \neq \mathbf{0})},\quad \text{TNR} = \frac{\sum_{t''\in S^c}\delta(\hat{\boldsymbol{\theta}}_{t''} = \mathbf{0})}{\sum_{t''\in S^c}\delta(\boldsymbol{\theta}^*_{t''} = \mathbf{0})},$$

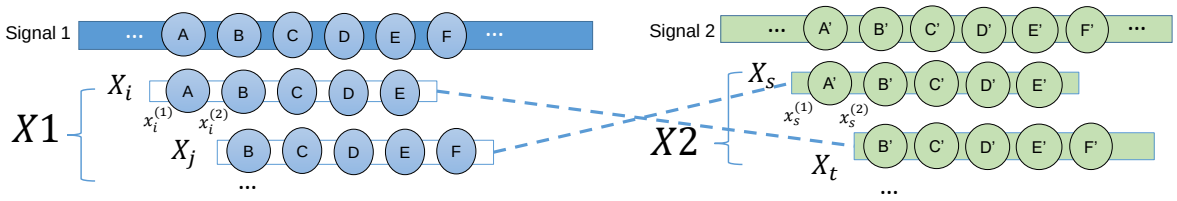

Figure 7. The illustration of sequence matching problem formulation.

where δ is the indicator function.

The differential learning method (Zhao et al., 2014) used in Section 6.1 learns the difference between two precision matrices. In our setting, if one can learn the difference between the precision matrices of $p(\boldsymbol{x})$ and $p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$, one can figure out all edges that go across two groups ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$).

This method requires sample covariance matrices of $p(\boldsymbol{x})$ and $p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$ respectively. The sample covariance of $p(\boldsymbol{x})$ is easy to compute given joint samples. However, to obtain the sample covariance of $p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$, we would again need the U-statistics (Hoeffding, 1963) introduced in line Section 4. We may approximate the u, v-th element of the covariance matrix of $p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$ using the formula: $\Sigma_{u,v} = \frac{1}{\binom{n}{2}} \sum_{j \neq k} x_v^{[j,k]} x_u^{[j,k]}$, assuming the joint distribution has zero mean.

F. Illustration of Sequence Matching

We plot the illustrations of our sequence matching problem formulation from two sequences in Figure 7.