APPENDIX A

PROOF OF LEMMA 1

To show the first property, we define the $|X_i| \times (|X_i| \cdot |X_2| \cdots |X_d|)$ matrix $B_i$, for $i = 1, \ldots, d$, as

$$B_i(x'_i; x^d) = \begin{cases} \sqrt{P_{X_i} (x'_i)} & \text{if } x'_i = x_i \\ 0 & \text{otherwise.} \end{cases}$$

Then, one can verify that

$$B = \begin{bmatrix} B_1 \\ \vdots \\ B_1^T \cdots B_d^T \end{bmatrix},$$

which implies that $B$ is positive semidefinite.

To establish the second property, note that $\psi(0)$ is an eigenvector of $B$ with eigenvalue $d$, and thus $(\psi(0))^T B \psi(0) = d$.

Moreover, it is shown in [23] that the largest singular value of $B_{ij}$ is 1, i.e., $\|B_{ij}\|_s = 1$, where $\|\cdot\|_s$ denotes the spectral norm of its matrix argument.

Therefore, for $\psi = [\psi_1^T, \ldots, \psi_d^T]^T$ with each $\psi_i$ being an $|X_i|$-dimensional vector, we have

$$\psi^T B \psi = \sum_{i=1}^d \sum_{j=1}^d \psi_i^T B_{ij} \psi_j \leq \sum_{i=1}^d \sum_{j=1}^d \|\psi_i\| \cdot \|B_{ij}\|_s \cdot \|\psi_j\| = d \|\psi\|^2,$$

where the second inequality follows from the fact that the arithmetic mean is no greater than the quadratic mean. Hence, we have

$$\max_{\psi: \|\psi\| = 1} \psi^T B \psi = d,$$

i.e., the largest eigenvalue of $B$ is $d$.

To verify the third property, we construct $\psi'$ as

$$\psi' = \begin{bmatrix} \psi_1' \\ \vdots \\ \psi_1' \end{bmatrix},$$

where $\psi_1' \in \mathbb{R}^{|X_1|}$ is chosen such that $\langle \psi_1', v_1 \rangle = 0$ and $\|\psi_1'\|^2 = 1$, and where 0 is the $(m - |X_1|)$-dimensional zero vector. Therefore, we have $\langle \psi', \psi(0) \rangle = 0$ and $\|\psi'\|^2 = 1$.

Note that the second eigenvalue $\lambda^{(1)}$ of $B$ can be written as

$$\lambda^{(1)} = \max_{\psi: \|\psi\| = 1, \langle \psi, \psi(0) \rangle = 0} \psi^T B \psi,$$

which implies that $\lambda^{(1)} \geq (\psi')^T B \psi' = \|\psi'\|^2 = 1$.

To verify the fourth property, we define the $(d - 1)$-dimensional subspace $S_{\text{eig}}$ as

$$S_{\text{eig}} = \left\{ \psi = [\alpha_1 v_1^T, \ldots, \alpha_d v_d^T]^T: \sum_{i=1}^d \alpha_i = 0 \right\}.$$ (A-1)

Then, for all $\psi \in S_{\text{eig}}$, from $B_{ij} v_j = v_i$, it is straightforward to verify that $B \psi = 0_m$, where $0_m$ is the zero vector in $\mathbb{R}^m$. Therefore, $S_{\text{eig}}$ is an eigenspace of $B$ associated with $d - 1$ zero eigenvalues. Since $B$ is positive semidefinite, without loss of generality we can assume that $S_{\text{eig}}$ is spanned by $\psi(m-d+1), \ldots, \psi(m-1)$, which correspond to eigenvalues $\lambda(m-d+1) = \cdots = \lambda(m-1) = 0$.

Finally, to establish the last property, for each $\ell = 1, \ldots, m - d$, from $\langle \psi^{(\ell)}, \psi^{(0)} \rangle = 0$ we have

$$\sum_{i=1}^d \langle \psi^{(\ell)}, v_i \rangle = 0.$$

Therefore, from the third property, we have

$$\psi' = \begin{bmatrix} \psi_1' \psi_1 \psi_1' \cdots \psi_d' \psi_d \end{bmatrix} \in S_{\text{eig}}.$$

Hence, we obtain $\langle \psi', \psi(0) \rangle = 0$, i.e.,

$$\sum_{i=1}^d \langle \psi^{(\ell)}, v_i \rangle = 0,$$

which implies that $\langle \psi^{(\ell)}, v_i \rangle = 0$ for $i = 1, \ldots, d$.

APPENDIX B

PROOF OF THEOREM 1

First, we replace $\delta$ by $\frac{1}{2} \epsilon^2$ for the convenience of presentation when applying the local geometric approach. With this notation, the constraint (2) becomes

$$I(U; X^d) \leq \frac{1}{2} \epsilon^2,$$ (A-2)

with $\epsilon$ assumed to be small. Then, it follows from (3) and (A-2) that for all $u$, the conditional distribution $P_{X^d|U=u}$ can be written as a perturbation to the marginal distribution:

$$P_{X^d|U}(x^d|u) = P_{X^d}(x^d) + \epsilon \sqrt{P_{X^d}(x^d)} \phi_u(x^d)$$ (A-3)

where $\phi_u$ can be viewed as an $|X_1| \cdots |X_d|$-dimensional vector. Moreover, it follows from the second order Taylor’s expansion for the K-L divergence that

$$I(U; X^d) = \mathbb{E}_U \left[ D(P_{X^d|U} \| P_{X^d}) \right] = \frac{1}{2} \epsilon^2 \mathbb{E}_U \left[ \| \phi_U \|^2 \right] + o(\epsilon^2),$$

where $\| \cdot \|$ denotes the $l_2$-norm. Thus, by ignoring the higher order term of $\epsilon$ as we assume $\epsilon$ to be small, the constraint $I(U; X^d) \leq \frac{1}{2} \epsilon^2$ can be reduced to

$$\mathbb{E}_U[\| \phi_U \|^2] \leq 1.$$ (A-4)

In addition, the objective function $\ell(X^d|U)$ can also be expressed in terms of mutual information:

$$D(P_{X^d}||P_{X_1} \cdots P_{X_d}) - D(P_{X^d}||P_{X_1} \cdots P_{X_d}|U) = \sum_{i=1}^d I(U; X_i) - I(U; X^d)$$ (A-5)
and for each $i$, the mutual information $I(U; X_i)$ can be again approximated as the $l_2$-norm square

$$I(U; X_i) = \frac{1}{2} \epsilon^2 \mathbb{E}_U[\|\psi_{i,U}\|^2] + o(\epsilon^2),$$

where for $U = u$, the vector $\psi_{i,u}$ is the $|\mathcal{X}_i|$-dimensional perturbation vector defined as

$$\psi_{i,u}(x_i) = \frac{P_{X_i|U}(x_i|u) - P_{X_i}(x_i)}{\epsilon \sqrt{P_{X_i}(x_i)}} \quad (A-6)$$

Then, by ignoring the higher order terms of $\epsilon$, the optimization problem we want to solve can be transferred to a linear algebraic problem

$$\max_{\|\phi_U\|^2 \leq 1} \sum_{i=1}^{d} \mathbb{E}_U[\|\psi_{i,U}\|^2] = \mathbb{E}_U[\|\phi_U\|^2]. \quad (A-7)$$

To solve (A-7), observe that $P_{X_i}$ and $P_{X_i|U}$ are marginal distributions of $P_{X\times A}$ and $P_{X\times A|U}$, thus there is a correlation between $\phi_U$ and $\psi_{i,U}$:

$$\psi_{i,u}(x_i) = \frac{\sqrt{P_{X_i}(x_i)}}{\sqrt{P_{X_i}(x_i)}} \phi_U(x^d)$$

which can be represented in matrix form as $\psi_{i,u} = B_i \cdot \phi_u$, where $B_i$ is an $|\mathcal{X}_i| \times (|\mathcal{X}_1| \times |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$ matrix with entries

$$B_i(x'_i; (x_1, \ldots, x_d)) = \begin{cases} \sqrt{P_{X_i}(x'_i)} & \text{if } x'_i = x_i \\ \sqrt{P_{X_i}(x_i)} & \text{otherwise.} \end{cases} \quad (A-8)$$

Therefore, if we define an $(|\mathcal{X}_1| + \cdots + |\mathcal{X}_d|) \times (|\mathcal{X}_1| \cdots |\mathcal{X}_m|)$-dimensional matrix

$$B_0 = \begin{bmatrix} B_1 \\ \vdots \\ B_d \end{bmatrix}, \quad (A-9)$$

then since

$$\sum_{i=1}^{d} \mathbb{E}_U[\|\psi_{i,U}\|^2] = \sum_{i=1}^{d} \mathbb{E}_U[\|B_i \cdot \phi_U\|^2] = \mathbb{E}_U[\|B_0 \cdot \phi_U\|^2],$$

we can rewrite (A-7) as

$$\max_{\|\phi_U\|^2 \leq 1} \mathbb{E}_U[\|B_0 \cdot \phi_U\|^2] = \mathbb{E}_U[\|\phi_U\|^2]. \quad (A-10)$$

Moreover, since $\phi_U$ is a perturbation vector of probability distributions, by summing over all $x^d$ for both sides of (A-3), it has to satisfy an extra constraint

$$\sum_{x^d} \sqrt{P_{X\times A}(x^d)} \phi_U(x^d) = 0, \quad (A-11)$$

which implies that $\phi_U$ is orthogonal to an $(|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$-dimensional vector $\phi^{(0)}$, whose entries are $\sqrt{P_{X\times A}(x^d)}$. In particular, it is shown in [23] that $\phi^{(0)}$ is the right singular vector of $B_0$ with the largest singular value $\sigma_0 = \sqrt{d}$, and the corresponding left singular vector is $\psi^{(0)}$. In addition, it can be verified that $B_0$ satisfies $B_0B_0^T = B$ with $B$ as defined in (4). Therefore, the second largest singular value of $B_0$ is $\sigma_1 = \sqrt{\lambda^{(2)}} \geq 1$, and the optimal solution of (A-10) is to align the vectors $\phi_{U=0}$, for all $u$, along the second largest right singular vector of $B_0$.

It turns out that it is easier to compute the second largest left singular vector of $B_0$ instead of the right one. This is equivalent to computing the second largest eigenvector of the matrix $B_0B_0^T = B$.

Now, the second largest right singular vector $\phi^{(1)}$ of $B_0$ can be computed as

$$\phi^{(1)}(x^d) = \frac{1}{\sqrt{\lambda^{(1)}}(B_0^T \psi^{(1)})(x^d)} \frac{1}{\sqrt{\lambda^{(1)}}} \left( \sqrt{P_{X\times A}(x^d)} \sum_{i=1}^{d} \frac{\psi^{(1)}(x_i)}{\sqrt{P_{X_i}(x_i)}} \right)$$

where $B_0^T \psi^{(1)}(x^d)$ is a vector and $(B_0^T \psi^{(1)})(x^d)$ is the $x^d$-th entry of this vector. Since all the $\phi_{U=u}$ should be aligned to $\phi^{(1)}$, there exists a function $h: U \mapsto \mathbb{R}$, such that

$$P_{X\times A}(x^d|u) = P_{X\times A}(x^d) \left( 1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o(\epsilon),$$

where the term $o(\epsilon)$ comes from the local approximation we made for (A-4). Therefore, the optimal joint distributions for our optimization problem can be written as

$$P_{U\times A}(u, x^d) = P_U(u)P_{X\times A}(x^d) \left( 1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o(\epsilon). \quad (A-13)$$

Note that if we sum both sides of (A-13) over all $u \in U$, then we have $\sum_{u \in U} P_U(u)h(u) = 0$, which implies that $h(U)$ is a zero-mean function. Moreover, it is easy to compute from (A-4) that the variance $\mathbb{E}[h^2(U)] = 1$. Finally, note that the exponential family $P_{\exp}^{(\delta)}$, when $\delta$ is small, can be written as

$$P_{\exp}^{(\delta)} = \left\{ P_U(u)P_{X\times A}(x^d) \left( 1 + \frac{\sqrt{\lambda^{(1)}}}{\sqrt{2\delta h(u)}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o(\sqrt{\delta}) : h \in \mathcal{H}_\delta \right\}. \quad (A-14)$$

Since $\delta = \frac{1}{2} \epsilon^2$ the proof is completed by comparing (A-14) and (A-13).

**APPENDIX C**

**PROOF OF THEOREM 2**

First, we introduce a useful lemma (see, e.g., [24, Corollary 4.3.39, p. 248]).

**Lemma A.1.** Given an arbitrary $k_1 \times k_2$ matrix $A$ and any $k \in \{1, \ldots, \min\{k_1, k_2\}\}$, we have

$$\max_{M \in \mathbb{R}^{k\times k}} \|AM\|_F^2 = k \sum_{i=1}^{k} \sigma_i^2, \quad (A-15)$$
where $\| \cdot \|_F$ denotes the Frobenius norm, and where $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$ denotes the singular values of $A$. Moreover, the maximum in (A-15) can be achieved by $M = [v_1 \cdots v_k] Q$, where $v_i$ denotes the right singular vector of $A$ corresponding to $\sigma_i$, for $i = 1, \ldots, \min(m,n)$, and $Q \in \mathbb{R}^{k \times k}$ is an (arbitrary) orthogonal matrix.

To begin the proof, similar to Theorem 1, we replace $\delta$ by $\frac{1}{2} \epsilon^2$ and write the conditional distribution $P_{X^d|U^k=x^d}$ as a perturbation to the joint distribution $P_{X^d}$:

$$P_{X^d|U^k}(x^d|u^k) = P_{X^d}(x^d) + \epsilon \sqrt{P_{X^d}(x^d)} \Phi_{u^k}(x^d),$$  \hspace{1cm} (A-16)

where $\Phi_{u^k}$ is an $(|X_1| \cdot |X_2| \cdots |X_d|)$-dimensional vector. Again, it follows from the second-order Taylor series expansion of the K-L divergence that

$$I(U^k; X^d) = \mathbb{E}_{U^k}[D(P_{X^d|U^k}||P_{X^d})]$$

$$= \frac{1}{2} \epsilon^2 \mathbb{E}_{U^k} \left[ \| \Phi_{u^k} \|^2 \right] + o(\epsilon^2).$$

Similarly, for all $i = 1, \ldots, k$, the conditional distribution $P_{X^d|U_i=u_i}$ can be written as

$$P_{X^d|U_i}(x^d|u_i) = P_{X^d}(x^d) + \epsilon \sqrt{P_{X^d}(x^d)} \Phi_{u_i}(x^d),$$  \hspace{1cm} (A-17)

and we have

$$I(U_i; X^d) = \frac{1}{2} \epsilon^2 \mathbb{E}_{U_i} \left[ \| \Phi_{u_i} \|^2 \right] + o(\epsilon^2).$$

Therefore, by ignoring the higher order terms of $\epsilon$, the first constraint can be reduced to

$$1 \geq \mathbb{E}_{U_i} \left[ \| \Phi_{u_i} \|^2 \right] \geq \cdots \geq \mathbb{E}_{U_k} \left[ \| \Phi_{u_k} \|^2 \right].$$

Moreover, due to the independence and conditional independence among the $U^k$, $\Phi_{u^k}$ and $\Phi_{u_i}$ satisfy

$$\Phi_{u^k} = \sum_{i=1}^k \Phi_{u_i} + o(1)$$  \hspace{1cm} (A-18)

and

$$\langle \Phi_{u_i}, \Phi_{u_j} \rangle = 0, \hspace{0.5cm} \text{for all } i \neq j, u_i \in U_i, u_j \in U_j.$$  \hspace{1cm} (A-19)

Indeed, we have

$$P_{X^d|U^k}(x^d|u^k) = \frac{P_{X^d}(x^d)P_{U^k|X^d}(u^k|x^d)}{P_{U^k}(u^k)}$$

$$= P_{X^d}(x^d) \prod_{i=1}^k \frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)},$$

which implies

$$\frac{P_{X^d|U_i}(x^d|u_i)}{P_{X^d}(x^d)} = \prod_{i=1}^k \frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)} \frac{k}{P_{X^d}(x^d)}.$$  \hspace{1cm} (A-20)

Substituting (A-16) and (A-17) into (A-20) then yields

$$1 + \epsilon \frac{\Phi_{u^k}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \prod_{i=1}^k \left[ 1 + \epsilon \frac{\Phi_{u_i}(x_i)}{\sqrt{P_{X^d}(x_i)}} \right],$$

and via comparing the $\epsilon$-order terms for both sides we obtain

$$e^2 \langle \phi_{u^k}, \phi_{u_i} \rangle$$

$$= \epsilon^2 \sum_{x^d} \phi_{u_i}(x^d) \phi_{u_j}(x^d)$$  \hspace{1cm} (A-21)

$$= \sum_{x^d} \left( \frac{1}{P_{X^d}(x^d)} \cdot [P_{X^d|U_i}(x^d|u_i) - P_{X^d}(x^d)] \right) \cdot [P_{X^d|U_j}(x^d|u_j) - P_{X^d}(x^d)]$$  \hspace{1cm} (A-22)

$$= \sum_{x^d} \frac{P_{X^d|U_i}(x^d|u_i)P_{X^d|U_j}(x^d|u_j)}{P_{X^d}(x^d)} - 1$$  \hspace{1cm} (A-23)

$$= \sum_{x^d} \frac{P_{X^d}(x^d)}{P_{U_i}(u_i)} \sum_{x^d} P_{X^d}(x^d) P_{U_i|X^d}(u_i|x^d) P_{U_j|X^d}(u_j|x^d) - 1$$  \hspace{1cm} (A-24)

$$= 0,$$

where to obtain (A-24) we have again exploited the independence and conditional independence of $U_i$ and $U_j$.

In addition, the objective function $L(X^d|U^k)$ can be expressed as [cf. (A-5)]

$$D(P_{X^d||X_i} \cdots P_{X_d}) - D(P_{X^d||X_1} \cdots P_{X_d|U^k})$$

$$= \sum_{i=1}^d I(U^k; X_i) - I(U^k; X^d)$$

$$= \sum_{i=1}^d I(U^k; X_i) - \sum_{j=1}^k I(U_j; X^d),$$  \hspace{1cm} (A-25)

where to obtain the last equality we have used the fact that

$$I(U^k; X^d) = \mathbb{E}_{U^k,X^d} \left[ \log \frac{P_{U^k|X^d}(U^k|X^d)}{P_{U^k}(U^k)} \right]$$  \hspace{1cm} (A-26)

$$= \mathbb{E}_{U^k,X^d} \left[ \sum_{j=1}^k \log \frac{P_{U_j|X^d}(U_j|X^d)}{P_{U_j}(U_j)} \right]$$  \hspace{1cm} (A-27)

$$= \sum_{j=1}^k I(U_j; X^d),$$  \hspace{1cm} (A-28)

and where (A-27) follows from the facts that $U_1, \ldots, U_k$ are mutually independent and are conditionally independent given $X^d$.

For each $i$, the mutual information $I(U^k; X_i)$ can be approximated as

$$I(U^k; X_i) = \frac{1}{2} \epsilon^2 \mathbb{E}_{U^k} \left[ \| \psi_{i,u^k} \|^2 \right] + o(\epsilon^2),$$

where for $U_k = u^k$, the vector $\psi_{i,u_k}$ is an $|X_i|$-dimensional perturbation vector defined as

$$\psi_{i,u^k}(x_i) = \frac{P_{X|U^k}(x_i|u^k) - P_{X_i}(x_i)}{\epsilon \sqrt{P_{X_i}(x_i)}}.$$
Therefore, by ignoring the higher order terms of $\epsilon$, the maximization of total correlation can be rewritten as

$$
\max_{\phi_{u,k}} \sum_{i=1}^{d} \mathbb{E}_{U^k} \left[ \| \psi_{i,U^k} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-29a)
$$

subject to: $1 \geq \mathbb{E}_{U_i} \left[ \| \phi_{U_i} \|^2 \right] \geq \cdots \geq \mathbb{E}_{U_k} \left[ \| \phi_{U_k} \|^2 \right] \geq 1$ \quad (A-29b)

$$
\left( \phi_{u_i}, \phi_{u_j} \right) = 0, \quad i \neq j, u_i \in \mathcal{U}_i, u_j \in \mathcal{U}_j \quad (A-29c)
$$

$$
\left\{ \phi_{u_j}, \phi^{(0)} \right\} = 0, \forall u_j \in \mathcal{U}_j, j = 1, \ldots, k \quad (A-29d)
$$

$$
\phi_{u,k} = \frac{1}{k} \sum_{j=1}^{k} \phi_{u_j}, \forall u^k \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_k. \quad (A-29e)
$$

To solve (A-29), first observe that we have $\psi_{i,U^k} = B_i \phi_{U^k}$, where $B_i$ is as defined in (A-8). Then, the objective function (A-29a) can be rewritten as

$$
\sum_{i=1}^{d} \mathbb{E}_{U^k} \left[ \| \psi_{i,U^k} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] = \sum_{i=1}^{d} \mathbb{E}_{U^k} \left[ \| B_i \phi_{U^k} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-30)
$$

$$
= \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| B_0 \phi_{U^j} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-31)
$$

$$
= \mathbb{E}_{U^k} \left[ \sum_{j=1}^{k} \| B_0 \phi_{U^j} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-32)
$$

$$
= \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| B_0 \phi_{U^j} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-33)
$$

$$
= \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| B_0 \phi_{U^j} \|^2 \right] - \sum_{j=1}^{k} \mathbb{E}_{U^j} \left[ \| \phi_{U^j} \|^2 \right] \quad (A-34)
$$

where $B_0$ is as defined in (A-9). To obtain (A-33), we have used the fact that, for $i \neq j$,

$$
\mathbb{E}_{U^k} \left[ \phi_{U_i}^T B_0^T B_0 \phi_{U_j} \right] = \mathbb{E}_{U_i} \left[ \phi_{U_i} \right]^T B_0^T B_0 \mathbb{E}_{U_j} \left[ \phi_{U_j} \right] = 0,
$$

where the first equality follows from the fact that $U_i$ and $U_j$ are independent, and the second equality follows from that $\mathbb{E}_{U_i} \left[ \phi_{U_i} \right] = 0$.

To maximize (A-34), $\phi_{u_i}$ should be aligned to the same direction for all $u_i \in \mathcal{U}_i$. Otherwise, we can align all $\phi_{u_i}$ to

$$
\arg \max_{\phi_{u_i}, \forall u_i \in \mathcal{U}_i} \frac{\| B_0 \phi_{u_i} \|^2}{\| \phi_{u_i} \|^2}
$$

while keeping $\mathbb{E}_{U^i} \left[ \| \phi_{U_i} \|^2 \right]$ fixed, which yields a larger value for the objective function.

Therefore, for each $i$ and $u_i \in \mathcal{U}_i$, we can write $\phi_{u_i}$ as

$$
\phi_{u_i} = h_i(u_i) \phi_i, \quad (A-35)
$$

where $h_i: \mathcal{U}_i \rightarrow \mathbb{R}$ and $\phi_i$ is a unit-norm vector. Then, we have

$$
\mathbb{E}_{U_i} \left[ \phi_{U_i} \right] = \mathbb{E}_{U_i} \left[ h_i(u_i) \right] \phi_i = 0
$$

$$
\mathbb{E}_{U_i} \left[ \| \phi_{U_i} \|^2 \right] = \mathbb{E}_{U_i} \left[ h_i^2(u_i) \right] \| \phi \|^2
$$

$$
\mathbb{E}_{U_i} \left[ \| B_0 \phi_{U_i} \|^2 \right] = \mathbb{E}_{U_i} \left[ h_i^2(u_i) \right] \| B_0 \phi_i \|^2
$$

Now, the constraint (A-29b) can be reduced to

$$
1 \geq \mathbb{E}_{U_i} \left[ h_i^2(U_i) \right] \geq \cdots \geq \mathbb{E}_{U_k} \left[ h_k^2(U_k) \right]. \quad (A-37)
$$

In addition, it follows from (A-36) that $\mathbb{E}_{U_i} \left[ \| B_0 \phi_{U_i} \|^2 \right] = \mathbb{E}_{U_i} \left[ h_i^2(U_i) \right] \| B_0 \phi_i \|^2 - 1$. As a result, to maximize (A-34), $h_i$ should be chosen such that

$$
\mathbb{E}_{U_i} \left[ h_i^2(U_i) \right] = \begin{cases} 1 & \text{if } \| B_0 \phi_i \|^2 > 1 \\ 0 & \text{otherwise.} \end{cases} \quad (A-38)
$$

Then, from (A-37) there exists $k_0 \in \{1, \ldots, k\}$ such that

$$
\mathbb{E}_{U_i} \left[ h_i^2(U_i) \right] = \begin{cases} 1 & i = 1, \ldots, k_0 \\ 0 & i > k_0. \end{cases}
$$

and the objective function (A-34) can be reduced to

$$
\sum_{j=1}^{k_0} \mathbb{E}_{U^j} \left[ \| B_0 \phi_{U^j} \|^2 - \| \phi_{U^j} \|^2 \right] = k_0 \| B_0 \phi_i \|^2 - k_0
$$

$$
= \| B_0 \Phi_0 \|^2 - k_0,
$$

where we have defined $\Phi_0 \triangleq [\phi_1 \cdots \phi_{k_0}]$.

As a result, the optimization problem (A-29) is equivalent to

$$
\max_{\Phi_0} \| B_0 \Phi_0 \|^2 - k_0 \quad (A-39a)
$$

subject to: $\Phi_0 \Phi_0^T = I_{k_0}, \quad (A-39b)$

$$
\Phi_0 \neq 0_{k_0}, \quad (A-39c)
$$

where $I_{k_0}$ is the identity matrix of order $k_0$, and $0_{k_0}$ is the zero vector in $\mathbb{R}^{k_0}$. In addition, since $\phi^{(0)}$ is the first right singular vector of $B_0$, (A-39) can be further reduced to

$$
\max_{\Phi_0} \| B_0 \Phi_0 \|^2 - k_0 \quad (A-40a)
$$

subject to: $\Phi_0^T \Phi_0 = I_{k_0}, \quad (A-40b)$

where $\Phi_0 \triangleq B_0 - \sqrt{\lambda^{(0)} \psi^{(0)}} (\phi^{(0)})^T$.

From Lemma A-1, the optimal value of (A-40) is

$$
\sum_{i=1}^{k_0} \lambda^{(i)} - k_0 = \sum_{i=1}^{k_0} \left[ \lambda^{(i)} - 1 \right]. \quad (A-41)
$$

To maximize (A-41), $k_0$ should be chosen as the largest $i$ such that $\lambda^{(i)} > 1$, i.e., $k_0 = \min \{k, k^*\}$. In addition, the optimal $\Phi_0$ is $\Phi_0 = [\phi^{(1)} \cdots \phi^{(k_0)}]$ for $Q \in \mathbb{R}^{k_0 \times k_0}$ with $Q^T Q = I_{k_0}$. Hence, we have

$$
\phi_i = \sum_{j=1}^{k_0} q_{ij} \phi^{(j)}.
$$

Following the same derivation as that for (A-12), we can
express $\phi^{(j)}$ as
\[
\frac{\phi^{(j)}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \frac{1}{\sqrt{\lambda(j)}} \sum_{i=1}^{d} f_i^{(j)}(x_i).
\]
and thus
\[
\frac{\phi_{u_k}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \frac{\phi_{u_{\ell}}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{i=1}^{d} f_i^{(j)}(x_i).
\]

Then, it follows from (A-35) that
\[
\frac{\phi_{u_k}(x^d)}{\sqrt{P_{X^d}(x^d)}} = h_{\ell}(u_{\ell}) \frac{\phi_{u_{\ell}}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \frac{1}{\sqrt{\lambda(j)}} \sum_{i=1}^{d} f_i^{(j)}(x_i) \quad (A-42)
\]
for $\ell = 1, \ldots, k_0$. Moreover, from (A-29e), we have
\[
\phi_{u_k} = \sum_{\ell=1}^{k} \phi_{u_{\ell}} = \sum_{\ell=1}^{k} \phi_{u_{\ell}}
\]
where the second equality follows from the consequence of (A-35) and (A-38) that $\phi_{u_{\ell}} = 0$ for $\ell > k_0$.

Therefore,
\[
\frac{\phi_{u_k}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{\ell=1}^{k_0} \phi_{u_{\ell}}(x^d) = \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i),
\]
which implies
\[
P_{X^d|U^k}(x^d|u^k)
= P_{X^d}(x^d) \left[ 1 + \epsilon \frac{\phi_{u_k}(x^d)}{\sqrt{P_{X^d}(x^d)}} \right] + o(\epsilon)
= P_{X^d}(x^d) \left[ 1 + \epsilon \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] + o(\epsilon)
\]
and
\[
P_{X^d|U^k}(x^d, u^k)
= P_{X^d}(x^d) \left[ \prod_{j=1}^{k_0} P_{U_{\ell}}(u_{\ell}) \right] \left[ 1 + \epsilon \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] + o(\epsilon).
\]

Finally, note that the exponential family $\mathcal{F}_{\text{exp},k}^{(j)}$, when $\delta$ is small, can be written as
\[
\mathcal{F}_{\text{exp},k}^{(j)} = \left\{ P_{X^d}(x^d) \prod_{j=1}^{k_0} P_{U_{\ell}}(u_{\ell}) \left[ 1 + \epsilon \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] \right\}
= \left\{ P_{X^d}(x^d) \left[ 1 + \epsilon \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] \right\}
= \left\{ P_{X^d}(x^d) \left[ 1 + \epsilon \sum_{\ell=1}^{k_0} h_{\ell}(u_{\ell}) \sum_{j=1}^{k_0} q_{j\ell} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] \right\}
\]

Since $\delta = \frac{1}{\epsilon^2}$, the proof is completed by comparing (A-44) and (A-43).

**APPENDIX D**

**JOINT CORRELATION MAXIMIZATION**

For functions $f_i : \mathcal{X}_i \rightarrow \mathbb{R}^{k}, i = 1, \ldots, d$, we define $\Psi_i \in \mathbb{R}^{|\mathcal{X}_i| \times k}$ such that the row vectors of $\Psi_i$ are $\sqrt{P_{X_i}(x_i)} f_i^T(x_i)$, for all $x_i \in \mathcal{X}_i$. Furthermore, we define the $m \times k$ matrix $\Psi$ as $\Psi = [\Psi^T_1 \cdots \Psi^T_d]$. Then the optimization problem (12) can be rewritten as
\[
\max_{\Psi : \Psi \in \mathbb{R}^{m \times k}} \text{tr} \{ \Psi^T B \Psi \} \quad (A-45a)
\]
subject to: $\Psi^T v_i = 0_k$, for all $i$ \quad (A-45b)
\[
\Psi^T \Psi = I_k, \quad (A-45c)
\]
where $0_k$ is the zero vector in $\mathbb{R}^k$, and $I_k$ is the $k \times k$ identity matrix. To establish the equivalence of (12) and (A-45), note that we have
\[
\Psi^T \Psi = \sum_{i=1}^{d} \sum_{j=1}^{d} P_{X_i}(x_i) f_i(x_i) f_j^T(x_i) = \sum_{i=1}^{d} \mathbb{E} \left[ f_i^T(x_i) f_i(x_i) \right]
= \mathbb{E} \left[ \sum_{i=1}^{d} f_i^T(x_i) f_i(x_i) \right]
\]
and
\[
\text{tr} \{ \Psi^T B \Psi \} = \sum_{i=1}^{d} \sum_{j=1}^{d} \text{tr} \{ \Psi^T B_{ij} \Psi \}
= \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{E} \left[ f_i^T(x_i) f_j^T(x_i) \right] + \sum_{i \neq j} \text{tr} \{ \mathbb{E} \left[ f_i^T(x_i) f_j^T(x_j) \right] \}
= \text{tr} \left\{ \sum_{i=1}^{d} \mathbb{E} \left[ f_i^T(x_i) f_i(x_i) \right] \right\} + \sum_{i \neq j} \text{tr} \{ \mathbb{E} \left[ f_i^T(x_i) f_j^T(x_j) \right] \}
= k + \mathbb{E} \left[ \sum_{i \neq j} f_i^T(x_i) f_j^T(x_j) \right].
\]

From Lemma 1, for $k < m - d$, the solution of (A-45) can be represented as $\Psi = [\Psi^{(1)} \cdots \Psi^{(k)}] Q$, where $Q \in \mathbb{R}^{k \times k}$ is an orthogonal matrix. Therefore, the optimal solution of (12) corresponds to $f_i^{(\ell)}$ with $i = 1, \ldots, d$ and $\ell = 1, \ldots, k$. 
APPENDIX E
COMMON BITS PATTERNS EXTRACTION

First, we define $\ell_{\text{max}}$ as the largest $\ell$ such that $w(J_\ell) > 0$, i.e., $\ell_{\text{max}} \triangleq \max\{\ell: 0 \leq \ell \leq 2^r - 1, w(J_\ell) > 0\}$. Then, $w(J_\ell) > 0$ is equivalent to $\ell \leq \ell_{\text{max}}$, and (14) can be equivalently expressed as

$$\lambda^{(\ell)} = w(J_\ell), \quad \ell \leq \ell_{\text{max}}, \quad (A-46)$$

and

$$\lambda^{(\ell)} = 0, \quad \ell > \ell_{\text{max}}. \quad (A-47)$$

Note that (7) establishes a one-to-one correspondence between the functions $f_i^{(\ell)} (i = 1, \ldots, d)$ and the vector $\psi^{(\ell)}$. With this correspondence, we use $\tilde{\psi}^{(\ell)}$ to denote the vector corresponding to the functions $f_i^{(\ell)}$ as defined in (15). Then the proof can be accomplished in two steps. First, we show that $\tilde{\psi}^{(\ell)} (\ell = 0, \ldots, \ell_{\text{max}})$ are $(\ell_{\text{max}} + 1)$ orthogonal eigenvectors of $B$ associated with eigenvalues $w(J_\ell) (\ell = 0, \ldots, \ell_{\text{max}})$, i.e., for all $0 \leq \ell \leq \ell_{\text{max}}$ and $0 \leq \ell' \leq \ell_{\text{max}}$, the $\tilde{\psi}^{(\ell)}$’s satisfy

$$B \tilde{\psi}^{(\ell)} = w(J_\ell) \tilde{\psi}^{(\ell)} \quad \text{and} \quad \langle \tilde{\psi}^{(\ell)}, \tilde{\psi}^{(\ell')} \rangle = \delta_{\ell \ell'}. \quad (A-48)$$

where $\delta_{\ell \ell'}$ is the Kronecker delta. Then, it suffices to verify that all other eigenvalues of $B$ are zeros [cf. (A-47)].

To begin, we equivalently express (A-48) using $f_i^{(\ell)}$ as

$$\sum_{j=1}^{d} \mathbb{E} \left[ f_j^{(\ell)} (X_i) \right] X_i = w(J_\ell) f_i^{(\ell)} (X_i), \quad 1 \leq i \leq d, \quad (A-49)$$

and

$$\sum_{i=1}^{d} \mathbb{E} \left[ f_i^{(\ell)} (X_i) f_i^{(\ell')} (X_i) \right] = \delta_{\ell \ell'}. \quad (A-50)$$

Then, since we have [cf. (13)]

$$\sum_{i=1}^{d} \mathbb{E} \left[ f_j^{(\ell)} (X_i) \right] X_i = \mathbb{E} \left[ f_j^{(\ell)} (X_i) \right], \quad 1 \leq i \leq d,$$

it suffices to show that

$$\mathbb{E} \left[ f_j^{(\ell)} (X_i) \right] = f_j^{(\ell)} (X_i), \quad 1 \leq i, j \leq d, \quad (A-51)$$

and

$$\mathbb{E} \left[ f_i^{(\ell)} (X_i) f_i^{(\ell')} (X_i) \right] = \frac{\mathbb{E} \left[ f_j^{(\ell)} (X_i) \right]}{w(J_\ell)} \cdot \delta_{\ell \ell'}, \quad 1 \leq i \leq d. \quad (A-52)$$

To obtain (A-51), note that if $J_\ell \not\subset I_j$, it follows from (15) that $f_j^{(\ell)} (X_i) = 0$, and thus (A-51) holds. Otherwise, we have $J_\ell \subset I_j$ and

$$\mathbb{E} \left[ f_j^{(\ell)} (X_i) \right] = \frac{1}{w(J_\ell)} \mathbb{E} \left[ \prod_{s \in J_\ell} b_s \right] X_i \quad (A-53)$$

Since $X_i = b_{I_j}$ is composed of all the $b_s$’s with indices in $I_j$, we have

$$\mathbb{E} \left[ \prod_{s \in J_\ell} b_s \right] X_i = \begin{cases} 1 & \text{if } J_\ell \subset I_j \\ 0 & \text{otherwise.} \end{cases} \quad (A-54)$$

Therefore, we obtain

$$\mathbb{E} \left[ f_j^{(\ell)} (X_i) \right] = \begin{cases} 1 & \text{if } J_\ell \subset I_j \\ 0 & \text{otherwise} \end{cases} \quad (A-55)$$

Likewise, (A-52) follows immediately from (15) when $\ell = \ell'$, and it suffices to consider the case $\ell \neq \ell'$ and prove that

$$\mathbb{E} \left[ f_i^{(\ell)} (X_i) f_i^{(\ell')} (X_i) \right] = 0. \quad (A-56)$$

Indeed, when $J_\ell \not\subset I_i$ or $J_{\ell'} \not\subset I_i$, (A-56) is trivially true. Otherwise, we have $J_\ell \subset I_i$ and $J_{\ell'} \subset I_i$, and it follows from (15) that

$$f_i^{(\ell)} (X_i) f_i^{(\ell')} (X_i) = \frac{1}{w(J_\ell) w(J_{\ell'})} \prod_{j \in J_\ell \triangle J_{\ell'}} b_j,$$

where “$\triangle$” denotes the symmetric difference of two sets, i.e., $A \triangle B = (A \setminus B) \cup (B \setminus A)$. Therefore, we have

$$\mathbb{E} \left[ f_i^{(\ell)} (X_i) f_i^{(\ell')} (X_i) \right] = \frac{1}{w(J_\ell) w(J_{\ell'})} \prod_{j \in J_\ell \triangle J_{\ell'}} \mathbb{E}[b_j] = 0,$$

where we have used the fact that the set $(J_\ell \triangle J_{\ell'})$ is non-empty, since $J_\ell \not\subset J_{\ell'}$.

Finally, to prove (A-47), i.e., eigenvalues other that $w(J_\ell)$ ($\ell = 0, \ldots, \ell_{\text{max}}$) are all zeros, note that

$$\sum_{\ell=0}^{\ell_{\text{max}}} w(J_\ell) = \sum_{\ell=0}^{2^r-1} w(J_\ell) = \sum_{\ell=0}^{2^r-1} w(I) \quad (A-57)$$

On the other hand, we have the sum of all eigenvalues

$$\sum_{\ell=0}^{m-1} \lambda^{(\ell)} = \text{tr} \{ B \} = m. \quad (A-58)$$

From Lemma 1, all eigenvalues of $B$ are non-negative, which implies (A-47).
To begin, we write the matrix $\tilde{B}$ of (6) as a block matrix

$$
\tilde{B} = \begin{bmatrix}
\tilde{B}_{11} & \tilde{B}_{12} & \cdots & \tilde{B}_{1d} \\
\tilde{B}_{21} & \tilde{B}_{22} & \cdots & \tilde{B}_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{B}_{d1} & \tilde{B}_{d2} & \cdots & \tilde{B}_{dd}
\end{bmatrix},
$$

(A-55)

where each block $\tilde{B}_{ij}$ is an $(|X_i| \times |X_j|)$-dimensional matrix. Then, we can rewrite $\|\tilde{B} - \Psi \Psi^T\|_F^2$ as

$$
\sum_{i=1}^{d} \sum_{j=1}^{d} \|\tilde{B}_{ij} - \Psi_i \Psi_j^T\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} \left( \|\tilde{B}_{ij}\|_F^2 - 2 \text{tr} \left( \Psi_i^T \tilde{B}_{ij} \Psi_j \right) + \|\Psi_i \Psi_j\|_F^2 \right) = \sum_{i=1}^{d} \sum_{j=1}^{d} \left( \|\tilde{B}_{ij}\|_F^2 - 2 \text{tr} \left( \Psi_i^T \tilde{B}_{ij} \Psi_j \right) + \|\Psi_i \Psi_j\|_F^2 \right)
$$

where we have used the fact that

$$
\text{tr} \left( \Psi_i^T \tilde{B}_{ij} \Psi_j \right) = \frac{1}{2} \|\Psi_i \Psi_j\|_F^2
$$

$$
= \mathbb{E} \left[ f_i^T(X_i) f_j(X_j) \right] - \left( \mathbb{E} \left[ f_i(X_i) \right] \right)^T \mathbb{E} \left[ f_j(X_j) \right] - \frac{1}{2} \text{tr} \left( \mathbb{E} \left[ f_i(X_i) f_i^T(X_i) \right] \mathbb{E} \left[ f_j(X_j) f_j^T(X_j) \right] \right)
$$

$$
= H \left( f_i(X_i), f_j(X_j) \right).
$$

References


