Appendix

9. Symmetrization

We presented the kernel algorithm for learning the multi-view latent variable model where the views have identical conditional distributions. In this section, we will extend it to the general case where the views are different. Without loss of generality, we will consider recover the operator $\mu_{X_3|h}$ for conditional distribution $\mathbb{P}(X_3|h)$. The same strategy applies to other views. The idea is to reduce the multi-view case to the identical-view case based on a method by (Anandkumar et al., 2012b).

Given the observations $\mathcal{D}_{X_1X_2X_3} = \{(x_1^i, x_2^i, x_3^i)\}_{i\in[m]}$ drawn *i.i.d.* from a multi-view latent variable model $\mathbb{P}(X_1, X_2, X_3)$, let the kernel matrix associated with X_1, X_2 and X_3 be K, L and G respectively and the corresponding feature map be ϕ, ψ and υ respectively. Furthermore, let the corresponding feature matrix be $\widetilde{\Phi} = (\phi(x_1^1), \ldots, \phi(x_1^m))$, $\widetilde{\Psi} = (\phi(x_2^1), \ldots, \phi(x_2^m))$ and $\widetilde{\Upsilon} = (\phi(x_3^1), \ldots, \phi(x_3^m))$. Then, we have the empirical estimation of the second/third-order embedding as

$$\widehat{\mathcal{C}}_{X_1X_2} = \frac{1}{m}\widetilde{\Phi}\widetilde{\Psi}^\top,\ \widehat{\mathcal{C}}_{X_3X_1} = \frac{1}{m}\widetilde{\Upsilon}\widetilde{\Phi}^\top,\ \widehat{\mathcal{C}}_{X_2X_3} = \frac{1}{m}\widetilde{\Psi}\widetilde{\Upsilon}^\top$$
$$\widehat{\mathcal{C}}_{X_1X_2X_3} := \frac{1}{m}\boldsymbol{I}_n \times_1 \widetilde{\Phi} \times_2 \widetilde{\Psi} \times_3 \widetilde{\Upsilon}$$

Find two arbitrary matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{k\times\infty}$, so that $\boldsymbol{A}\widehat{\mathcal{C}}_{X_1X_2}\boldsymbol{B}^\top$ is invertible. Theoretically, we could randomly select k columns from Φ and Ψ and set $\boldsymbol{A} = \Phi_k^\top, \boldsymbol{B} = \Psi_k^\top$. In practial, the first k leading eigenvector directions of respect *RKHS* works better. Then, we have

$$\begin{aligned}
\widetilde{\mathcal{C}}_{X_1X_2} &= \frac{1}{m}\widetilde{\Phi}_k^\top\widetilde{\Phi}\widetilde{\Psi}^\top\widetilde{\Psi}_k = \frac{1}{m}K_{nk}^\top L_{nk} \\
\widetilde{\mathcal{C}}_{X_3X_1} &= \widehat{\mathcal{C}}_{X_3X_1}\widetilde{\Phi}_k = \frac{1}{m}\widetilde{\Upsilon}K_{nk} \\
\widetilde{\mathcal{C}}_{X_3X_2} &= \widehat{\mathcal{C}}_{X_3X_2}\widetilde{\Psi}_k = \frac{1}{m}\widetilde{\Upsilon}L_{nk} \\
\widetilde{\mathcal{C}}_{X_1X_2X_3} &= \widehat{\mathcal{C}}_{X_1X_2X_3} \times_1 \widetilde{\Phi}_k^\top \times_2 \widetilde{\Psi}_k^\top = \frac{1}{m}\boldsymbol{I}_n \times_1 K_{nk}^\top \times_2 L_{nk}^\top \times_3 \widetilde{\Upsilon}
\end{aligned}$$

Based on these matrices, we could reduce to a single view

$$\begin{aligned}
Pair_3 &= \widetilde{\mathcal{C}}_{X_3X_1}(\widetilde{\mathcal{C}}_{X_1X_2}^\top)^{-1}\widetilde{\mathcal{C}}_{X_3X_2} \\
&= \frac{1}{m}\widetilde{\Upsilon}K_{nk}(L_{nk}^\top K_{nk})^{-1}L_{nk}^\top\widetilde{\Upsilon}^\top = \frac{1}{m}\widetilde{\Upsilon}H\widetilde{\Upsilon}^\top
\end{aligned}$$

where $H = K_{nk}(\mathcal{L}_{nk}^\top K_{nk})^{-1}L_{nk}^\top$.

Assume the leading k eigenvectors ν_k lie in the span of the column of Υ, i.e., $\nu_k = \Upsilon\beta_k$ where $\beta_k \in \mathbb{R}^{m\times 1}$

$$\begin{aligned}
Pair_3\nu = \lambda\nu &\Rightarrow (Pair_3)^\top Pair_3\nu = \lambda^2\nu \\
&\Rightarrow \frac{1}{m^2}\widetilde{\Upsilon}H^\top\widetilde{\Upsilon}^\top\widetilde{\Upsilon}H\widetilde{\Upsilon}^\top\nu = \lambda^2\nu \\
&\Rightarrow \frac{1}{m^2}\widetilde{\Upsilon}H^\top GHG\boldsymbol{\beta} = \lambda^2\widetilde{\Upsilon}\boldsymbol{\beta} \\
&\Rightarrow \frac{1}{m^2}GH^\top GHG\beta = \lambda^2 G\beta
\end{aligned}$$

Then, we symmetrize and whiten the third-order embedding

$$Triple_3 = \frac{1}{m}\widetilde{\mathcal{C}}_{X_1X_2X_3} \times_1 [\widetilde{\mathcal{C}}_{X_3X_2}\widetilde{\mathcal{C}}_{X_1X_2}^{-1}] \times_2 [\widetilde{\mathcal{C}}_{X_3X_1}\widetilde{\mathcal{C}}_{X_2X_1}^{-1}] \tag{13}$$

Plug $\widetilde{\mathcal{C}}_{X_3X_2}\widetilde{\mathcal{C}}_{X_1X_2}^{-1} = \widetilde{\Upsilon}L_{nk}(K_{nk}^\top L_{nk})^{-1}$ and $\widetilde{\mathcal{C}}_{X_3X_1}\widetilde{\mathcal{C}}_{X_2X_1}^{-1} = \widetilde{\Upsilon}K_{nk}(L_{nk}^\top K_{nk})^{-1}$, we have

$$Triple_3 = \frac{1}{m}\boldsymbol{I}_n \times_1 \widetilde{\Upsilon} L_{nk}(K_{nk}^\top L_{nk})^{-1} K_{nk}^\top \\ \times_2 \widetilde{\Upsilon} K_{nk}(L_{nk}^\top K_{nk})^{-1} L_{nk}^\top \times_3 \Upsilon$$

We multiply each mode with $\Upsilon\beta\widehat{S}_k^{-\frac{1}{2}}$ to whitening the data and apply power method to decompose it

$$\begin{aligned}\widehat{\mathcal{T}} &= Triple_3 \times_1 \widehat{S}_k^{-\frac{1}{2}}\beta^\top\widetilde{\Upsilon}^\top \times_2 \widehat{S}_k^{-\frac{1}{2}}\beta^\top\widetilde{\Upsilon}^\top \times_3 \widehat{S}_k^{-\frac{1}{2}}\beta^\top\widetilde{\Upsilon}^\top \\ &= \frac{1}{m}\boldsymbol{I}_n \times_1 \widehat{S}_k^{-\frac{1}{2}}\beta^\top G\mathcal{L}_{nk}(K_{nk}^\top L_{nk})^{-1}K_{nk}^\top \times_2 \\ &\quad \widehat{S}_k^{-\frac{1}{2}}\beta^\top G K_{nk}(L_{nk}^\top K_{nk})^{-1}L_{nk}^\top \times_3 \widehat{S}_k^{-\frac{1}{2}}\beta^\top G\end{aligned}$$

10. Robust Tensor Power Method

We recap the robust tensor power method for finding the tensor eigen-pairs in Algorithm 2, analyzed in detail in (Anandkumar et al., 2013a) and (Anandkumar et al., 2012a). The method computes the eigenvectors of a tensor through deflation, using a set of initialization vectors. Here, we employ random initialization vectors. This can be replaced with better initialization vectors, in certain settings, e.g. in the community model, the neighborhood vectors provide better initialization and lead to stronger guarantees (Anandkumar et al., 2013a). Given the initialization vector, the method then runs a tensor power update, and runs for N iterations to obtain an eigenvector. The successive eigenvectors are obtained via deflation.

Algorithm 2 $\{\lambda, M\} \leftarrow$TensorEigen$(\mathcal{T}, \{v_i\}_{i\in[k]}, N)$

Input: Tensor $\mathcal{T} \in \mathbb{R}^{k\times k\times k}$, set of k initialization vectors $\{v_i\}_{i\in[k]}$, number of iterations N.

Output: the estimated eigenvalue/eigenvector pairs $\{\lambda, M\}$, where $\lambda = (\lambda_1, \ldots, \lambda_k)^\top$ is the vector of eigenvalues and $M = (v_1, \ldots, v_k)$ is the matrix of eigenvectors.

for $i = 1$ to k **do**
  for $\tau = 1$ to k **do**
    $\theta_0 \leftarrow v_\tau$.
    for $t = 1$ to N **do**
      $\widetilde{\mathcal{T}} \leftarrow \mathcal{T}$.
      for $j = 1$ to $i-1$ (when $i > 1$) **do**
        if $|\lambda_j \left\langle \theta_t^{(\tau)}, v_j \right\rangle| > \xi$ **then**
          $\widetilde{\mathcal{T}} \leftarrow \widetilde{\mathcal{T}} - \lambda_j \phi_j^{\otimes 3}$.
        end if
      end for
      Compute power iteration update $\theta_t^{(\tau)} := \frac{\widetilde{\mathcal{T}}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})}{\|\widetilde{\mathcal{T}}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})\|}$
    end for
  end for
  Let $\tau^* := \arg\max_{\tau\in L}\{\widetilde{\mathcal{T}}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
  Do N power iteration updates starting from $\theta_N^{(\tau^*)}$ to obtain eigenvector estimate v_i, and set $\lambda_i := \widetilde{\mathcal{T}}(v_i, v_i, v_i)$.
end for
return the estimated eigenvalue/eigenvectors (λ, M).

11. Proof of Theorem 2

11.1. Recap of Perturbation Bounds for the Tensor Power Method

We now recap the result of Anandkumar et al. (2013a, Thm. 13) that establishes bounds on the eigen-estimates under good initialization vectors for the above procedure. Let $\mathcal{T} = \sum_{i\in[k]} \lambda_i v_i$, where v_i are orthonormal vectors and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$. Let $\widehat{\mathcal{T}} = \mathcal{T} + E$ be the perturbed tensor with $\|E\| \leq \epsilon_T$. Recall that N denotes the number of iterations of the tensor power method. We call an initialization vector u to be (γ, R_0)-good if there exists v_i such that $\langle u, v_i\rangle > R_0$ and

$|\langle u, v_i\rangle| - \max_{j<i} |\langle u, v_j\rangle| > \gamma |\langle u, v_i\rangle|$. Choose $\gamma = 1/100$.

Theorem 3 *There exists universal constants $C_1, C_2 > 0$ such that the following holds.*

$$\epsilon_T \leq C_1 \cdot \lambda_{\min} R_0^2, \qquad N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{\lambda_{\max}}{\epsilon_T}\right)\right), \tag{14}$$

Assume there is at least one good initialization vector corresponding to each v_i, $i \in [k]$. The parameter ξ for choosing deflation vectors in each iteration of the tensor power method in Procedure 2 is chosen as $\xi \geq 25\epsilon_T$. We obtain eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{v}_1), (\hat{\lambda}_2, \hat{v}_2), \ldots, (\hat{\lambda}_k, \hat{v}_k)$ such that there exists a permutation η on $[k]$ with

$$\|v_{\eta(j)} - \hat{v}_j\| \leq 8\epsilon_T/\lambda_{\eta(j)}, \qquad |\lambda_{\eta(j)} - \hat{\lambda}_j| \leq 5\epsilon_T, \quad \forall j \in [k],$$

and

$$\left\| \mathcal{T} - \sum_{j=1}^{k} \hat{\lambda}_j \hat{v}_j^{\otimes 3} \right\| \leq 55\epsilon_T.$$

In the sequel, we establish concentration bounds that allows us to translate the above condition on tensor perturbation (14) to sample complexity bounds.

11.2. Concentration Bounds

11.2.1. Analysis of Whitening

Recall that we use the covariance operator $\mathcal{C}_{X_1X_2}$ for whitening the 3rd order embedding $\mathcal{C}_{X_1,X_2,X_3}$. We first analyze the perturbation in whitening when sample estimates are employed.

Let $\widehat{\mathcal{C}}_{X_1X_2}$ denote the sample covariance operator between variables X_1 and X_2, and let

$$B := 0.5(\widehat{\mathcal{C}}_{X_1X_2} + \widehat{\mathcal{C}}_{X_1X_2}^\top) = \widehat{\mathcal{U}}\widehat{S}\widehat{\mathcal{U}}^\top$$

denote the SVD. Let $\widehat{\mathcal{U}}_k$ and $\widehat{S}_k$ denote the restriction to top-k eigen-pairs, and let $B_k := \widehat{\mathcal{U}}_k\widehat{S}_k\widehat{\mathcal{U}}_k^\top$. Recall that the whitening matrix is given by $\widehat{\mathcal{W}} := \widehat{\mathcal{U}}_k\widehat{S}_k^{-1/2}$. Now $\widehat{\mathcal{W}}$ whitens B_k, i.e. $\widehat{\mathcal{W}}^\top B_k \widehat{\mathcal{W}} = I$.

Now consider the SVD of

$$\widehat{\mathcal{W}}^\top \mathcal{C}_{X_1X_2}\widehat{\mathcal{W}} = ADA^\top,$$

and define

$$\mathcal{W} := \widehat{\mathcal{W}}AD^{-1/2}A^\top,$$

and $\mathcal{W}$ whitens $\mathcal{C}_{X_1X_2}$ since $\mathcal{W}^\top \mathcal{C}_{X_1X_2} W = I$. Recall that by exchangeability assumption,

$$\mathcal{C}_{X_1,X_2} = \sum_{j=1}^{k} \pi_j \cdot \mu_{X|j} \otimes \mu_{X|j} = M \operatorname{Diag}(\pi) M^\top, \tag{15}$$

where the j^{th} column of M, $M_j = \mu_{X|j}$.

We now establish the following perturbation bound on the whitening procedure. Recall from (25), $\epsilon_{pairs} := \left\| \mathcal{C}_{X_1,X_2} - \widehat{\mathcal{C}}_{X_1,X_2} \right\|$. Let $\sigma_1(\cdot) \geq \sigma_2(\cdot) \ldots$ denote the singular values of an operator.

Lemma 4 (Whitening perturbation) *Assuming that $\epsilon_{pairs} < 0.5\sigma_k(\mathcal{C}_{X_1X_2})$,*

$$\epsilon_W := \|\operatorname{Diag}(\pi)^{1/2} M^\top (\widehat{\mathcal{W}} - \mathcal{W})\| \leq \frac{4\epsilon_{pairs}}{\sigma_k(\mathcal{C}_{X_1X_2})} \tag{16}$$

Remark: Note that $\sigma_k(\mathcal{C}_{X_1X_2}) = \sigma_k^2(M)$.

Proof: The proof is along the lines of Lemma 16 of (Anandkumar et al., 2013a), but adapted to whitening using the covariance operator here.

$$\begin{aligned} \|\operatorname{Diag}(\pi)^{1/2} M^\top (\widehat{\mathcal{W}} - \mathcal{W})\| &= \|\operatorname{Diag}(\pi)^{1/2} M^\top W (AD^{1/2}A^\top - I)\| \\ &\leq \|\operatorname{Diag}(\pi)^{1/2} M^\top \mathcal{W}\| \|D^{1/2} - I\|. \end{aligned}$$

Since $\mathcal{W}$ whitens $\mathcal{C}_{X_1X_2} = M\,\mathrm{Diag}(\pi)M^\top$, we have that $\|\mathrm{Diag}(\pi)^{1/2}M^\top\mathcal{W}\| = 1$. Now we control $\|D^{1/2} - I\|$. Let $\widetilde{E} := \mathcal{C}_{X_1,X_2} - B_k$, where recall that $B = 0.5(\widehat{\mathcal{C}}_{X_1,X_2} + \widehat{\mathcal{C}}_{X_1X_2}^\top)$ and B_k is its restriction to top-k singular values. Thus, we have $\|\widetilde{E}\| \le \epsilon_{pairs} + \sigma_{k+1}(B) \le 2\epsilon_{pairs}$. We now have

$$\begin{aligned}\|D^{1/2} - I\| &\le \|(D^{1/2} - I)(D^{1/2} + I)\| \le \|D - I\| \\ &= \|ADA^\top - I\| = \|\widehat{\mathcal{W}}^\top\mathcal{C}_{X_1X_2}\widehat{\mathcal{W}} - I\| \\ &= \|\widehat{\mathcal{W}}^\top\widetilde{E}\widehat{\mathcal{W}}\| \le \|\widehat{\mathcal{W}}\|^2(2\epsilon_{pairs}).\end{aligned}$$

Now

$$\|\widehat{\mathcal{W}}^2\| \le \frac{1}{\sigma_k(\widehat{\mathcal{C}}_{X_1X_2})} \le \frac{2}{\sigma_k(\mathcal{C}_{X_1X_2})},$$

when $\epsilon_{pairs} < 0.5\sigma_k(\mathcal{C}_{X_1X_2})$. □

11.2.2. Tensor Concentration Bounds

Recall that the whitened tensor from samples is given by

$$\widehat{\mathcal{T}} := \widehat{\mathcal{C}}_{X_1X_2X_3} \times_1 (\widehat{\mathcal{W}}^\top) \times_2 (\widehat{\mathcal{W}}^\top) \times_3 (\widehat{\mathcal{W}}^\top).$$

We want to establish its perturbation from the whitened tensor using exact statistics

$$\mathcal{T} := \mathcal{C}_{X_1X_2X_3} \times_1 (\mathcal{W}^\top) \times_2 (\mathcal{W}^\top) \times_3 (\mathcal{W}^\top).$$

Further, we have

$$\mathcal{C}_{X_1X_2X_3} = \sum_{h\in[k]} \pi_h \cdot \mu_{X|h} \otimes \mu_{X|h} \otimes \mu_{X|h} \tag{17}$$

Let $\epsilon_{triples} := \|\widehat{\mathcal{C}}_{X_1X_2X_3} - \mathcal{C}_{X_1X_2X_3}\|$. Let $\pi_{\min} := \min_{h\in[k]} \pi_h$.

Lemma 5 (Tensor perturbation bound) *Assuming that $\epsilon_{pairs} < 0.5\sigma_k(\mathcal{C}_{X_1X_2})$, we have*

$$\epsilon_T := \|\widehat{\mathcal{T}} - \mathcal{T}\| \le \frac{2\sqrt{2}\epsilon_{triples}}{\sigma_k(\mathcal{C}_{X_1X_2})^{1.5}} + \frac{\epsilon_W^3}{\sqrt{\pi_{\min}}}. \tag{18}$$

Proof: Define intermediate tensor

$$\widetilde{\mathcal{T}} := \mathcal{C}_{X_1X_2X_3} \times_1 (\widehat{\mathcal{W}}^\top) \times_2 (\widehat{\mathcal{W}}^\top) \times_3 (\widehat{\mathcal{W}}^\top).$$

We will bound $\|\widehat{\mathcal{T}} - \widetilde{\mathcal{T}}\|$ and $\|\widehat{\mathcal{T}} - \mathcal{T}\|$ separately.

$$\|\widehat{\mathcal{T}} - \widetilde{\mathcal{T}}\| \le \|\widehat{\mathcal{C}}_{X_1,X_2,X_2} - \mathcal{C}_{X_1,X_2,X_3}\|\|\widehat{\mathcal{W}}\|^3 \le \frac{2\sqrt{2}\epsilon_{triples}}{\sigma_k(\mathcal{C}_{X_1X_2})^{1.5}},$$

using the bound on $\|\widehat{\mathcal{W}}\|$ in Lemma 4. For the other term, first note that

$$\mathcal{C}_{X_1,X_2,X_3} = \sum_{h\in[k]} \pi_h \cdot M_h \otimes M_h \otimes M_h,$$

$$\begin{aligned}\|\widehat{\mathcal{T}} - \mathcal{T}\| &= \|\mathcal{C}_{X_1X_2X_3} \times_1 (\widehat{\mathcal{W}} - \mathcal{W})^\top \times_2 (\widehat{\mathcal{W}} - \mathcal{W})^\top \times_3 (\widehat{\mathcal{W}} - \mathcal{W})^\top\| \\ &\le \frac{\|\mathrm{Diag}(\pi)^{1/2}M^\top(\widehat{\mathcal{W}} - \mathcal{W})\|^3}{\sqrt{\pi_{\min}}} \\ &= \frac{\epsilon_W^3}{\sqrt{\pi_{\min}}}\end{aligned}$$

□

Proof of Theorem 2: We obtain a condition on the above perturbation ϵ_T in (18) by applying Theorem 3 as $\epsilon_T \le C_1\lambda_{\min}R_0^2$. Here, we have $\lambda_i = 1/\sqrt{\pi_i} \ge 1$. For random initialization, we have that $R_0 \sim 1/\sqrt{k}$, with probability $1-\delta$ using $\mathrm{poly}(k)\,\mathrm{poly}(1/\delta)$ trials, see Thm. 5.1 in (Anandkumar et al., 2012a). Thus, we require that $\epsilon_T \le \frac{C_1}{k}$. Summarizing,

we require for the following conditions to hold

$$\epsilon_{pairs} \leq 0.5\sigma_k(\mathcal{C}_{X_1X_2}), \quad \epsilon_T \leq \frac{C_1}{k}. \tag{19}$$

We now substitute for ϵ_{pairs} and $\epsilon_{triples}$ in (18) using Lemma 6 and Lemma 7.

From Lemma 6, we have that

$$\epsilon_{pairs} \leqslant \frac{2\sqrt{2}\rho\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}},$$

with probability $1-\delta$. It is required that $\epsilon_{pairs} < 0.5\sigma_k(\mathcal{C}_{X_1,X_2})$, which yields that

$$m > \frac{32\rho^2\log\frac{2}{\delta}}{\sigma_k^2(\mathcal{C}_{X_1,X_2})}. \tag{20}$$

Further we require that $\epsilon_T \leq C_1/k$, which implies that each of the terms in (18) is less than C/k, for some constant C. Thus, we have

$$\frac{2\sqrt{2}\epsilon_{triples}}{\sigma_k^{1.5}(\mathcal{C}_{X_1,X_2})} < \frac{C}{k} \quad \Rightarrow \quad m > \frac{C_3k^2\rho^3\log\frac{2}{\delta}}{\sigma_k^3(\mathcal{C}_{X_1,X_2})},$$

for some constant C_3 with probability $1-\delta$ from Lemma 7. Similarly for the second term in (18), we have

$$\frac{\epsilon_W^3}{\sqrt{\pi_{\min}}} < \frac{C}{k},$$

and from Lemma 4, this implies that

$$\epsilon_{pairs} \leq \frac{C'\pi_{\min}^{1/6}\sigma_k(\mathcal{C}_{X_1,X_2})}{k^{1/3}},$$

Thus, we have

$$m > \frac{C_4k^{\frac{2}{3}}\rho^2\log\frac{2}{\delta}}{\pi_{\min}^{\frac{1}{3}}\sigma_k^2(\mathcal{C}_{X_1,X_2})},$$

for some other constant C_4 with probability $1-\delta$. Thus, we have the result in Theorem 2.

□

11.2.3. Concentration bounds for Empirical Operators

Concentration results for the singular value decomposition of empirical operators.

Lemma 6 (Concentration bounds for pairs) *Let $\rho := \sup_{x\in\Omega} k(x,x)$, and $\|\cdot\|$ be the Hilbert-Schmidt norm, we have for*

$$\epsilon_{pairs} := \left\|\mathcal{C}_{X_1X_2} - \widehat{\mathcal{C}}_{X_1X_2}\right\|, \tag{21}$$

$$\Pr\left\{\epsilon_{pairs} \leqslant \frac{2\sqrt{2}\rho\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}}\right\} \geqslant 1-\delta. \tag{22}$$

Proof We will use similar arguments as in (Rosasco et al., 2010) which deals with symmetric operator. Let ξ_i be defined as

$$\xi_i \;=\; \phi(x_1^i)\otimes\phi(x_2^i) - \mathcal{C}_{X_1,X_2}. \tag{23}$$

It is easy to see that $\mathbb{E}[\xi_i] = 0$. Further, we have

$$\sup_{x_1,x_2}\|\phi(x_1)\otimes\phi(x_2)\|^2 = \sup_{x_1,x_2} k(x_1,x_1)k(x_2,x_2) \leqslant \rho^2, \tag{24}$$

which implies that $\|\mathcal{C}_{X_1X_2}\| \leqslant \rho$, and $\|\xi_i\| \leqslant 2\rho$. The result then follows from the Hoeffding's inequality in Hilbert space. ■

Similarly, we have the concentration bound for 3rd order embedding.

Lemma 7 (Concentration bounds for triples) *Let $\rho := \sup_{x\in\Omega} k(x,x)$, and $\|\cdot\|$ be the Hilbert-Schmidt norm, we have for*

$$\epsilon_{triples} := \left\| \mathcal{C}_{X_1X_2X_3} - \widehat{\mathcal{C}}_{X_1X_2X_3} \right\|, \tag{25}$$

$$\Pr\left\{ \epsilon_{triples} \leqslant \frac{2\sqrt{2}\rho^{3/2}\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}} \right\} \geqslant 1-\delta. \tag{26}$$

Proof We will use similar arguments as in (Rosasco et al., 2010) which deals with symmetric operator. Let ξ_i be defined as

$$\xi_i \;=\; \phi(x_1^i)\otimes\phi(x_2^i)\otimes\phi(x_3^i) - \mathcal{C}_{X_1X_2X_3}. \tag{27}$$

It is easy to see that $\mathbb{E}[\xi_i] = 0$. Further, we have

$$\sup_{x_1,x_2,x_3} \|\phi(x_1)\otimes\phi(x_2)\otimes\phi(x_3)\|^2 = \sup_{x_1,x_2,x_3} k(x_1,x_1)k(x_2,x_2)k(x_3,x_3) \leqslant \rho^3, \tag{28}$$

which implies that $\|\mathcal{C}_{X_1X_2X_3}\| \leqslant \rho^{3/2}$, and $\|\xi_i\| \leqslant 2\rho^{3/2}$. The result then follows from the Hoeffding's inequality in Hilbert space. ■

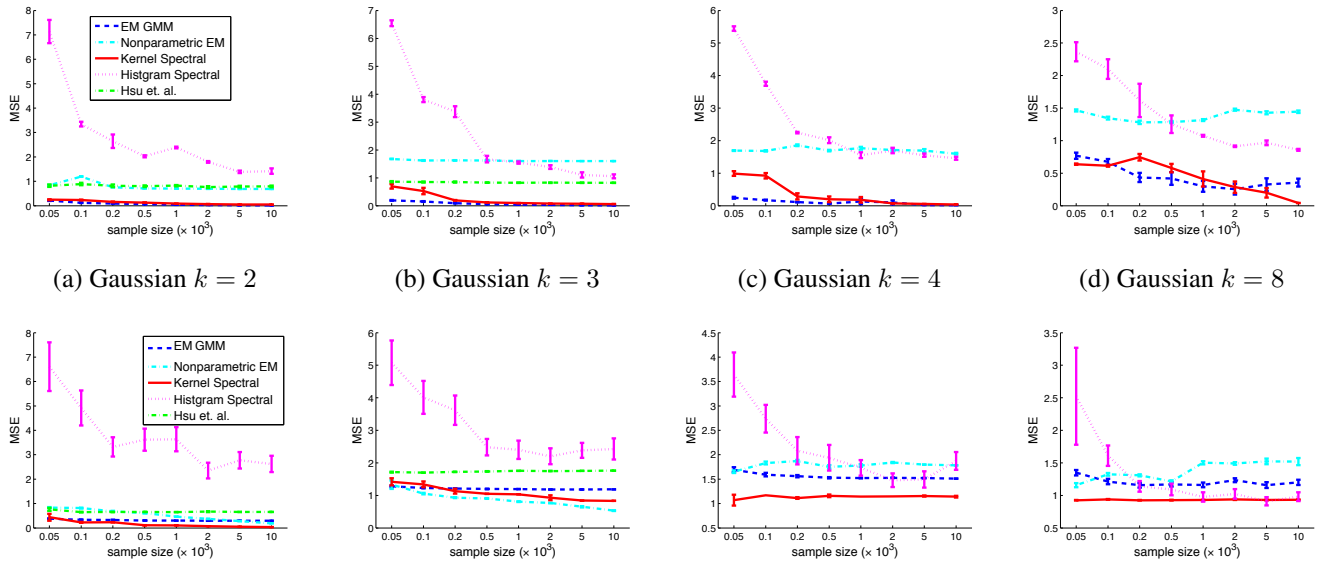
12. Experiment on Single Conditional Distribution

We also did some experiments for three-dimensional synthetic data that each view has the same conditional distribution. We generated the data from two settings:

1. Mixture of Gaussian conditional density;
2. Mixture of Gaussian and shifted Gamma conditional density.

The mixture proportion and other experiment settings are exact same as the experiment in the main text. The only difference is that the conditional densities for each view here are the identical. We use the same measure to evaluate the performance. The empirical results are plotted in Figure 5.

As we expected, the behavior of the proposed method is similar to the results in different conditional densities case. In mixture of Gaussians, our algorithm converges to the EM GMM resuls. And in the mixture of Gaussian/shift Gamma, our algorithm consistently better to other alternatives in most cases, except $k = 3$ where our method achieve comparable to nonparametric EM algorithm.



(a) Gaussian $k = 2$ (b) Gaussian $k = 3$ (c) Gaussian $k = 4$ (d) Gaussian $k = 8$

(e) Gaussian/Gamma $k = 2$ (f) Gaussian/Gamma $k = 3$ (g) Gaussian/Gamma $k = 4$ (h) Gaussian/Gamma $k = 8$

Figure 5. (a)-(d) Mixture of Gaussian distributions with $k = 2, 3, 4, 8$ components. (e)-(h) Mixture of Gaussian/Gamma distribution with $k = 2, 3, 4, 8$. For the former case, the performance of kernel spectral algorithm converge to those of EM algorithm for mixture of Gaussian model. For both cases, the performance of kernel spectral algorithm are consistently the best or comparable. Spherical Gaussian spectral algorithm does not work for $k = 4, 8$, and hence not plotted.