Nonparametric Estimation of Multi-view Latent Variable Models

Abstract

In this paper, we propose a nonparametric kernel method for estimating the parameters of the multi-view latent variable models.

1. Introduction

Recent interest in spectral algorithms for latent variable model: fast and provable guarantee

However, most methods only work for discrete and Gaussian algorithm.

We propose an algorithm to estimate multi-view latent variables where the conditional density can be nonparametric.

We will kernel embedding of distributions, covariance operators and tensor power methods.

We provide both theoretical guarantee for our method and empirical evidence of the method.

2. Notation

We denote by X a random variable with domain $\mathcal{X}$, and refer to instantiations of X by the lower case character, x. We endow $\mathcal{X}$ with some σ-algebra $\mathscr{A}$ and denote a distributions (with respect to $\mathscr{A}$) on $\mathcal{X}$ by $\mathbb{P}(X)$. We will also deal with multiple random variables, $X_1, X_2, \ldots, X_\ell$, with joint distribution $\mathbb{P}(X_1, X_2, \ldots, X_\ell)$. For simplicity of notation, we assume that the domains of all $X_t, t \in [\ell]$ are the same, but the methodology applies to the cases where they have different domains. Furthermore, we denote by H hidden variables with domain $\mathcal{H}$ and distribution $\mathbb{P}(H)$.

A *reproducing kernel Hilbert space (RKHS)* $\mathcal{F}$ on $\mathcal{X}$ with a kernel $k(x, x')$ is a Hilbert space of functions $f(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Its element $k(x, \cdot)$ satisfies the reproducing property: $\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{F}} = f(x)$, and consequently, $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{F}} = k(x, x')$, meaning that we can view the evaluation of a function f at any point $x \in \mathcal{X}$ as an inner product. Alternatively, $k(x, \cdot)$ can be viewed as an implicit feature map $\phi(x)$ where $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$. Popular kernel functions on $\mathbb{R}^n$ include the polynomial kernel $k(x, x') = (\langle x, x' \rangle + c)^d$ and the Gaussian RBF kernel $k(x, x') = \exp(-s \|x - x'\|^2)$. Kernel functions have also been defined on graphs, time series, dynamical systems, images and other structured objects (Schölkopf et al., 2004). Thus the methodology presented below can readily be generalized to a diverse range of data types as long as kernel functions are defined for them. Similarly, we denote by $\mathcal{G}$ an RKHS on $\mathcal{H}$ with kernel $l(h, h')$, and by $\psi(h)$ the corresponding feature map.

3. Kernel Embedding of Distributions

We begin by providing an overview of kernel embeddings of distributions, which are *implicit* mappings of distributions into potentially *infinite* dimensional RKHS.[1] The kernel embedding approach represents a distribution by an element in the RKHS associated with a kernel function (Smola et al., 2007; Sriperumbudur et al., 2008),

$$\mu_X := \mathbb{E}_X [\phi(X)] = \int_{\mathcal{X}} \phi(x)\, \mathbb{P}(dx), \tag{1}$$

where the distribution is mapped to its expected feature map, *i.e.*, to a point in a potentially infinite-dimensional and implicit feature space. The kernel embedding μ_X has the property that the expectation of any RKHS function f can be evaluated as an inner product in $\mathcal{F}$, $\mathbb{E}_X[f(X)] = \langle \mu_X, f \rangle_{\mathcal{F}}, \forall f \in \mathcal{F}$.

Kernel embeddings can be readily generalized to joint distributions of two or more variables using tensor product feature maps. For instance, we can embed a joint distribution of two variables X_1 and X_2 into a tensor product feature space $\mathcal{F} \times \mathcal{F}$ by

$$\begin{aligned} \mathcal{C}_{X_1 X_2} &:= \mathbb{E}_{X_1 X_2}[\phi(X_1) \otimes \phi(X_2)] & (2) \\ &= \int_{\mathcal{X} \times \mathcal{X}} \phi(x_1) \otimes \phi(x_2)\, \mathbb{P}(dx_1 \times dx_2) & (3) \end{aligned}$$

where the reproducing kernel for the tensor product features satisfies $\langle \phi(x_1) \otimes \phi(x_2), \phi(x_1') \otimes \phi(x_2') \rangle_{\mathcal{F} \times \mathcal{F}} = k(x_1, x_1')\, k(x_2, x_2')$.

Kernel embedding of distributions have both rich representational power. The mapping is injective for characteristic kernels (Sriperumbudur et al., 2008). That is, if two

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] By "implicit", we mean that we do not need to explicitly construct the feature spaces, and the actual computations boil down to kernel matrix operations.

distributions, $\mathbb{P}(X)$ and $\mathbb{Q}(X)$, are different, they will be mapped to two distinct points in the RKHS. For domain $\mathbb{R}^d$, many commonly used kernels are characteristic, such as the Gaussian RBF kernel $\exp(-\sigma\|x-x'\|^2)$ and Laplace kernel $\exp(-\sigma\|x-x'\|)$. This injective property of kernel embeddings has been exploited to design state-of-the-art two-sample tests (Gretton et al., 2012) and independence tests (Gretton et al., 2008).

Kernel embeddings as multilinear operators. The joint embeddings can also be viewed as an uncentered cross-covariance operator $\mathcal{C}_{X_1X_2} : \mathcal{F} \mapsto \mathcal{F}$ by the standard equivalence between a tensor product feature and a linear map. That is, given two functions $f_1, f_2 \in \mathcal{F}$, their covariance can be computed by $\mathbb{E}_{X_1X_2}[f_1(X_1)f_2(X_2)] = \langle f_1, \mathcal{C}_{X_1X_2} f_2\rangle_{\mathcal{F}}$, or equivalently $\langle f_1 \otimes f_2, \mathcal{C}_{X_1X_2}\rangle_{\mathcal{F}\times\mathcal{F}}$, where in the former we view $\mathcal{C}_{XY}$ as an operator while in the latter we view it as an element in tensor product feature space. By analogy, $\mathcal{C}_{X_1X_2X_3} := \mathbb{E}_{X_1X_2X_3}[\phi(X_1)\otimes\phi(X_2)\otimes\phi(X_3)]$ can also be defined, which can be regarded as a multi-linear operator from $\mathcal{F}\times\mathcal{F}\times\mathcal{F}$ to $\mathbb{R}$. It will be clear from the context whether we use $\mathcal{C}_{XY}$ as an operator between two spaces or as an element from a tensor product feature space.

More generally, the kernel embedding $\mathcal{C}_{X_{1:\ell}}$ for a joint distribution $\mathbb{P}(X_1, X_2, \ldots, X_\ell)$ can be viewed as a multilinear operator (tensor) of order ℓ mapping from $\mathcal{F}\times\ldots\times\mathcal{F}$ to $\mathbb{R}$. (For generic introduction to tensor and tensor notation, please see (Kolda & Bader, 2009)). The operator is linear in each argument (mode) when fixing other arguments. Furthermore, the application of the operator to a set of elements $\{f_i \in \mathcal{F}\}_{i\in[\ell]}$ can be defined using the inner product from the tensor product feature space, *i.e.*,

$$\mathcal{C}_{X_{1:\ell}} \times_1 f_1 \times_2 \ldots \times_\ell f_\ell := \langle \mathcal{C}_{X_{1:\ell}},\ f_1 \otimes \ldots \otimes f_\ell\rangle_{\mathcal{F}^d} = \mathbb{E}_{X_{1:\ell}}\left[\prod_{i\in[\ell]} \langle \phi(X_i), f_i\rangle_{\mathcal{F}}\right], \tag{4}$$

where $\times_i$ means applying f_i to the i-th argument of $\mathcal{C}_{X_{1:\ell}}$. Furthermore, we can define the Hilbert-Schmidt norm $\|\cdot\|$ of $\mathcal{C}_{X_{1:\ell}}$ as

$$\|\mathcal{C}_{X_{1:\ell}}\|^2 = \sum_{i_1=1}^{\infty}\cdots\sum_{i_\ell=1}^{\infty}(\mathcal{C}_{X_{1:\ell}} \times_1 u_{i_1} \times_2 \ldots \times_\ell u_{i_\ell})^2$$

using a collection of orthonormal bases $\{u_{i_1}\}_{i_1=1}^{\infty}, \ldots, \{u_{i_\ell}\}_{i_\ell=1}^{\infty}$. We can also define the inner product for the space of such operator that $\|\mathcal{C}_{X_{1:\ell}}\| < \infty$ as

$$\left\langle \mathcal{C}_{X_{1:\ell}}, \widetilde{\mathcal{C}}_{X_{1:\ell}}\right\rangle = \sum_{i_1=1}^{\infty}\cdots\sum_{i_\ell=1}^{\infty}(\mathcal{C}_{X_{1:\ell}} \times_1 u_{i_1} \times_2 \ldots \times_\ell u_{i_\ell})\left(\widetilde{\mathcal{C}}_{X_{1:\ell}} \times_1 u_{i_1} \times_2 \ldots \times_\ell u_{i_\ell}\right). \tag{5}$$

The joint embedding, $\mathcal{C}_{X_1X_2}$, is a 2nd order tensor, and we can essentially use notation and operations for matrices. For instance, we can perform singular value decomposition

$$\mathcal{C}_{X_1X_2} = \sum_{i=1}^{\infty}\sigma_i \cdot u_{i_1} \otimes u_{i_2}$$

where $\sigma_i \in \mathbb{R}$ are singular values ordered in nonincreasing manner, and $\{u_{i_1}\}_{i_1=1}^{\infty} \subset \mathcal{F}, \{u_{i_2}\}_{i_2=1}^{\infty} \subset \mathcal{F}$ are singular vectors and orthonormal bases. The rank of $\mathcal{C}_{X_1X_2}$ is the smallest k such that $\sigma_i = 0$ for $i > k$.

Example 1. The probability vector of a discrete variable $X \in [n]$, and the joint probability table of two discrete variables $X_1 \in [n]$ and $X_2 \in [n]$, are both kernel embeddings. To see this, let the kernel be the Kronecker delta kernel $k(x, x') = \delta(x, x')$. The corresponding feature map $\phi(x)$ is then the standard basis of e_x in $\mathbb{R}^n$. Then

$$\mu_X = \mathbb{E}_X[e_X] = \begin{pmatrix} \mathbb{P}(x=1) \\ \vdots \\ \mathbb{P}(x=n)\end{pmatrix},\ \mathcal{C}_{X_1X_2} = \mathbb{E}_{X_1X_2}[e_{X_1} \otimes e_{X_2}] = \Bigg(\ \mathbb{P}(x \tag{6}$$

Finite sample estimate. While we rarely have access to the true underlying distribution, $\mathbb{P}(X)$, we can readily estimate its embedding using a finite sample average. Given a sample $\mathcal{D}_X = \{x^1, \ldots, x^m\}$ of size m drawn *i.i.d.* from $\mathbb{P}(X)$, the empirical kernel embedding is

$$\widehat{\mu}_X := \frac{1}{m}\sum\nolimits_{i=1}^{m}\phi(x^i). \tag{7}$$

This empirical estimate converges to its population counterpart in RKHS norm, $\|\widehat{\mu}_X - \mu_X\|_{\mathcal{F}}$, with a rate of $O_p(m^{-\frac{1}{2}})$ (Smola et al., 2007). We note that this rate is independent of the dimension of X, meaning that statistics based on kernel embeddings circumvent the curse of dimensionality.

Kernel embeddings of joint distributions inherit the previous two properties of general embeddings: injectivity and easy empirical estimation. Given m pairs of training examples $\mathcal{D}_{XY} = \{(x_1^i, x_2^i)\}_{i\in[m]}$ drawn *i.i.d.* from $\mathbb{P}(X_1, X_2)$, the covariance operator can then be estimated as

$$\widehat{\mathcal{C}}_{X_1X_2} = \frac{1}{m}\sum_{i=1}^{m}\phi(x_1^i)\otimes\phi(x_2^i). \tag{8}$$

By virtue of the kernel trick, most of the computation required for statistical inference using kernel embeddings can be reduced to the Gram matrix manipulation. The entries in the Gram matrix K correspond to the kernel value between data points x^i and x^j, *i.e.*, $K_{ij} = k(x^i, x^j)$, and therefore its size is determined by the number of data points in the sample. The size of the Gram matrix is in general much smaller than the dimension of the feature spaces (which can be infinite). This enables efficient nonparametric methods using the kernel embedding representation. If the sample size is large, the computation in kernel embedding methods may be expensive. In this case, a popular solution is to use a low-rank approximation of the Gram matrix, such as incomplete Cholesky factorization (Fine & Scheinberg, 2001), which is known to work very effectively in reducing computational cost of kernel methods, while maintaining the approximation accuracy.

Relation between kernel embedding and the density function. Basically kernel embeddings maps the density to a function in the RKHS. See Steinwardt and Vert (need to say more).

4. Kernel Embeddings of Conditional Distributions

The kernel embedding of a conditional distribution $\mathbb{P}(X|h)$ is defined as (Song et al., 2009)

$$\mu_{X|h} := \mathbb{E}_{X|h}[\phi(X)] = \int_{\mathcal{X}} \phi(x)\, \mathbb{P}(dx|h). \tag{9}$$

Given this embedding, the conditional expectation of a function $f \in \mathcal{F}$ can be computed as $\mathbb{E}_{X|h}[f(X)] = \left\langle f, \mu_{X|h} \right\rangle_{\mathcal{F}}$. This may be compared with the property of the mean embedding in the last section, where the *unconditional* expectation of a function may be written as an inner product with the embedding. Unlike the embeddings discussed in the previous section, an embedding of conditional distribution is not a single element in the RKHS, but will instead sweep out a family of points in the RKHS, each indexed by a fixed value h of the conditioning variable H. It is only by fixing H to a particular value h, that we will be able to obtain a single RKHS element, $\mu_{X|h} \in \mathcal{F}$. In other words, we need to define an operator, denoted by $\mathcal{C}_{X|H}$, which can take as input an h and output an embedding. More specifically, we will want it to satisfy

$$\mu_{X|h} = \mathcal{C}_{X|H}\psi(h). \tag{10}$$

Based on the relation between conditional expectation and covariance operators, Song et al. (Song et al., 2009) show that, under the assumption $\mathbb{E}_{X|\cdot}[f(X)] \in \mathcal{G}$,

$$\mathcal{C}_{X|H} := \mathcal{C}_{XH}\mathcal{C}_{HH}^{-1}, \tag{11}$$

satisfy the requirement in (10), and hence $\mu_{X|h} = \mathcal{C}_{XH}\mathcal{C}_{HH}^{-1}\psi(h)$. We remark that the assumption $\mathbb{E}_{X|\cdot}[f(X)] \in \mathcal{G}$ always holds for finite domains with characteristic kernels, but it is not necessarily true for continuous domains (Fukumizu et al., 2004). In practice, the inversion of the operator can be replaced by the regularized inverse $(\mathcal{C}_{HH} + \lambda I)^{-1}$.

Example 2. The definition of the conditional embedding operator in (11) is very general, and the conditional probability for discrete variables becomes a special case. We again use a Kronecker delta kernel for both $x \in [n]$ and $h \in [k]$, and a feature space construction similar to (6). We obtain

$$\underbrace{\Big(\ \mathbb{P}(x=s|h=t)\ \Big)_{s\in[n],t\in[k]}}_{\mathcal{C}_{X|H}} = \underbrace{\Big(\ \mathbb{P}(x=s,h=t)\ \Big)_{s\in[n],t\in[k]}}_{\mathcal{C}_{XH}} \Big(\ \mathbb{P}(h \tag{12}$$

Example 3. As another example, let X from a general domain $\mathcal{X}$ while $H \in [k]$ being discrete. Then, there are k different conditional distributions, $\mathbb{P}(X|h=t)$, $t \in [k]$, one for each value of the discrete conditioning variable H. Using Kronecker delta kernel for H, the conditional embedding operator is simply a column concatenation of the embeddings for each $\mathbb{P}(X|h)$, *i.e.*,

$$\mathcal{C}_{X|H} = \left(\mu_{X|h=1},\ \mu_{X|h=2},\ \ldots,\ \mu_{X|h=k}\right), \text{ and } \mu_{X|h} = \mathcal{C}_{X|H}e_h. \tag{13}$$

k-restricted conditional embedding operator. Let $\{u_i\}_{i=1}^{\infty}$ and $\{v_i\}_{i=1}^{\infty}$ be orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$ respectively, such that the conditional embedding operator $\mathcal{C}_{X|H}$ have the following singular value decomposition

$$\mathcal{C}_{X|H} = \sum_{i=1}^{\infty} \sigma_i \cdot u_i \otimes v_i, \tag{14}$$

where the singular values are ordered in non-increasing manner. Let $\{\tilde{v}_i\}_{i\in[k]}$ be a set of k orthonormal vectors in $\mathcal{G}$, we may approximate the conditional embedding operator by

$$\mathcal{C}_{X|H} \approx \mathcal{C}_{X|H}\,\mathcal{V}\mathcal{V}^{\top} \tag{15}$$

where $\mathcal{V} := \sum_{i\in[k]} \tilde{v}_i \otimes e_i$ is an operator mapping from $\mathbb{R}^k$ to $\mathcal{F}$, and $\mathcal{V}^{\top}$ is its ajoint. The composition of $\mathcal{V}$ and $\mathcal{V}^*$ form a projection operator which essentially reduces a vector in the potentially infinite dimensional RKHS $\mathcal{G}$ to one in k dimensional space, and them bring it back to $\mathcal{G}$. We will define

$$\widetilde{\mathcal{C}}_{X|H} := \mathcal{C}_{X|H}\,\mathcal{V} \tag{16}$$

as k-restricted conditional embedding operator which maps from a k-dimensional subspace of $\mathcal{G}$ to $\mathcal{F}$.

Finite sample estimate. Given a dataset $\mathcal{D}_{XH} = \left\{(x^i, h^i)\right\}_{i\in[m]}$ of size m drawn *i.i.d.* from $\mathbb{P}(X,H)$, we can estimate the conditional embedding operator as

$$\widehat{\mathcal{C}}_{X|H} := \Phi(L+\lambda I)^{-1}\Upsilon^{\top} \tag{17}$$

where $\Phi := (\phi(x^1), \ldots, \phi(x^m))$ and $\Upsilon := (\phi(h^1), \ldots, \phi(h^m))$ are implicitly formed feature matrix, and $L = \Upsilon^\top \Upsilon$ is the Gram matrix for samples from variable H. Furthermore, we need the additional regularization parameter λ to avoid overfitting. Then $\widehat{\mu}_{X|h} = \widehat{\mathcal{C}}_{X|H}\phi(h)$ becomes a weighted sum of feature mapped data points from X,

$$\widehat{\mu}_{X|h} = \sum_{i=1}^{m} \beta^i(h)\phi(x^i) = \Phi\,\beta(h) \quad \text{where} \tag{18}$$
$$\beta(h) := \left(\beta^1(h), \ldots, \beta^m(h)\right)^\top = (L + \lambda I)^{-1} L_{:h},$$

and $L_{:h} = (l(h, h^1), \ldots, l(h, h^m))^\top$. The empirical estimator of the conditional embedding is similar to the estimator of the ordinary embedding from equation (1). The difference is that, instead of applying uniform weights $\frac{1}{m}$, the former applies *non-uniform* weights, $\beta^i(h)$, on observations which are, in turn, determined by the value h of the conditioning variable. These non-uniform weights reflect the effects of conditioning on the embeddings. It can also be shown that this empirical estimate converges to its population counterpart in RKHS norm, $\left\|\widehat{\mu}_{X|h} - \mu_{X|h}\right\|_{\mathcal{F}}$, with rate of $O_p(m^{-\frac{1}{4}})$ if one decreases the regularization λ with rate $O(m^{-\frac{1}{2}})$.

5. Multiview Latent Variable Models

Multi-view latent variable models (*e.g.*, naïve Bayes models) are a special class of Bayesian networks in which observed variables $X_1, X_2, \ldots, X_\ell$ are conditionally independent given a latent variable H, and the conditional distributions, $\mathbb{P}(X_t|H)$, of the $X_t, t \in [\ell]$ given the hidden variable H can be different. The conditional independent structure of a multiview latent variable model can be found in Figure **??**(a), and many complicated graphical models, such as the hidden Markov model in Figure **??**(b), can be reduced to a multiview latent variable model.

For simplicity of exposition, we now consider a simple model with three observed variables, X_1, X_2 and X_3 which are conditionally independent given H, and furthermore the conditional distributions, $\mathbb{P}(X_t|H)$, are the same for $t \in \{1, 2, 3\}$. That is the random variables are *exchangeable*. Our analysis can be easily extended to the general multi-view setting, see (Anandkumar et al., 2012) for details. Then the joint distributions, $\mathbb{P}(X_1, X_2)$ and $\mathbb{P}(X_1, X_2, X_3)$, can be factorized respectively as

$$\mathbb{P}(dx_1 \times dx_2) = \int_{\mathcal{H}} \mathbb{P}(dx_1|h)\,\mathbb{P}(dx_2|h)\,\mathbb{P}(dh) \tag{19}$$

$$\mathbb{P}(dx_1 \times dx_2 \times dx_3) = \int_{\mathcal{H}} \mathbb{P}(dx_1|h)\,\mathbb{P}(dx_2|h)\,\mathbb{P}(dx_3|h)\,\mathbb{P}(dh), \tag{20}$$

and the corresponding kernel embeddings can be factorized respectively as

$$\mathcal{C}_{X_1X_2} = \int_{\mathcal{X}\times\mathcal{X}\times\mathcal{H}} \phi(x_1) \otimes \phi(x_2)\,\mathbb{P}(dx_1|h)\,\mathbb{P}(dx_2|h)\,\mathbb{P}(dh) \tag{21}$$
$$= \int_{\mathcal{H}} \left(\int_{\mathcal{X}} \phi(x_1)\,\mathbb{P}(dx_1|h)\right) \otimes \left(\int_{\mathcal{X}} \phi(x_2)\,\mathbb{P}(dx_2|h)\right)\mathbb{P}(dh) \tag{22}$$
$$= \int_{\mathcal{H}} \left(\mathcal{C}_{X|H}\psi(h)\right) \otimes \left(\mathcal{C}_{X|H}\psi(h)\right)\,\mathbb{P}(dh) \tag{23}$$
$$= \mathcal{C}_{X|H}\,\left(\int_{\mathcal{H}} \psi(h) \otimes \psi(h)\,\mathbb{P}(dh)\right)\,\mathcal{C}_{X|H}^\top \tag{24}$$
$$= \mathcal{C}_{X|H}\,\mathcal{C}_{HH}\,\mathcal{C}_{X|H}^\top \tag{25}$$
$$\mathcal{C}_{X_1X_2X_3} = \int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}\times\mathcal{H}} \phi(x_1) \otimes \phi(x_2) \otimes \phi(x_3)\,\mathbb{P}(dx_1|h)\,\mathbb{P}(dx_2|h)\,\mathbb{P}(dx \tag{26}$$
$$= \mathcal{C}_{HHH} \times_1 \mathcal{C}_{X|H} \times_2 \mathcal{C}_{X|H} \times_3 \mathcal{C}_{X|H} \tag{27}$$

k-restricted factorization. If we use k-restricted conditional embedding operator to form approximations of the above operators, then we have

$$\widetilde{\mathcal{C}}_{X_1X_2} := \widetilde{\mathcal{C}}_{X|H}\,(\mathcal{V}^\top \mathcal{C}_{HH}\mathcal{V})\,\widetilde{\mathcal{C}}_{X|H}^\top \tag{28}$$
$$\widetilde{\mathcal{C}}_{X_1X_2X_3} := \left(\mathcal{C}_{HHH} \times_1 \mathcal{V}^\top \times_2 \mathcal{V}^\top \times_3 \mathcal{V}^\top\right) \times_1 \widetilde{\mathcal{C}}_{X|H} \times_2 \widetilde{\mathcal{C}}_{X|H} \times_3 \widetilde{\mathcal{C}}_{X|H} \tag{29}$$

Let the singular value decomposition of $\mathcal{C}_{HH}$ be $\sum_{i=1}^{\infty} \sigma_i \cdot v_i \otimes v_i$, and we construct $\mathcal{V}$ using the leading k singular vectors. Then $\mathcal{V}^\top \mathcal{C}_{HH}\mathcal{V}$ is a diagonal matrix with entries

$$\mathcal{V}^\top \mathcal{C}_{HH}\mathcal{V} = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_k \end{pmatrix} \tag{30}$$

where we used $\int_{\mathcal{H}} v_i(h)\,v_{i'}(h)\,\mathbb{P}(dh) = 0$ for $i \neq i'$. But the 3rd order tensor

$$\mathcal{C}_{HHH} \times_1 \mathcal{V}^\top \times_2 \mathcal{V}^\top \times_3 \mathcal{V}^\top = \mathbb{E}_H\left[(\mathcal{V}^\top\psi(h)) \otimes (\mathcal{V}^\top\psi(h)) \otimes (\mathcal{V}^\top\psi(h))\right] \tag{31}$$
$$= \left(\int_{\mathcal{H}} v_i(h)\,v_{i'}(h)\,v_{i''}(h)\,\mathbb{P}(dh)\right)_{i,i',i''\in[} \tag{32}$$

is not a diagonal tensor in general. However, there are important special cases where both the 2nd and 3rd order tensors are simultaneously diagonalizable.

Discrete latent variable. When the hidden variable $H \in [k']$ is discrete, and we use Kronecker delta kernel $\delta(h, h')$ on $\mathcal{H}$, then both

$$\mathcal{C}_{HH} = \begin{pmatrix} \mathbb{P}(h=1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbb{P}(h=k') \end{pmatrix}, \text{ and } \mathcal{C}_{HHH} = \begin{pmatrix} \mathbb{P}(h=i)\delta(i,i')\delta(i',i'') \end{pmatrix}_{i,i',i'' \in [k']} \tag{33}$$

are diagonal tensors. The operator $\mathcal{V}$ used for constructing k-restricted conditional embedding operator can be contructed from the standard basis. That is $\mathcal{V} = \sum_{i \in [k]} e_i \otimes \tilde{e}_i$, where e_i and $\tilde{e}_i$ are the i-th standard basis vector in $\mathbb{R}^{k'}$ and $\mathbb{R}^k$ respectively.

Furthermore, in this case, $\widetilde{\mathcal{C}}_{X|H} = \left(\mu_{X|h=1},\ \mu_{X|h=2},\ \dots,\ \mu_{X|h=k}\right)$, and the approximate kernel embeddings for $\mathbb{P}(X_1, X_2)$ and $\mathbb{P}(X_1, X_2, X_3)$ are

$$\widetilde{\mathcal{C}}_{X_1X_2} = \sum_{h \in [k]} \pi_h \cdot \mu_{X|h} \otimes \mu_{X|h}, \tag{34}$$

$$\widetilde{\mathcal{C}}_{X_1X_2X_3} = \sum_{h \in [k]} \pi_h \cdot \mu_{X|h} \otimes \mu_{X|h} \otimes \mu_{X|h} \tag{35}$$

where $\pi_h := \mathbb{P}(h)$.

Identifiability. Allman et al. showed that, under mild conditions, a finite mixture of nonparametric product distributions is identifiable. The multiview latent variable model in (35) has the same form as a finite mixture of nonparametric product distribution, and therefore we can adapt Allman's results to the current setting.

Theorem 1 *Let $\mathbb{P}(X_1, X_2, X_3)$ be a multiview latent variable model of the form (35), such that the conditional distributions $\{\mathbb{P}(X|h)\}_{h \in [k]}$ are linearly independent. Then, the set of parameters $\left\{\pi_h, \mu_{X|h}\right\}_{h \in [k]}$ are identifiable from $\mathcal{C}_{X_1X_2X_3}$, up to label swapping of the hidden variable H.*

6. Kernel Algorithm

We will deal with the discrete latent variable case. The parameters $\left\{\pi_h, \mu_{X|h}\right\}_{h \in [k]}$, of the multiview latent variable model can be recovered from $\mathcal{C}_{X_1X_2}$ and $\mathcal{C}_{X_1X_2X_3}$ using the following simple algorithm.

1. Eigen-value decomposition for $\mathcal{C}_{X_1X_2}$,

$$\mathcal{C}_{X_1X_2} = \sum_{i=1}^{\infty} \sigma_i \cdot u_i \otimes u_i$$

   where the eigen-values are ordered in non-decreasing manner. Let the leading k eigenvectors corresponding to the largest k eigen-value be $\mathcal{U}_k := (u_1, u_2, \dots, u_k)$, and the eigen-value matrix of $S_k := \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_k)$.

2. Whiten the 3rd order kernel embedding $\mathcal{C}_{X_1X_2X_3}$ using whitening matrix $\mathcal{W} := \mathcal{U}_k S_k^{-1/2}$.

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

   where $\times_i$ denotes mode-i tensor-matrix multiplication.

3. Find the leading k tensor eigenvectors V_k for $\mathcal{T}$ using tensor power method.

4. Recover the conditional embedding operator by

$$\mathcal{C}_{X|H} = (\mu_{X|h=1}, \mu_{X|h=1}, \dots, \mu_{X|h=k}) = (\mathcal{W})^\dagger V_k$$

Finite sample estimate. Given m observation $\mathcal{D}_{X_1X_2X_3} = \{(x_1^i, x_2^i, x_3^i)\}_{i \in [m]}$ drawn *i.i.d.* from a multi-view latent variable model $\mathbb{P}(X_1, X_2, X_3)$, we now design a kernel algorithm to estimate the latent parameters from data. Although the empirical kernel embeddings has infinite dimensions, we can carry out the decomposition using just the kernel matrices.

1. We will perform a kernel eigenvalue decomposition of the empirical 2nd order embedding

$$\widehat{\mathcal{C}}_{X_1X_2} := \frac{1}{2m} \sum_{i=1}^{m} \left(\phi(x_1^i) \otimes \phi(x_2^i) + \phi(x_2^i) \otimes \phi(x_1^i) \right).$$

   Denote the implicitly formed feature matrix by

$$\Phi := (\phi(x_1^1), \phi(x_1^2), \dots, \phi(x_1^m), \phi(x_2^1), \phi(x_2^2), \dots, \phi(x_2^m)) \tag{36}$$

$$\Psi := (\phi(x_2^1), \phi(x_2^2), \dots, \phi(x_2^m), \phi(x_1^1), \phi(x_1^2), \dots, \phi(x_1^m)) \tag{37}$$

   respectively, and the corresponding kernel matrix be $K = \Phi^\top \Phi$ and $L = \Psi^\top \Psi$. Using the feature matrix, $\mathcal{C}_{X_1X_2}$ can be expressed as

$$\mathcal{C}_{X_1X_2} = \frac{1}{2m} \Phi \Psi^\top.$$

   Its leading k eigenvectors $\mathcal{U}_k = (u_1, \dots, u_k)$ will lie in the span of the column of Φ, *i.e.*, $\mathcal{U}_k = \Phi(\beta_1, \dots, \beta_k)$ where $\beta \in \mathbb{R}^m$. Then we can transform the eigen-value decomposition problem for an infinite dimensional matrix to a problem involving finite dimensional kernel matrices,

$$\mathcal{C}_{X_1X_2} \mathcal{C}_{X_1X_2}^\top u = \sigma\ u \quad \Rightarrow \quad \frac{1}{4m^2} \Phi \Psi^\top \Psi \Phi^\top \Phi \beta = \sigma\, \Phi \beta \quad \Rightarrow \quad \frac{1}{4m}$$

   Let the Cholesky decomposition of K be $R^\top R$. Then by redefining $\widetilde{\beta} = R\beta$, and solving an eigenvalue problem

$$\frac{1}{4m^2} R L R^\top \widetilde{\beta} = \sigma\, \widetilde{\beta}, \text{ and obtain } \beta = R^\dagger \widetilde{\beta}. \tag{38}$$

Algorithm 1 KernelSVD(K, L, k)

```
Out: $S_k$ and $(\beta_1, \ldots, \beta_k)$
1: Perform Cholesky decomposition: $K = R^\top R$
2: Solve eigen-decomposition problem: $\frac{1}{4m^2} RLR^\top \widetilde{\beta} = \sigma \widetilde{\beta}$
3: Let the $k$ leading eigen-values be: $S_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$
4: Let the corresponding $k$ leading eigenvectors be: $(\widetilde{\beta}_1, \ldots, \widetilde{\beta}_k)$
5: Compute: $(\beta_1, \ldots, \beta_k) = R^\dagger(\widetilde{\beta}_1, \ldots, \widetilde{\beta}_k)$
```

The resulting eigenvectors satisfy $u_i^\top u_{i'} = \beta_i^\top \Phi^\top \Phi \beta_{i'} = \beta_i^\top K \beta_{i'} = \widetilde{\beta}_i^\top \widetilde{\beta}_{i'} = \delta_{ii'}$. This algorithm is summarized in Algorithm 1.

2. We whiten the empirical 3rd order embedding

$$\widehat{\mathcal{C}}_{X_1X_2X_3} := \frac{1}{3m}\sum_{i=1}^{m}\left(\phi(x_1^i)\otimes\phi(x_2^i)\otimes\phi(x_3^i)+\phi(x_3^i)\otimes\phi(x_1^i)\otimes\phi(x_2^i)+\phi(x_2^i)\otimes\phi(x_3^i)\otimes\phi(x_1^i)\right)$$

to obtain

$$\widehat{\mathcal{T}} := \frac{1}{3m}\sum_{i=1}^{m}\left(\xi(x_1^i)\otimes\xi(x_2^i)\otimes\xi(x_3^i)+\xi(x_3^i)\otimes\xi(x_1^i)\otimes\xi(x_2^i)+\xi(x_2^i)\otimes\xi(x_3^i)\otimes\xi(x_1^i)\right), \tag{39}$$

where

$$\xi(x_1^i) := S_k^{-1/2}(\beta_1, \ldots, \beta_k)^\top \Phi^\top \phi(x_1^i) \quad \in \quad \mathbb{R}^k.$$

3. We run tensor power method in Algorithm 2 on the finite dimension tensor $\widehat{\mathcal{T}}$ to obtain its leading k eigenvectors $V_k := (v_1, \ldots, v_k)$.

4. The estimate for the conditional embedding is given by

$$\widehat{\mathcal{C}}_{X|H} = (\widehat{\mu}_{X|h=1}, \ldots, \widehat{\mu}_{X|h=k}) = \Phi(\beta_1, \ldots, \beta_k) S_k^{1/2} V_k \tag{40}$$

7. Interpretation with Parametric Family

For interpretable results, we can project the nonparametric representation to parametric family of distributions (*e.g.*, exponential families) as post-processing.

8. Algorithm and Sample Complexity Bounds

8.1. Robust Tensor Power Method

We recap the robust tensor power method for finding the tensor eigen-pairs, analyzed in detail in (Anandkumar et al., 2013). ### **AA:** I will add some more discussion later ###

Algorithm 2 $\{\lambda, \Phi\}$ ←TensorEigen(T, $\{v_i\}_{i\in[L]}$, N)

```
Input: Tensor $T \in \mathbb{R}^{k\times k\times k}$, set of $L$ initialization vectors $\{v_i\}_{i\in L}$, number of iterations $N$.
Output: the estimated eigenvalue/eigenvector pairs $\{\lambda, \Phi\}$, where $\lambda$ is the vector of eigenvalues and $\Phi$ is the matrix of eigenvectors.
  for $i = 1$ to $k$ do
    for $\tau = 1$ to $L$ do
      $\theta_0 \leftarrow v_\tau$.
      for $t = 1$ to $N$ do
        $\tilde{T} \leftarrow T$.
        for $j = 1$ to $i - 1$ (when $i > 1$) do
          if $|\lambda_j \left\langle \theta_t^{(\tau)}, \phi_j, | \right\rangle > \xi$ then
            $\tilde{T} \leftarrow \tilde{T} - \lambda_j \phi_j^{\otimes 3}$.
          end if
        end for
        Compute power iteration update $\theta_t^{(\tau)} := \frac{\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})}{\|\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})\|}$
      end for
    end for
    Let $\tau^* := \arg\max_{\tau\in L}\{\tilde{T}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
    Do $N$ power iteration updates starting from $\theta_N^{(\tau^*)}$ to obtain eigenvector estimate $\phi_i$, and set $\lambda_i := \tilde{T}(\phi_i, \phi_i, \phi_i)$.
  end for
  return the estimated eigenvalue/eigenvectors $(\lambda, \Phi)$.
```

8.2. Sample Bounds

Let $\kappa := \sup_{x\in\Omega} k(x,x)$, $\|\cdot\|$ be the Hilbert-Schmidt norm, $\pi_{\min} := \min_{i\in[k]} \pi_i$ and $\sigma_k(\mathcal{C}_{X_1,X_2})$ be the k^{th} singular value of $\mathcal{C}_{X_1,X_2}$.

Theorem 2 (Sample Bounds) *Pick any $\delta \in (0,1)$. When the number of samples m satisfies*

$$m > \frac{\rho\kappa^2 \log \frac{\delta}{2}}{\sigma_k(\mathcal{C}_{X_1,X_2})}, \quad \rho := \max\left(\frac{512}{\sigma_k^3(\mathcal{C}_{X_1,X_2})}, \frac{C'k^2}{\sigma_k^3(\mathcal{C}_{X_1,X_2})}, \frac{C''k^{\frac{2}{3}}}{\pi_{\min}^{\frac{1}{3}}}\right) \tag{41}$$

for some constants $C', C'' > 0$, and the number of iterations N and the number of random initialization vectors L (drawn uniformly on the sphere $\mathcal{S}^{k-1}$) satisfy

$$N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{1}{\sqrt{\pi_{\min}}\,\epsilon_T}\right)\right),$$

$$\sqrt{\frac{\ln(L/\log_2(k/\delta))}{\ln(k)}} \cdot \left(1 - \frac{\ln(\ln(L/\log_2(k/\delta))) + C_3}{4\ln(L/\log_2(k/\delta))} - \sqrt{\frac{\ln(8)}{\ln(L/\log_2(k/\delta))}}\right) \geq 1.02\left(1 + \sqrt{\frac{\ln(4)}{\ln(k)}}\right).$$

for constants $C_2, C_3 > 0$. (Note that the condition on L holds with $L = \operatorname{poly}(k)\log(1/\delta)$.) The robust power method in Algorithm 2 yields eigen-pairs $(\widehat{\lambda}_i, \widehat{\phi}_i)$ such that there exists a permutation η, with probability $1-4\delta$, we have

$$\|\pi_j^{-1/2}\mu_{X|h=j} - \widehat{\phi}_{\eta(j)}\| \leq 8\epsilon_T \cdot \pi_j^{-1/2}, \qquad |\pi_j^{-1/2} - \widehat{\lambda}_{\eta(j)}| \leq 5\epsilon_T, \quad \forall j \in [k],$$

and

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

where ϵ_T is the tensor perturbation bound

$$\epsilon_T := \|\widehat{\mathcal{T}} - \mathcal{T}\| \leq \frac{12\kappa\sqrt{\log\frac{\delta}{2}}}{\sqrt{m}\,\sigma_k^{1.5}(\mathcal{C}_{X_1,X_2})} + \frac{512\sqrt{2}\kappa^3\left(\log\frac{\delta}{2}\right)^{1.5}}{m^{1.5}\,\sigma_k^3(\mathcal{C}_{X_1,X_2})\sqrt{\pi_{\min}}}$$

Thus, the above result provides bounds on the estimated eigen-pairs using the robust tensor power method. The proof is in Appendix A.

A. Proof of Theorem 2

A.1. Recap of Perturbation Bounds for the Tensor Power Method

We now recap the result of (?)Thm. 13]AnandkumarEtal:community12 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.*

$$\epsilon_T \leq C_1 \cdot \lambda_{\min} R_0^2, \qquad N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{\lambda_{\max}}{\epsilon_T}\right)\right), \tag{42}$$

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 (42) to sample complexity bounds.

A.2. Concentration Bounds

A.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

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

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

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

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,

$$\mathcal{C}_{X_1,X_2} = \sum_{j=1}^{k} \pi_j \cdot \mu_{X|j} \otimes \mu_{X|j} + E_{X_1X_2} = M\,\mathrm{Diag}(\pi) M^\top + E_{X_1X_2}, \tag{43}$$

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 (53), $\epsilon_{pairs} := \left\|\mathcal{C}_{X_1,X_2} - \widehat{\mathcal{C}}_{X_1,X_2}\right\|$. Let $\sigma_1(\cdot) \geq \sigma_2(\cdot) \ldots$ denote the singular values of an operator.

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

$$\epsilon_W := \| \operatorname{Diag}(\pi)^{1/2} M^\top (\widehat{\mathcal{W}} - \mathcal{W})\| \leq \frac{2(2\epsilon_{pairs} + \sigma_{k+1}(\mathcal{C}_{X_1X_2}))}{\sigma_k(\mathcal{C}_{X_1X_2})} \cdot (1 + \sigma_{k+1}(\mathcal{C}_{X_1X_2})) \tag{44}$$

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

Proof: The proof is along the lines of (**?**)Lemma 16]AnandkumarEtal:community12, 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 $\mathcal{C}_{X_1X_2} = M \operatorname{Diag}(\pi) M^\top + E$, we have that $\| \operatorname{Diag}(\pi)^{1/2} M^\top \mathcal{W}\| = 1$ or $\| \operatorname{Diag}(\pi)^{1/2} M^\top \mathcal{W}\| \leq \|I - E\|^{1/2} \leq 1 + \sigma_{k+1}(\mathcal{C}_{X_1X_2})$. 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 $\|\widetilde{E}\|_{HS} \leq \epsilon_{pairs} + \sigma_{k+1}(B) \leq 2\epsilon_{pairs} + \sigma_{k+1}(\mathcal{C}_{X_1X_2})$. We now have

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

Now

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

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

A.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

$$\mathcal{C}_{X_1X_2X_3} = \sum_{h \in [k]} \pi_h \cdot \mu_{X|h} \otimes \mu_{X|h} \otimes \mu_{X|h} + E_{X_1,X_2,X_3} \tag{45}$$

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*

$$\epsilon_T := \|\widehat{\mathcal{T}} - \mathcal{T}\| \leq \frac{2\sqrt{2}\epsilon_{triples}}{\sigma_k(\mathcal{C}_{X_1X_2})^{1.5}} + \frac{\epsilon_W^3}{\sqrt{\pi_{\min}}} + \|E_{X_1X_2X_3}\| \frac{\epsilon_W^3}{\pi_{\min}^{1.5}\sigma_k(M)^3}. \tag{46}$$

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.

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

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

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

where $\|E_{X_1,X_2,X_3}\|$ is the residual and we need to bound this in non-parametric case.

$$\begin{aligned}\|\widehat{\mathcal{T}} - \mathcal{T}\| &= \|\mathcal{C}_{X_1X_2X_3} \times_1 (\widehat{\mathcal{W}} - \mathcal{W})^\top \times_2 (\widehat{\mathcal{W}} - \mathcal{W})^\top \times_3 (\widehat{\mathcal{W}} - \mathcal{W})^\top\| \\ &\leq \frac{\| \operatorname{Diag}(\pi)^{1/2} M^\top (\widehat{\mathcal{W}} - \mathcal{W})\|^3}{\sqrt{\pi_{\min}}} + \|E_{X_1X_2X_3}\| \|\widehat{\mathcal{W}} - \mathcal{W}\|^3 \\ &= \frac{\epsilon_W^3}{\sqrt{\pi_{\min}}} + \|E_{X_1X_2X_3}\| \frac{\epsilon_W^3}{\pi_{\min}^{1.5}\sigma_k(M)^3}\end{aligned}$$

□

Proof of Theorem 2:

We obtain a condition on the above perturbation ϵ_T in (46) by applying Theorem 3 as $\epsilon_T \leq C_1 \lambda_{\min} R_0^2$. Here, we have $\lambda_i = 1/\sqrt{\pi_i} \geq 1$. For random initialization, we have that $R_0 \sim 1/\sqrt{k}$, with probability $1 - \delta$ using $\operatorname{poly}(k) \operatorname{poly}(1/\delta)$ trials (**?**)Thm. 5.1]AnandkumarEtal:tensor12. Thus, we require that $\epsilon_T \leq \frac{C_1}{k}$. Summarizing, we require for the following conditions to hold

$$\epsilon_{pairs} \leq 0.5\sigma_k(\mathcal{C}_{X_1X_2}), \quad \epsilon_T \leq \frac{C_1}{k}. \tag{47}$$

We now substitute for ϵ_{pairs} and $\epsilon_{triples}$ in (46) using Lemma 6 and Lemma 7. First consider the case when H is exactly a k-way categorical variable.

Case of k-component mixture: Here, $E_{X_1,X_2} = 0$ and $E_{X_1,X_2,X_3} = 0$ in (43) and (45). We have that

$$\epsilon_W \leq \frac{4\epsilon_{pairs}}{\sigma_k(\mathcal{C}_{X_1,X_2})} \leq \frac{8\sqrt{2}\kappa\sqrt{\log\frac{\delta}{2}}}{\sqrt{m}\,\sigma_k(\mathcal{C}_{X_1,X_2})},$$

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

$$m > \frac{512\kappa^2 \log\frac{\delta}{2}}{\sigma_k^4(\mathcal{C}_{X_1,X_2})}. \tag{48}$$

Further we require that $\epsilon_T \leq C_1/k$, which implies that each of the terms in (46) is less than C/k, for some constant C. Thus, we have

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

for some constant C' with probability $1-\delta$. Similarly for the second term in (46), we have

$$\frac{\epsilon_W^3}{\sqrt{\pi_{\min}}} < \frac{C}{k} \quad \Rightarrow \quad m > \frac{C'' k^{\frac{2}{3}}\kappa^2 \log\frac{\delta}{2}}{\pi_{\min}^{\frac{1}{3}}\sigma_k(\mathcal{C}_{X_1,X_2})},$$

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

AA: let me know if the extension beyond k-mixture is interesting and I will add it.

□

A.2.3. Concentration bounds for Empirical Operators

AA: I changed the notation slightly

Concentration results for the singular value decomposition of empirical operators.

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

$$\epsilon_{pairs} := \left\|\mathcal{C}_{X_1X_2} - \widehat{\mathcal{C}}_{X_1X_2}\right\|, \tag{49}$$

$$\Pr\left\{\epsilon_{pairs} \leqslant \frac{2\sqrt{2}\kappa\sqrt{\log\frac{\delta}{2}}}{\sqrt{m}}\right\} \geqslant 1-\delta. \tag{50}$$

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

$$\xi_i = \phi(x_t^i) \otimes \phi(x_{t'}^i) - \mathcal{C}_{X_t,X_{t'}}. \tag{51}$$

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

$$\sup_{x_1,x_2} \|\phi(x_1)\otimes\phi(x_2)\|^2 \leqslant \kappa^2, \tag{52}$$

which implies that $\|\mathcal{C}_{X_1X_2}\| \leqslant \kappa$, and $\|\xi_i\| \leqslant 2\kappa$. 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 $\kappa := \sup_{x\in\Omega} k(x,x)$, and $\|\cdot\|$ be the Hilbert-Schmidt norm, we have for*

$$\epsilon_{triples} := \left\|\mathcal{C}_{X_1X_2X_3} - \widehat{\mathcal{C}}_{X_1X_2X_3}\right\|, \tag{53}$$

$$\Pr\left\{\epsilon_{triples} \leqslant \frac{3\sqrt{2}\kappa\sqrt{\log\frac{\delta}{2}}}{\sqrt{m}}\right\} \geqslant 1-\delta. \tag{54}$$

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor Methods for Learning Latent Variable Models. *Available at arXiv:1210.7559*, Oct. 2012.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. M. A Tensor Spectral Approach to Learning Mixed Membership Community Models. *ArXiv 1302.2684*, Feb. 2013.

Fine, S. and Scheinberg, K. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Gretton, A., Fukumizu, K., Teo, C.-H., Song, L., Schölkopf, B., and Smola, A. J. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pp. 585–592, Cambridge, MA, 2008. MIT Press.

Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

Kolda, Tamara. G. and Bader, Brett W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Rosasco, L., Belkin, M., and Vito, E.D. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

Schölkopf, B., Tsuda, K., and Vert, J.-P. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 4754, pp. 13–31. Springer, 2007.

Song, L., Huang, J., Smola, A. J., and Fukumizu, K. Hilbert space embeddings of conditional distributions. In *Proceedings of the International Conference on Machine Learning*, 2009.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Injective Hilbert space embeddings of probability measures. In *Proc. Annual Conf. Computational Learning Theory*, pp. 111–122, 2008.