Nonparametric Estimation of Multi-View Latent Variable Models

Le Song, Animashree Anandkumar, Bo Dai and Bo Xie
Georgia Institute of Technology, University of California Irvine

Introduction

Given samples only from the observed variables $\{X_t\}_{t\in[\ell]}$, could we recover the multi-view latent variable models generating the dataset?

$$\mathbb{P}\left(\{X_t\}_{t\in[\ell]}\right) = \sum\nolimits_{h\in[k]} \mathbb{P}(h)\cdot \prod\nolimits_{t\in[\ell]} \mathbb{P}(X_t|h), \quad l \geq 3$$

where h is discrete and $\{X_t\}_{t\in[\ell]}$ are conditional independent given h.

Naïve Bayes Model — Hidden Markov Model

Figure: Examples of multi-view latent variable models

We proposed a kernel method for obtaining sufficient statistics for the model with theoretical guarantee.

- Based on spectral algorithm for model estimation, our algorithm is computational efficient and with provable guarantees.
- Without the parametric assumption involved in $\mathbb{P}(X_t|h)$, our algorithm is more flexible and robust.

Kernel Embeddings of Distributions

Denote $k(\cdot,\cdot)$ is the kernel function of a RKHS whose elements are functions $f:\Omega\mapsto\mathbb{R}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$. $k(x,\cdot)$ can be viewed as an implicit feature map $\phi(x)$ where $k(x,x') = \langle\phi(x),\phi(x')\rangle_{\mathcal{F}}$
A kernel embedding represents a density by its expected features,

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

And the conditional distribution embedding is $\mu_{X|h} := \mathbb{E}_{X|h}[\phi(X)]$.
Kernel embeddings could be generalized to joint distribution of two or more variables using tensor product feature maps.

$$\mathcal{C}_{X_{1:d}} := \mathbb{E}_{X_{1:d}}\left[\otimes_{i=1}^d \phi(X_i)\right] = \int_{\mathcal{X}^d} \left(\otimes_{i=1}^d \phi(x_i)\right) p(x_1,\ldots,x_d) \prod_{i=1}^d dx_i$$

It could be viewed as a multi-linear operator of order d mapping from $\mathcal{F}\times\ldots\times\mathcal{F}$ to $\mathbb{R}$. And thus,

$$\mathcal{C}_{X_{1:d}} \times_1 f_1 \times_2 \ldots \times_d f_d := \left\langle \mathcal{C}_{X_{1:d}},\ \otimes_{i=1}^d f_d \right\rangle_{\mathcal{F}^d} = \mathbb{E}_{X_{1:d}}\left[\prod_{i=1}^d \langle\phi(X_i), f_i\rangle_{\mathcal{F}}\right]$$

Recap Multi-View Latent Variable Models

Tile these embeddings into a matrix, the conditional embedding operator is $\mathcal{C}_{X|H} = \left(\mu_{X|h=1}, \mu_{X|h=2}, \ldots, \mu_{X|h=k}\right)$.
Since we assume the hidden variable $H\in[k]$ is discrete, let $\pi_h := \mathbb{P}(h)$, then,

$$\mathcal{C}_{HH} = \mathbb{E}_H[e_H \otimes e_H] = \left(\pi_h \delta(h,h')\right)_{h,h'\in[l]},$$
$$\mathcal{C}_{HHH} = \mathbb{E}_H[e_H\otimes e_H\otimes e_H] = \left(\pi_h\,\delta(h,h')\,\delta(h',h'')\right)_{h,h',h''\in[l]}$$

We obatin the factorization of $\mathbb{P}(X_1,X_2)$ and $\mathbb{P}(X_1,X_2,X_3)$ respectively,

$$\mathcal{C}_{X_1X_2} = \mathcal{C}_{X|H}\mathcal{C}_{HH}\mathcal{C}_{X|H}^\top = \sum\nolimits_{h\in[k]} \pi_h\cdot\mu_{X|h}\otimes\mu_{X|h}$$
$$\mathcal{C}_{X_1X_2X_3} = \mathcal{C}_{HHH}\times_1\mathcal{C}_{X|H}\times_2\mathcal{C}_{X|H}\times_3\mathcal{C}_{X|H} = \sum\nolimits_{h\in[k]} \pi_h\cdot\mu_{X|h}\otimes\mu_{X|h}\otimes\mu_{X|h}$$

Under mild condition, the set $\{\pi_h, \mu_{X|h}\}$ is identifiable.

Spectral Algorithm

For simplicity of exposition, the algorithm is explained for symmetric view population case.

1. Perform eigen-decomposition of $\mathcal{C}_{X_1X_2}$ which is rank l. Denote the l-leading eigenvectors be $\mathcal{U}_l := (u_1,u_2,\ldots,u_l)$, and the eigen-value matrix be $S_l := \mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_l)$. We have the whitening operator $\mathcal{W} := \mathcal{U}_k S_k^{-1/2}$ which satisfies
$$\mathcal{W}^\top\mathcal{C}_{X_1X_2}\mathcal{W} = (\mathcal{W}^\top\mathcal{C}_{X|H}\mathcal{C}_{HH}^{1/2})(\mathcal{C}_{HH}^{1/2}\mathcal{C}_{X|H}^\top\mathcal{W}) = I$$
and $M := \mathcal{W}^\top\mathcal{C}_{X|H}\mathcal{C}_{HH}^{1/2}$ is an orthogonal matrix.
2. Apply the whiten operator to the 3rd order kernel embedding $\mathcal{C}_{X_1X_2X_3}$
$$\mathcal{T} := \mathcal{C}_{X_1X_2X_3}\times_1(\mathcal{W}^\top)\times_2(\mathcal{W}^\top)\times_3(\mathcal{W}^\top) = \mathcal{C}_{HHH}^{-1/2}\times_1 M\times_2 M\times_3 M,$$
which is a tensor with orthogonal factors.
3. Use tensor power method to find the leading l eigenvectors M for $\mathcal{T}$. The corresponding l eigenvalues $\lambda = (\lambda_1,\ldots,\lambda_l)^\top$ will then be equal to $(\mathbb{P}(h=1)^{-1/2},\ldots,\mathbb{P}(h=l)^{-1/2})$.
4. Recover the conditional embedding operator by undoing the whitening step $\mathcal{C}_{X|H} = (\mathcal{W}^\top)^\dagger M\,\mathrm{diag}(\lambda)$.

Kernelization

Given m 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 $\mathbb{P}(X_1,X_2,X_3)$, we denote the implicit feature matrix

$$\Phi := (\phi(x_1^1),\ldots,\phi(x_1^m),\phi(x_2^1),\ldots,\phi(x_2^m)),$$
$$\Psi := (\phi(x_2^1),\ldots,\phi(x_2^m),\phi(x_1^1),\ldots,\phi(x_1^m)),$$

and the corresponding kernel matrix by $K = \Phi^\top\Phi$ and $L = \Psi^\top\Psi$ respectively. $\otimes[\xi_1,\xi_2,\xi_3] := \xi_1\otimes\xi_2\otimes\xi_3 + \xi_3\otimes\xi_1\otimes\xi_2 + \xi_2\otimes\xi_3\otimes\xi_1$.

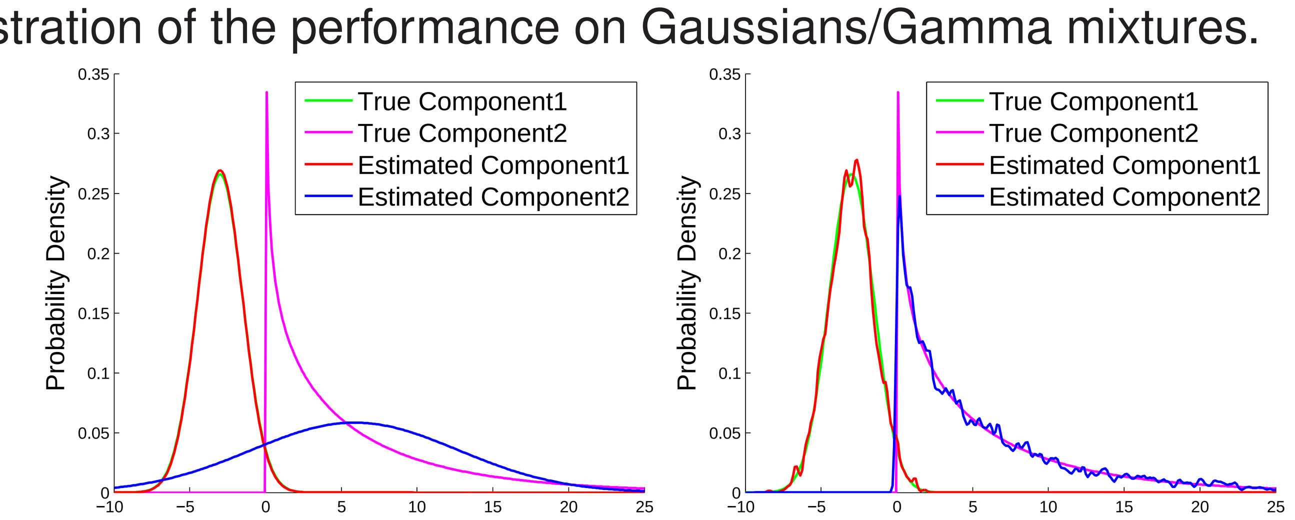
Kernel Spectral Algorithm

In: Kernel matrices K and L, and desired rank k
Out: A vector $\widehat{\pi}\in\mathbb{R}^k$ and a matrix $A\in\mathbb{R}^{2m\times k}$

1: Cholesky decomposition: $K = R^\top R$
2: Eigen-decomposition: $\frac{1}{4m^2}RLR^\top\widetilde{\beta} = \widehat{\sigma}^2\,\widetilde{\beta}$
3: Use k leading eigenvalues: $\widehat{S}_k = \mathrm{diag}(\widehat{\sigma}_1,\ldots,\widehat{\sigma}_k)$
4: Use k leading eigenvectors $(\widetilde{\beta}_1,\ldots,\widetilde{\beta}_k)$ to compute: $(\beta_1,\ldots,\beta_k) = R^\dagger(\widetilde{\beta}_1,\ldots,\widetilde{\beta}_k)$
5: Form tensor: $\widehat{\mathcal{T}} = \frac{1}{3m}\sum_{i=1}^m \otimes\left[\xi(x_1^i),\xi(x_2^i),\xi(x_3^i)\right]$ where $\xi(x_1^i) = \widehat{S}_k^{-1/2}(\beta_1,\ldots,\beta_k)^\top K_{x_1^i}$
6: Power method: eigenvectors $\widehat{M} := (\widehat{v}_1,\ldots,\widehat{v}_k)$, and the eigenvalues $\widehat{\lambda} := (\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k)^\top$ of $\widehat{\mathcal{T}}$
7: $A = (\beta_1,\ldots,\beta_k)\widehat{S}_k^{1/2}\widehat{M}\,\mathrm{diag}(\widehat{\lambda})$
8: $\widehat{\pi} = (\widehat{\lambda}_1^{-2},\ldots,\widehat{\lambda}_k^{-2})^\top$

Illustration

Illustration of the performance on Gaussians/Gamma mixtures.



(a) GMM (b) Kernel Spectral Algorithm

Sample Complexity

Theorem *Pick any $\delta\in(0,1)$. When the number of samples m satisfies $m > \frac{\theta\rho^2\log\frac{2}{\delta}}{\sigma_k^2(\mathcal{C}_{X_1X_2})}$, $\theta := \max\left(\frac{C_3k^2\rho}{\sigma_k(\mathcal{C}_{X_1X_2})}, \frac{C_4k^{2/3}}{\pi_{\min}^{1/3}}\right)$, for some constants $C_3, C_4 > 0$, and the number of iterations N and the number of random initialization vectors L (drawn uniformly on the sphere $\mathcal{S}^{k-1}$) satisfy*

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

for constant $C_2 > 0$ and $L = (k)\log(1/\delta)$, the robust power method yields eigen-pairs (λ_i, v_i) such that there exists a permutation η, with probability $1-4\delta$, we have

$$\|\pi_j^{-1/2}\mu_{X|h=j} - (\beta_1,\ldots,\beta_k)\widehat{S}_k^{1/2}v_{\eta(j)}\|_{\mathcal{F}} \leq 8\epsilon_{\mathcal{T}}\cdot\pi_j^{-1/2},$$
$$|\pi_j^{-1/2} - \lambda_{\eta(j)}| \leq 5\epsilon_{\mathcal{T}}, \quad \forall j\in[k],$$

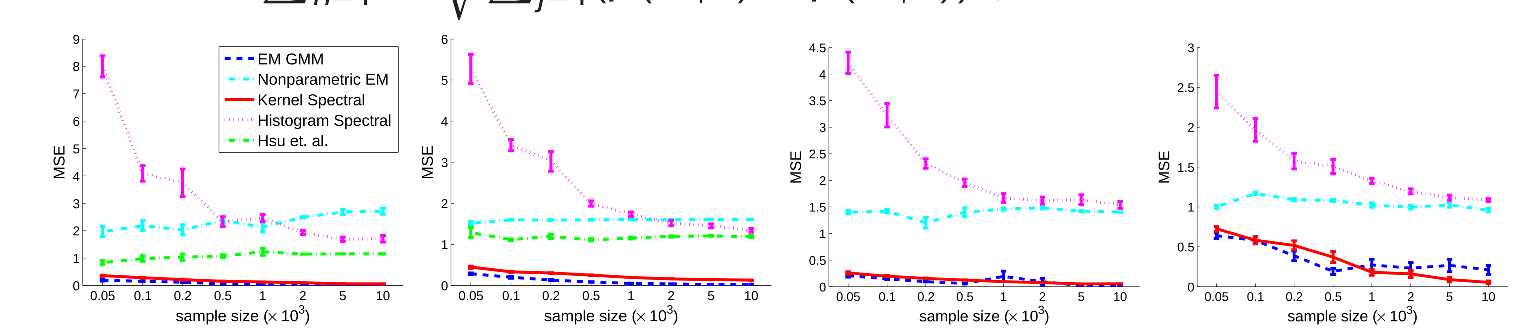
and $\|\mathcal{T} - \sum_{j=1}^k \hat{\lambda}_j\hat{\phi}_j^{\otimes 3}\| \leq 55_{\mathcal{T}}$ where $_{\mathcal{T}} := \|\mathcal{T} - \widehat{\mathcal{T}}\|$ is the tensor perturbation bound

$$_{\mathcal{T}} \leq \frac{8\rho^{1.5}\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}\,\sigma_k^{1.5}(\mathcal{C}_{X_1X_2})} + \frac{512\sqrt{2}\rho^3\left(\log\frac{2}{\delta}\right)^{1.5}}{m^{1.5}\,\sigma_k^3(\mathcal{C}_{X_1X_2})\sqrt{\pi_{\min}}}$$

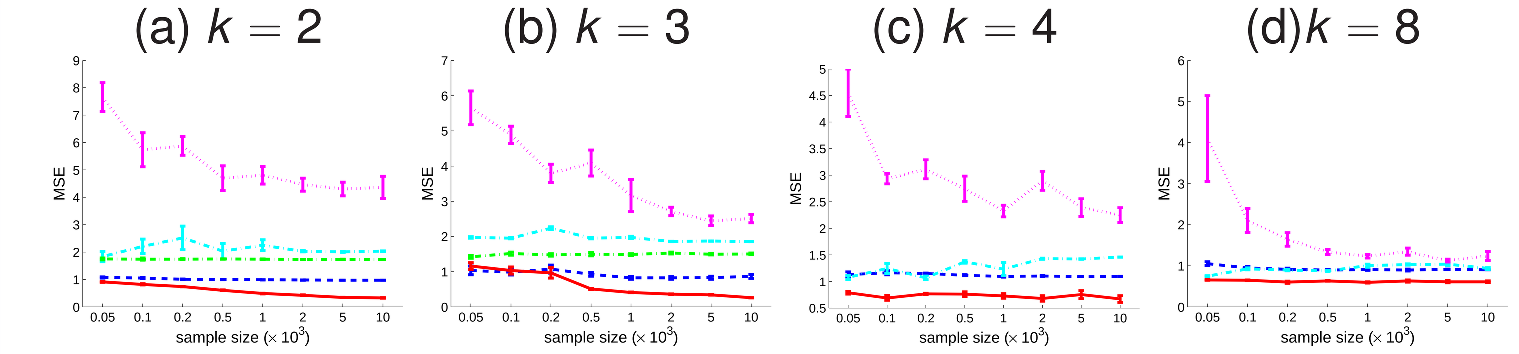
Remark: We note that the sample complexity is $(k,\rho,1/\pi_{\min},1/\sigma_k(\mathcal{C}_{X_1X_2}))$ of a low order, and in particular, it is $O(k^2)$, when the other parameters are fixed.

Synthetic Data: Model Estimation

We measured the performance of algorithms by the weighted ℓ_2 norm difference $\sum_{h=1}^k \pi_h\sqrt{\sum_{j=1}^{m'}(p(x^j|h) - \widehat{p}(x^j|h))^2}$.



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

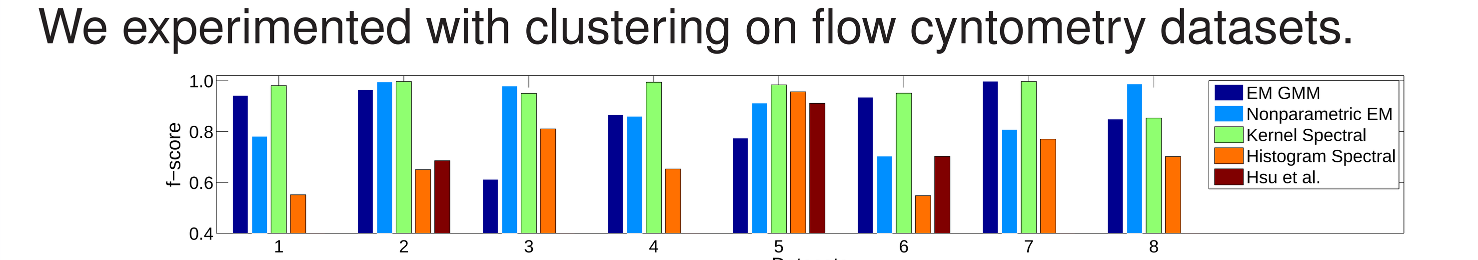


(e) $k=2$ (f) $k=3$ (g) $k=4$ (h) $k=8$

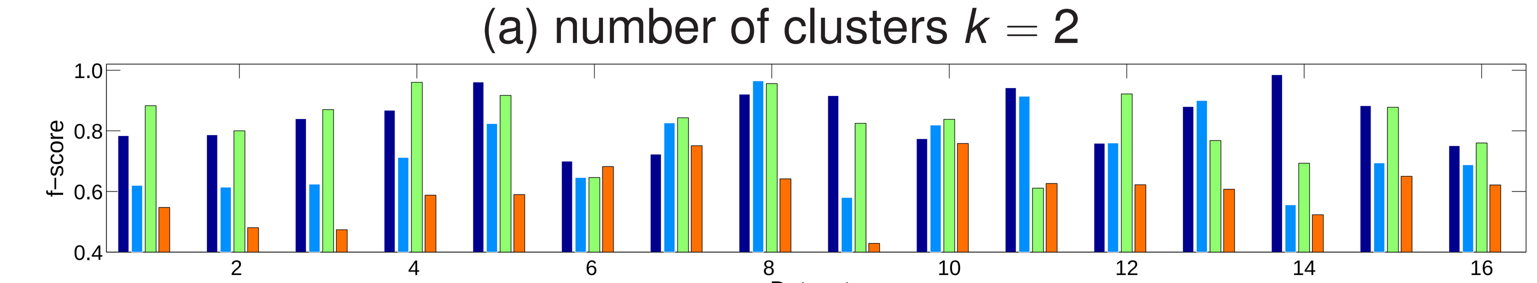
(a)-(d) Mixture of Gaussian distributions with $k=2,3,4,8$. (e)-(h) Mixture of Gaussian/Gamma distribution with $k=2,3,4,8$.

Real-World Data: Clustering Task

We experimented with clustering on flow cytometry datasets.



(a) number of clusters $k=2$



(b) number of clusters $k=3$
Clustering results on the DLBCL flow cytometry datasets.

Georgia Institute of Technology, University of California Irvine
lsong@cc.gatech.edu, a.anandkumar@uci.edu, bodai@gatech.edu, bxie33@gatech.edu