Spectral Learning in Latent Variable Models

Ayan Acharya, Rajiv Khanna

UT Austin

March 4, 2013

Parameter Estimation in Latent Variable Models

- Expectation maximization: problems – computational intractability, often no closed form solution of updates, local minima, uncertainty of solutions.
- General method of moments: problems – computational difficulty in solving multivariate polynomials.
- However, commonly used latent variable models have rich structure in their second order (matrix) and third order (tensor) moments.

Spectral Learning Techniques

- Methods of (Symmetric) Tensor Decomposition – the most general technique.
  - Power method.
  - Simultaneous diagonalization of matrices obtained from tensor.
- Subspace methods based on observable representation.

A Simple Example

- Toss a biased coin and based on the outcome, toss one of the two other biased coins and report the result – a mixture model with two components.
- Let's try method of moments on the independent observations.
- $\mathbb{E}[X] = \pi_1\mu_1 + \pi_2\mu_2$, where, $\pi_1 + \pi_2 = 1$.
- $\mathbb{E}[X_1X_2] = \mathbb{E}[X]^2$, $\mathbb{E}[X_1X_2X_3] = \mathbb{E}[X]^3, \cdots$.
- Higher order moments do not have any additional information.
- Can we leverage the structure of the problem in a more intelligent way?

A Simple Example Continued

- $\mathbb{E}[X] = \pi_1\mu_1 + \pi_2\mu_2$.
- $\mathbb{E}[X_1X_2|Z_1 = Z_2] = \pi_1\mu_1^2 + \pi_2\mu_2^2$.
- $\mathbb{E}[X_1X_2X_3|Z_1 = Z_2 = Z_3] = \pi_1\mu_1^3 + \pi_2\mu_2^3$.
- Lesson learnt: observations with related latent structure are useful for identifying parameters.
- Additionally, it is sufficient to know that the latent variables are drawn from the same distribution – they need not be the same.

Structure of Moments in pLSI

- k : number of mixture components, d : size of vocabulary, $\ell \geq 3$: minimum number of words per document.
- Let $\mathbf{w} = (w_i)_{i=1}^k$ denote the probability vector for topic selection. $\{\boldsymbol{\mu}_i\}_{i=1}^k$ be the topic-word distributions for different topics.
- One-hot encoding of the words in documents.
- Statistics based on words co-occurring in a given document:
  - $\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] = M_2 = \sum_{i=1}^{k} w_i(\mu_i \otimes \mu_i).$
  - $\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] = M_3 = \sum_{i=1}^{k} w_i(\mu_i \otimes \mu_i \otimes \mu_i).$
- Since $\boldsymbol{\mu}_\mathbf{i}$'s are not orthogonal, eigen decomposition of M_2 is not sufficient to recover w_i and $\boldsymbol{\mu}_\mathbf{i}$.

Rank of a Tensor

- The rank of a p^{th} order tensor $A \in \otimes^p \mathbb{R}^n$ is the minimum number k such that $A = \sum_{j=1}^{k} u_{1j} \otimes u_{2j} \otimes \cdots \otimes u_{pj}$ for $u_{ij} \in \mathbb{R}^n \ \forall i, j$.
- For $p = 2$, the above decomposition is the rank$-k$ approximation of the matrix A, a.k.a. SVD.
- For symmetric tensors the decomposition can be written as: $A = \sum_{j=1}^{k} \otimes^p u_j$ for $u_j \in \mathbb{R}^n \ \forall j$.
- Facts:
  - Rank of a tensor might not be finite!
  - There might not exist orthogonal eigen vectors!
  - Removal of the best rank-1 approximation might increase the rank of the residual tensor!

Eigen Decomposition of Symmetric Tensors

- Let $M(\mathbf{u}, \mathbf{u}) = \sum_{1 \leq i,j \leq d} M_{ij} (\mathbf{e}_i^\dagger \mathbf{u})(\mathbf{e}_j^\dagger \mathbf{u}) = \mathbf{u}^\dagger M \mathbf{u}$.
- Also, let $T(\mathbf{u}, \mathbf{u}, \mathbf{u}) = \sum_{1 \leq i,j,\ell \leq d} T_{ij\ell} (\mathbf{e}_i^\dagger \mathbf{u})(\mathbf{e}_j^\dagger \mathbf{u})(\mathbf{e}_\ell^\dagger \mathbf{u})$.
- Fixed Point Characterization of Eigen Vector:
  - Matrix: $M(\mathbf{I}, \mathbf{u}) = M\mathbf{u} = \lambda \mathbf{u}$.
  - Tensor: $T(\mathbf{I}, \mathbf{u}, \mathbf{u}) = \lambda \mathbf{u}$.
- Variational Characterization of Eigen Vector:
  - Matrix: $\sup_{\mathbf{u}} M(\mathbf{u}, \mathbf{u})$ s.t.$||\mathbf{u}||_2 = 1 \equiv \sup_{\mathbf{u}} \mathbf{u}^T M \mathbf{u}$ s.t.$||\mathbf{u}||_2 = 1$.
  - Tensor: $\sup_{\mathbf{u}} T(\mathbf{u}, \mathbf{u}, \mathbf{u})$ s.t.$||\mathbf{u}||_2 = 1$.

Further Problems with Symmetric Tensor

- Let $T = \sum_{i=1}^{k} \lambda_i (\mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i)$ with $\mathbf{v}_i$'s being orthogonal and $\lambda_i > 0 \forall i$.
- For any $S \subseteq \{1, 2, \cdots, k\}$ and for any $\mathbf{u} = \sum_{i \in S} \frac{\mathbf{v}_i}{\lambda_i}$, $T(\mathbf{I}, \mathbf{u}, \mathbf{u}) = \mathbf{u}$.
- There exists lot more eigen vectors than what the low rank structure suggests.
- Fortunately, there are only k "robust" eigen vectors.

Characterization of Robust Eigen Vectors

- Power method update for eigen decomposition: $\bar{\boldsymbol{\theta}} \mapsto \frac{M(\mathbf{I},\bar{\boldsymbol{\theta}})}{||M(\mathbf{I},\bar{\boldsymbol{\theta}})||}$.
- A unit vector $\mathbf{u}$ is a "robust eigenvector" of T if there exists an $\epsilon > 0$ such that $\forall \theta \in \{\mathbf{u}' : ||\mathbf{u}' - \mathbf{u}|| \leq \epsilon\}$, repeated iteration of the map $\bar{\boldsymbol{\theta}} \mapsto \frac{T(\mathbf{I},\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})}{||T(\mathbf{I},\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\theta}})||}$, converges to $\mathbf{u}$ starting from $\boldsymbol{\theta}$.
- Let T have an orthogonal decomposition. Then,
  1. The set of $\boldsymbol{\theta}$ which do not converge to some $\mathbf{v}_i$ under repeated tensor power method iteration has measure zero.
  2. The set of robust eigenvectors of T is equal to $\{\mathbf{v}_i\}_{i=1}^k$.
- Implication: start from somewhere and the power iteration takes to *one* of the robust eigen vectors!

Properties of Robust Eigen Vectors

- Let T have an orthogonal decomposition, and consider the optimization problem $\sup_{\mathbf{u}} T(\mathbf{u}, \mathbf{u}, \mathbf{u})$ s.t.$||\mathbf{u}|| = 1$.
  1. The stationary points are eigenvectors of T.
  2. A stationary point $\mathbf{u}$ is an isolated local maximizer if and only if $\mathbf{u} = \mathbf{v}_i$ for some $i \in \{1, 2, .. \cdots, k\}$.
- Stationary points other than robust eigen vectors can be discarded from the test of $T(\mathbf{I}, \mathbf{I}, \mathbf{u})$.

Reduction to Orthogonally Decomposable Tensor

- Non-degeneracy condition: the vectors $\{\boldsymbol{\mu}_i\}_{i=1}^k$ are linearly independent, and the scalars $w_i > 0 \forall i$ are strictly positive.
- Basic idea: use SVD of M_2 to construct an orthonormal basis for the span of $\{\boldsymbol{\mu}_i\}_{i=1}^k$, and in that basis some transformation of M_3 has a unique orthogonal decomposition whose eigenvectors determine $\{\boldsymbol{\mu}_i\}_{i=1}^k$.
- Let $W \in \mathbb{R}^{d \times k}$ be such that $M_2(W, W) = W^\dagger M_2 W = \mathbf{I}$.
- In particular, we can take $W = UD^{-1/2}$.
- $M_2(W, W) = \sum_{i=1}^{k} W^\dagger(\sqrt{w_i}\boldsymbol{\mu}_i)(\sqrt{w_i}\boldsymbol{\mu}_i)^\dagger W = \sum_{i=1}^{k} \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_i^\dagger = \mathbf{I}$.
- $\tilde{\boldsymbol{\mu}}_i$'s are orthogonal where $\tilde{\boldsymbol{\mu}}_i = \sqrt{w_i} W^\dagger \boldsymbol{\mu}_i$.

Eigen Decomposition of $\tilde{M}_3$

- Define $\tilde{M}_3 = M_3(W, W, W) = \sum_{i=1}^{k} w_i (W^{\dagger} \boldsymbol{\mu}_i)^{\otimes 3} = \sum_{i=1}^{k} \frac{\tilde{\boldsymbol{\mu}}_i^{\otimes 3}}{\sqrt{w_i}}$.
- The set of robust eigenvectors of $\tilde{M}_3$ is equal to $\{\tilde{\boldsymbol{\mu}}_i\}_{i=1}^{k}$.
- The eigenvalue corresponding to the robust eigenvector $\tilde{\boldsymbol{\mu}}_i$ of $\tilde{M}_3$ is equal to $1/\sqrt{w_i} \forall i$.
- If $B \in \mathbb{R}^{d \times k}$ is the Moore-Penrose pseudo-inverse of $W^{\dagger}$, and $(\mathbf{v}, \lambda)$ is a robust eigenvector/eigenvalue pair of $\tilde{M}_3$, then $\lambda B \mathbf{v} = \boldsymbol{\mu}_i$ for some $i \in \{1, 2, \cdots, k\}$.

Multi-view Models

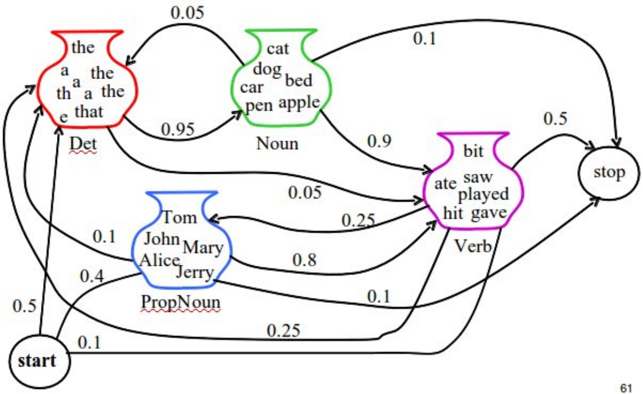
- $\ell \geq 3$ different views – $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\ell$ conditionally independent given $\mathbf{z}$.
- Similar to pLSI – only the conditional distributions are different.
- $\mathbb{E}[\mathbf{x}_t \otimes \mathbf{x}_{t'}] = \sum_{i=1}^{k} w_i(\mu_{ti} \otimes \mu_{t'i}) \; \forall t, t'$.
- $\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] = \sum_{i=1}^{k} w_i(\mu_{1i} \otimes \mu_{2i} \otimes \mu_{3i})$.
- $\tilde{\mathbf{x}}_1 = \mathbb{E}[\mathbf{x}_3 \otimes \mathbf{x}_2]\mathbb{E}[\mathbf{x}_3 \otimes \mathbf{x}_2]^{-1}\mathbf{x}_1$, $\tilde{\mathbf{x}}_2 = \mathbb{E}[\mathbf{x}_3 \otimes \mathbf{x}_1]\mathbb{E}[\mathbf{x}_2 \otimes \mathbf{x}_1]^{-1}\mathbf{x}_2$.
- $\mathbb{E}[\tilde{\mathbf{x}}_1 \otimes \tilde{\mathbf{x}}_2] = M_2 = \sum_{i=1}^{k} w_i(\mu_{3i} \otimes \mu_{3i})$.
- $\mathbb{E}[\tilde{\mathbf{x}}_1 \otimes \tilde{\mathbf{x}}_2 \otimes \tilde{\mathbf{x}}_3] = M_3 = \sum_{i=1}^{k} w_i(\mu_{3i} \otimes \mu_{3i} \otimes \mu_{3i})$.

Comparison with Other Methods

- Dasgupta and Schulman, 2007; Vempala and Wang, 2002; Chaudhuri and Rao, 2008; Brubaker and Vempala, 2008.
- Recovers the parameters provided that the distance between means is sufficiently large (roughly either d^c or k^c times the standard deviation of the Gaussians, for some $c > 0$).
- Techniques have been developed for learning GMM without any separation condition (Kalai et al., 2010; Belkin and Sinha, 2010; Moitra and Valiant, 2010).
- The computational and sample complexities of these methods grow exponentially with k – modern implementations of traditional method of moments.

Hidden Markov Model

- Discrete state, discrete observation HMM.
- Hidden state-observation pair $\{h_t, \mathbf{x}_t\}_t$.
- Number of hidden states: m and number of different outcomes of the the observations n with $m \leq n$.
- Parameters to be learnt:
  - Transition matrix T of dimension $m \times m$.
  - Observation probability matrix O of dimension $n \times m$.
  - Initial state distribution $\boldsymbol{\pi}$ – a vector of length m.



Courtesy: Dr. Ray Mooney.

Observable Operator View of HMM

- For $\mathbf{x} = \{1, 2, \cdots, n\}$ define $A_x = Tdiag(O_{x_1}, \cdots, O_{x_m})$. For any $t : \Pr[\mathbf{x}_1, \cdots, \mathbf{x}_t] = \mathbf{1}_m^\dagger A_{\mathbf{x}_t} \cdots A_{\mathbf{x}_1} \boldsymbol{\pi}$.
- Assumption 1: $\boldsymbol{\pi} > \mathbf{0}$, and O and T are rank m.
- $[P_1]_i = \Pr[x_1 = i]$, $[P_{2,1}]_{ij} = \Pr[x_2 = i, x_1 = j]$, $[P_{3,x,1}]_{ij} = \Pr[x_3 = i, x_2 = x, x_1 = j] \; \forall \mathbf{x} \in \{1, 2, \cdots, n\}$,
- Assumption 2: $U^\dagger O$ is invertible for some $U \in \mathbb{R}^{n \times m}$.
- Assume $\boldsymbol{\pi} > \mathbf{0}$ and that O and T have column rank m. Then rank$(P_{2,1}) = m$. Moreover, if U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then range(U) =range(O), so U obeys assumption 2.
- With the above two assumptions,
  - $\mathbf{b}_1 = U^\dagger P_1 = (U^\dagger O)\boldsymbol{\pi}$.
  - $\mathbf{b}_\infty^\dagger = (P_{2,1}^\dagger U)^{\dagger\dagger} P_1 = \mathbf{1}_m^\dagger (U^\dagger O)^{-1}$.
  - $B_x = (U^\dagger P_{3,x,1})(U^\dagger P_{2,1})^{\dagger\dagger} = (U^\dagger O) A_\mathbf{x} (U^\dagger O)^{-1} \; \forall \mathbf{x}$.
  - $\Pr[\mathbf{x}_{1:t}] = \mathbf{b}_\infty^\dagger B_{\mathbf{x}_{t:1}} \mathbf{b}_1 \;\; \forall t, \mathbf{x}$.

Algorithm LearnHMM(m, N):

Inputs: m - number of states, N - sample size

Returns: HMM model parameterized by $\{\widehat{b}_1, \widehat{b}_\infty, \widehat{B}_x \ \forall x \in [n]\}$

1. Independently sample N observation triples (x_1, x_2, x_3) from the HMM to form empirical estimates $\widehat{P}_1, \widehat{P}_{2,1}, \widehat{P}_{3,x,1} \ \forall x \in [n]$ of $P_1, P_{2,1}, P_{3,x,1} \ \forall x \in [n]$.
2. Compute the SVD of $\widehat{P}_{2,1}$, and let $\widehat{U}$ be the matrix of left singular vectors corresponding to the m largest singular values.
3. Compute model parameters:
   (a) $\widehat{b}_1 = \widehat{U}^\top \widehat{P}_1$,
   (b) $\widehat{b}_\infty = (\widehat{P}_{2,1}^\top \widehat{U})^+ P_1$,
   (c) $\widehat{B}_x = \widehat{U}^\top \widehat{P}_{3,x,1} (\widehat{U}^\top \widehat{P}_{2,1})^+ \ \forall x \in [n]$.

- To predict the probability of a sequence:

$$\widehat{\Pr}[x_1, \ldots, x_t] = \widehat{b}_\infty^\top \widehat{B}_{x_t} \ldots \widehat{B}_{x_1} \widehat{b}_1.$$

- Given an observation x_t, the 'internal state' update is:

$$\widehat{b}_{t+1} = \frac{\widehat{B}_{x_t} \widehat{b}_t}{\widehat{b}_\infty^\top \widehat{B}_{x_t} \widehat{b}_t}.$$

- To predict the conditional probability of x_t given $x_{1:t-1}$:

$$\widehat{\Pr}[x_t | x_{1:t-1}] = \frac{\widehat{b}_\infty^\top \widehat{B}_{x_t} \widehat{b}_t}{\sum_x \widehat{b}_\infty^\top \widehat{B}_x \widehat{b}_t}.$$

Remark 5. *If U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then U acts much like the observation probability matrix O in the following sense:*

Given a conditional state $\vec{b}_t$, $\Pr[x_t = i \mid x_{1:t-1}] = [U\vec{b}_t]_i$.	*Given a conditional hidden state $\vec{h}_t$,* $\Pr[x_t = i \mid x_{1:t-1}] = [O\vec{h}_t]_i$.

To see this, note that $UU^\top$ is the projection operator to range(U). *Since* range(U) = range(O) *(Lemma 2), we have $UU^\top O = O$, so $U\vec{b}_t = U(U^\top O)\vec{h}_t = O\vec{h}_t$.*

Simultaneous Diagonalization for Tensor Decomposition

- Let $V = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k]$, $W = \text{diag}(w_1, w_2, \cdots, w_k)$, and $D(\boldsymbol{\eta}) = \text{diag}(\boldsymbol{\mu}_1^\dagger \boldsymbol{\eta}, \boldsymbol{\mu}_2^\dagger \boldsymbol{\eta}, \cdots, \boldsymbol{\mu}_k^\dagger \boldsymbol{\eta})$.
- $M_2 = VWV^\dagger$, $M_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\eta}) = VWD(\boldsymbol{\eta})V^\dagger$.
- Find a matrix X such that $X^\dagger M_2 X$ and $X^\dagger M_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\eta})X$ (for all $\boldsymbol{\eta}$) are diagonal.

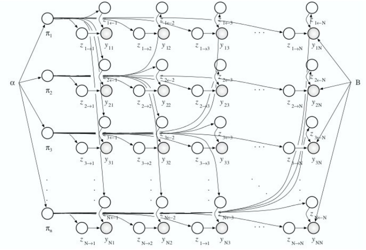
Algorithm I for Simultaneous Diagonalization

- Get $\tilde{M}_2$ and $\tilde{M}_3$ from data.
- Let $\tilde{A}$ and $\tilde{B}$ be the top$-k$ left and right singular vectors of $\tilde{M}_2$.
- Define $C(\boldsymbol{\eta}) = (A^\dagger M_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\eta})B)(A^\dagger M_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\eta})B)^{-1}$.
- Also $C(\boldsymbol{\eta}) = (A^\dagger V)\text{diag}(V^\dagger \boldsymbol{\eta})(A^\dagger V)^{-1}$.
- Get empirical estimate of $C(\boldsymbol{\eta})$.
- It can be shown that $A^\dagger V$ is invertible – eigen decomposition of $C(\boldsymbol{\eta})$ can be performed to recover $A^\dagger V$ and hence V.
- Since $M_2 = VWV^\dagger$, W can be recovered from a knowledge of V.
- Little bit more work needed for multi-view models.
- Algorithm II – two SVDs for LDA.

On the choice of $\boldsymbol{\eta}$

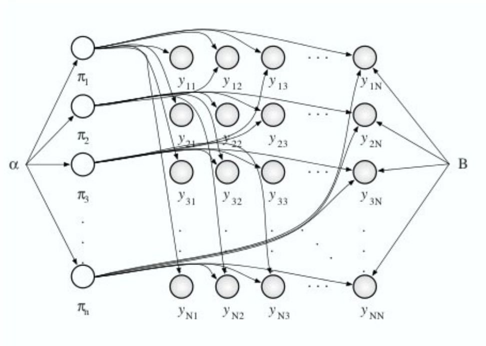
- Components of $\boldsymbol{\eta}$ have to be distinct.
- Can be taken to be a unit basis vector if there is some prior information about the distinct probabilities of a word in topics.
- Else, $\boldsymbol{\eta}$ can be chosen as $\boldsymbol{\eta} = \tilde{A}\boldsymbol{\theta}$ where $\boldsymbol{\theta} \in \mathbb{R}^k$ is a unit vector sampled randomly from a sphere in dimension k.

Mixed Membership Stochastic Block Model

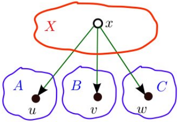


- For each node $p \in \mathcal{N}$:
  - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim$ Dirichlet ($\vec{\alpha}$).
- For each pair of nodes $(p,q) \in \mathcal{N} \times \mathcal{N}$:
  - Draw membership indicator for the initiator, $\vec{z}_{p\rightarrow q} \sim$ Multinomial ($\vec{\pi}_p$).
  - Draw membership indicator for the receiver, $\vec{z}_{q\rightarrow p} \sim$ Multinomial ($\vec{\pi}_q$).
  - Sample the value of their interaction, $Y(p,q) \sim$ Bernoulli ($\vec{z}^{\top}_{p\rightarrow q} B\, \vec{z}_{p\leftarrow q}$).

Fixed Membership Stochastic Block Model



Graph Moments



- Adjacency matrix: G, submatrix going from X to A: $G_{X,A}$, community connectivity matrix P.
- $F = \pi^\dagger P^\dagger \in \mathbb{R}^{n \times k}$, $F_A = \pi_A^\dagger P^\dagger$ denoting the submatrix of F corresponding to nodes in A.
- $T_{X \to \{A,B,C\}} = \frac{1}{|X|} \sum_{i \in X} [G_{i,A}^\dagger \otimes G_{i,B}^\dagger \otimes G_{i,C}^\dagger]$.
- $\mathbb{E}[G_{X,A}^\dagger | \pi_X, \pi_A] = F_A \pi_X$.
- $\mathbb{E}[T_{X \to \{A,B,C\}} | \pi_A, \pi_B, \pi_C] = \sum_{i \in [k]} \hat{\alpha}_i (F_A)_i \otimes (F_B)_i \otimes (F_C)_i$.

Simultaneous Diagonalization

- We have looked at several tensor decomposition methods
- Is there a subclass in which simpler special case formulations exist?

Setup

- Sequence of *exchangeable* RVs: $x_1, x_2, \dots, x_n \in \mathbb{R}^d$
- Latent variable vector: $h \in \mathbb{R}^k$
- Topic matrix: $O \in \mathbb{R}^{dxk}$
- Structure :

$$\mathbb{E}(x_v|h) = Oh$$

- **Goal:** Recover O after observing x_v
- Assumption 0: Some info about distribution of h
- Assumption 1: $d \geq k$
- Assumption 2 : O is full column-rank

Method for Independent (skewed) factor model

- Product distribution: Each h_i independent from the rest
- Variance of h_i: $\sigma_i^2 = \mathbb{E}[(h_i - \mathbb{E}(h_i))^2]$
- Higher moment: $\mu_{i,l} = \mathbb{E}[(h_i - \mathbb{E}(h_i))^l]$
- As before, define moments of $\mathbf{x_v}$:
  - $\mu := \mathbb{E}(\mathbf{x_1})$
  - Pairs $:= \mathbb{E}(\mathbf{x_1} - \mu)(\mathbf{x_2} - \mu)^\dagger$
  - Triples $:= \mathbb{E}[(\mathbf{x_1} - \mu) \otimes (\mathbf{x_2} - \mu) \otimes (\mathbf{x_3} - \mu)]$
  - Triples(η) $:= \mathbb{E}[(\mathbf{x_1} - \mu)(\mathbf{x_2} - \mu)^\dagger \langle (\mathbf{x_3} - \mu), \eta \rangle]$

Identifying structure

- Easy to show relationship between O and and the moments of $\mathbf{x}$ based on $\mathbb{E}(x_v|h) = Oh$
- Pairs $= O\ \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)\ O^\dagger$
- Triples$(\eta) = O\mathrm{diag}(O^\dagger \eta)\ \mathrm{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3})\ O^\dagger$
- Triples structure hints at a possible SVD. Need an "appropriate" η, and O not orthogonal
- 2-svd helps in obtaining it.
- Operate in matrices - tensor decompositions bypassed.

SVD 1 – Whiten Pairs

- $\text{Triples}(\eta) = O\text{diag}(O^\dagger \eta)\ \text{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3})\ O^\dagger$.
- Need η that is not in left null space of O.
- Identifiability issues for any η.
- Assumptions 1 and 2 come into play.
  - Pairs is $d \times d$, but has rank k.
  - $\exists W$ s.t. $W^\dagger \text{Pairs} W = I_{k\,x\,k}$
- Set $\eta = W\theta$, definitely not in left null space of O.
- Claim: For a randomly drawn $\theta \in \mathcal{S}^{k-1}$, an SVD for $W^\dagger \text{Triples}(W\theta) W$ recovers $W^\dagger O$ as singular vectors.

SVD 2

- Can show $W^{\dagger}\text{Triples}(W\theta)W = M\text{diag}(M^{\dagger}\theta)\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)M^{\dagger}$
- γ_i is skewness
- $M = W^{\dagger}O$
- M is orthogonal $\implies$ SVD struct.
- θ is randomly chosen, $M^{\dagger}\theta$ is a rotation
- With probability 1, singular values are unique i.e. singular vectors are identifiable (upto sign and permutation)
- For a random θ, only the singular values change!

Algorithm - Independent Skewed Factors

Algorithm 1 ECA, with skewed factors

Input: vector $\theta \in \mathbb{R}^k$; the moments Pairs and Triples(η)

1. **Dimensionality Reduction:** Find a matrix $U \in \mathbb{R}^{d \times k}$ such that

$$\text{Range}(U) = \text{Range}(\text{Pairs}).$$

   (See Remark 1 for a fast procedure.)

2. **Whiten:** Find $V \in \mathbb{R}^{k \times k}$ so $V^\top (U^\top \text{Pairs}\, U) V$ is the $k \times k$ identity matrix. Set:

$$W = UV$$

3. **SVD:** Let Λ be the set of (left) singular vectors, with *unique* singular values, of

$$W^\top \text{Triples}(W\theta) W$$

4. **Reconstruct:** Return the set $\widehat{O}$:

$$\widehat{O} = \{\ (W^+)^\top \lambda\ : \lambda \in \Lambda\}$$

   where W^+ is the pseudo-inverse (see Eq 1).

Identifiability of O

- Can rescale h and columns of O to get the same model
- Canonicalize : Set $\sigma_i = 1$
- Re-running with different θ recovers upto permutation and sign.
- h being product crucial above, in general we only recover range of O.
- O recovered above identifiable upto sign and permutation of columns

Comments

- $M^{\dagger}\theta$ can have a 0 entry, corresponding O_{*j} not recovered.
  - Rerun.
- What if h is not skewed (third moment is 0).
  - Use fourth moments(kurtosis). Slightly more work, algorithm and proofs similar.
- *Possible Extension* Are there other such classes of models amenable to simpler special case algorithms?
- Can be "embedded" into more complicated models. e.g. LDA, multiview, altered mixture models.
- Moments of x built empirically.
  - For LDA : (properly permuted) With prob $1-\delta$, for $N \geq O(ln(\delta))$

$$||O_{*j} - \hat{O}_{*j}|| \leq O(\frac{\ln(1/\delta)}{N})^{\frac{1}{2}}$$

Example: Coding LDA

- Each x_v is the v^{th} word in a document
- d is vocabulary size
- $x_v = e_j$ if x_v is the j^{th} word from the vocabulary.
- $h \in \Delta^{k-1}$. Distributed Dirichlet($\boldsymbol{\alpha}$).
- O_{*j} = word distribution of j^{th} topic.
- $\Pr([x_v]_j = 1|h) = [Oh]_j \implies E(x_v|h) = Oh$
- Parameter $\alpha_0 = \sum_k \alpha_k$ supplied externally

- Define moments
  - $\text{Pairs}_{\alpha_0} := \mathbb{E}(x_1 x_2^\dagger) - \frac{\alpha_0}{\alpha_0+1}\mu\mu^\dagger$
  - $\text{Triples}_{\alpha_0}(\eta) :=$
    $\mathbb{E}(x_1 x_2^\dagger \langle \eta, x_3 \rangle) - \frac{\alpha_0}{\alpha_0+2}\left(\mathbb{E}[x_1 x_2^\dagger]\eta\mu^\dagger + \mu\eta^\dagger\mathbb{E}[x_1 x_2^\dagger] + \langle \eta, \mu \rangle \mathbb{E}[x_1 x_2^\dagger]\right) +$
    $\frac{2\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\langle \eta, \mu \rangle \mu\mu^\dagger$
- Structure
  - $\text{Pairs}_{\alpha_0} = \frac{1}{(\alpha_0+1)\alpha_0} O\text{diag}(\boldsymbol{\alpha})O^\dagger$
  - $\text{Triples}_{\alpha_0}(\eta) = \frac{1}{(\alpha_0+1)(\alpha_0+2)\alpha 0} O\text{diag}(O^\dagger\eta)\text{diag}(\boldsymbol{\alpha})O^\dagger$

Algorithm for LDA

Algorithm 3 ECA for latent Dirichlet allocation

Input: a vector $\theta \in \mathbb{R}^k$; the moments Pairs_{α_0} and $\text{Triples}_{\alpha_0}$

1. **Dimensionality Reduction:** Find a matrix $U \in \mathbb{R}^{d \times k}$ such that

$$\text{Range}(U) = \text{Range}(\text{Pairs}_{\alpha_0}).$$

   (See Remark 1 for a fast procedure.)

2. **Whiten:** Find $V \in \mathbb{R}^{k \times k}$ so $V^\top (U^\top \text{Pairs}_{\alpha_0} U) V$ is the $k \times k$ identity matrix. Set:

$$W = UV$$

3. **SVD:** Let Λ be the set of (left) singular vectors, with *unique* singular values, of

$$W^\top \text{Triples}_{\alpha_0}(W\theta)W$$

4. **Reconstruct and Normalize:** Return the set $\widehat{O}$:

$$\widehat{O} = \left\{ \frac{(W^+)^\top \lambda}{\vec{1}^\top (W^+)^\top \lambda} : \lambda \in \Lambda \right\}$$

   where $\vec{1} \in \mathbb{R}^d$ is a vector of all ones and W^+ is the pseudo-inverse (see Eq 1).

Multiview extension

- O is not the same for every v
- $\mathbb{E}[x_v|h] = O_v h$
- Define moments. For $v \in 1, 2, 3$,
  - $\text{Pairs}_{v,v'} := \mathbb{E}[(x_v - \mu)(x_{v'} - \mu)^\dagger]$
  - $\text{Triples}_{132}(\eta) := \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\dagger \langle \eta, x_3 - \mu \rangle]$
- Structure - For $v \in 1, 2, 3$,
  - $\text{Pairs}_{v,v'} = O_v \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) O_{v'}^\dagger$
  - $\text{Triples}_{132}(\eta) = O_1 \text{diag}(O_3^\dagger) \text{diag}(\mu_{1,3}, \ldots, \mu_{k,3}) O_2^\dagger$
- Generate a single view from the three views and use the first algorithm.

Multiview Algorithm

Algorithm 4 ECA; the multi-view case

Input: vector $\theta \in \mathbb{R}^k$; the moments $\mathrm{Pairs}_{v,v'}$ and $\mathrm{Triples}_{132}(\eta)$

1. **Project views 1 and 2:** Find matrices $A \in \mathbb{R}^{k \times d_1}$ and $B \in \mathbb{R}^{k \times d_2}$ such that $A\,\mathrm{Pairs}_{12}\,B^\top$ is invertible. Set:

$$\begin{aligned}
\widetilde{\mathrm{Pairs}}_{12} &:= A\,\mathrm{Pairs}_{12}\,B^\top \\
\widetilde{\mathrm{Pairs}}_{31} &:= \mathrm{Pairs}_{31}\,A^\top \\
\widetilde{\mathrm{Pairs}}_{32} &:= \mathrm{Pairs}_{32}\,B^\top \\
\widetilde{\mathrm{Triples}}_{132}(\eta) &:= A\,\mathrm{Triples}_{132}(\eta) B^\top
\end{aligned}$$

   (See Remark 10 for a fast procedure.)

2. **Symmetrize:** Reduce to a single view:

$$\begin{aligned}
\mathrm{Pairs}_3 &:= \widetilde{\mathrm{Pairs}}_{31}(\widetilde{\mathrm{Pairs}}_{12}^\top)^{-1}\widetilde{\mathrm{Pairs}}_{23} \\
\mathrm{Triples}_3(\eta) &:= \widetilde{\mathrm{Pairs}}_{32}(\widetilde{\mathrm{Pairs}}_{12})^{-1}\widetilde{\mathrm{Triples}}_{132}(\eta)(\widetilde{\mathrm{Pairs}}_{12})^{-1}\widetilde{\mathrm{Pairs}}_{13}
\end{aligned}$$

3. **Estimate O_3 with ECA:** Call Algorithm 1, with θ, Pairs_3, and $\mathrm{Triples}_3(\eta)$.

New Research Direction

- Online implementation of tensor decomposition – application – large scale learning of topic models.
- Constrained tensor decomposition based on auxiliary/side information – application – supervised topic models.
- More efficient implementation (optimization) of tensor decomposition.

References

- Tensor decompositions for learning latent variable models.
- A Spectral Algorithm for Learning Hidden Markov Models.
- A Method of Moments for Mixture Models and Hidden Markov Models.
- A Tensor Spectral Approach to Learning Mixed Membership Community Models.
- A Spectral Algorithm for Latent Dirichlet Allocation.
- NIPS12 Spectral Learning Workshop.