Spectral Methods
Outline

● Example Applications
● Spectral Methods - PCA
● Latent Variable Models - Gaussian Mixture Model
● Tensor factorization
● Eigen Analysis
● Conclusion
Applications

- Gaussian mixture models
- Hidden Markov Models
- Community Detection
- Topic Models
- Recommender systems
- Feature Learning
Latent Variable Models

Difficulties in learning:

- Identifiability
- Maximum likelihood is NP-hard
- Practice: EM, Variational Bayes have no consistency guarantees.
- Efficient computational and sample complexities
PCA - Spectral method on covariance matrices

Optimization problem

For (centered) points $x_i \in \mathbb{R}^d$, find projection $P$ with $\text{Rank}(P) = k$ s.t.

$$\min_{P \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i \in [n]} \|x_i - Px_i\|^2.$$  

Result: If $S = \text{Cov}(X)$ and $S = U \Lambda U^\top$ is eigen decomposition, we have $P = U_{(k)} U_{(k)}^\top$, where $U_{(k)}$ are top-$k$ eigen vectors.
Gaussian mixture models

- \( k \) Gaussians: each sample is \( x = Ah + z \).
- \( h \in \left[ e_1, \ldots, e_k \right] \), the basis vectors. \( \mathbb{E}[h] = w \).
- \( A \in \mathbb{R}^{d \times k} \): columns are component means.
- Let \( \mu := Aw \) be the mean.
- \( z \sim \mathcal{N}(0, \sigma^2 I) \) is white Gaussian noise.
Gaussian mixture models

\[
\mathbb{E}[(x - \mu)(x - \mu)^\top] = \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^\top + \sigma^2 I.
\]

Aim: Given the points \( x \), learn \( A \)
Conventional Method: Expectation Maximization
Problem: Converges to local minima
Idea: Use higher order moments
Higher order moments for GMM

For the GMM example,

\[ \mathbb{E}[x \otimes x \otimes x] = \sum_i w_i a_i \otimes a_i \otimes a_i + \sigma^2 \sum_i (\mu \otimes e_i \otimes e_i + \ldots) \]

\[ M_3 = \sum_i w_i a_i \otimes a_i \otimes a_i \]

\[ M_2 = \sum_i w_i a_i \otimes a_i. \]
Tensor factorization

Multilinear transformation of tensor

\[ M_3(B, C, D) := \sum_i w_i (B^\top a_i) \cdot (C^\top a_i) \cdot (D^\top a_i) \]

If the columns of A are orthogonal,

\[ M_3(I, a_1, a_1) = \sum_i w_i \langle a_i, a_1 \rangle^2 a_i = w_1 a_1 \]

\( a_i \) are eigenvectors of tensor \( M_3 \)
Whitening

Problem: A is not orthogonal in general
Solution:

Find whitening matrix $W$ s.t. $W^\top A = V$ is an orthogonal matrix.

$$T = M_3(W, W, W) = \sum_i w_i (W^\top a_i)^\otimes 3 = \sum_{i \in [k]} w_i \cdot v_i \otimes v_i \otimes v_i$$
Whitening

\[
M_2 = U \text{Diag}(\tilde{\lambda}) U^\top \quad \tilde{W} = U \text{Diag}(\tilde{\lambda}^{-1/2})
\]

V is an orthogonal matrix; T is an orthogonal tensor.

\[
T(I, v_1, v_1) = \sum_i \lambda_i \langle v_i, v_1 \rangle^2 v_i = \lambda_1 v_1
\]

\(v_i\) are eigenvectors of tensor \(T\).
Tensor power method

- Randomly initialize the power method. Run to convergence to obtain $v$ with eigenvalue $\lambda$.
- Deflate: $T - \lambda v \otimes v \otimes v$ and repeat.

Is there convergence? Does the convergence depend on initialization?
Matrix EigenAnalysis

Eigen vectors are fixed points: $Mv = \lambda v$

Uniqueness (Identifiability): Iff. $\lambda_i$ are distinct.

Power method: $v \mapsto \frac{M(I,v)}{\|M(I,v)\|}$

$v_1$ is the only local optimum

Let initialization $v = \sum_i c_i v_i$.

If $c_1 \neq 0$, power method converges to $v_1$
Tensor EigenAnalysis

Bad news:

- Decomposition may not always exist for general tensors.
- Finding the decomposition is NP hard in general.

For an orthogonal tensor, no spurious local optima! \( \{v_i\} \) are the only local optima.

Converges to \( v_i \) for which \( v_i|c_i| = \max! \) could be any of them.
Tensor EigenAnalysis

- Matrix power method - Linear convergence;
- Tensor power method - Quadratic convergence
- Matrix power method: Requires gap between largest and second-largest eigenvalue
- Tensor power method: Requires gap between largest and second-largest $\lambda_i c_i$
- Tensor Power method - robust to noise
Putting it together

- Gaussian mixture: \( x = Ah + z \), where \( \mathbb{E}[h] = w \).
- \( z \sim \mathcal{N}(0, \sigma^2 I) \).

\[
M_2 = \sum_i w_i a_i \otimes a_i, \quad M_3 = \sum_i w_i a_i \otimes a_i \otimes a_i.
\]

- Obtain whitening matrix \( W \) from SVD of \( M_2 \).
- Use \( W \) for multilinear transform: \( T = M_3(W, W, W) \).
- Find eigenvectors of \( T \) through power method and deflation.
Conclusion

- Good method for guaranteed convergence to global minima (not guaranteed by EM)
- Numerous applications to latent variable models
- Scalability issues: requires computing SVDs of large matrices. Storage and decomposition of large tensors. In practice: use SGD techniques. Don’t know if there are guarantees.
- Weak robustness results
- Higher sample complexity