IFT 6760A - Lecture 13
Latent Variable Models and Method of Moments

Scribe(s): Bogdan Mazoure, Alex Zhang, Adam Ibrahim **Instructor:** Guillaume Rabusseau

1 Summary

This lecture covers latent variable models, in which observed variables are linked with some hidden representation which we want to infer. As opposed to traditional maximum likelihood methods such as Expectation Maximization, we might opt for a parametric Method of Moments (MoM) approach[1]. As an example, Gaussian mixture models and single topics models are used to show the method.

2 Latent Variable Models

A latent variable model is a statistical model which relates a set of observed variables to a set of latent, or hidden variables.
Notable examples of latent variable models include hidden Markov models (HMM) and mixture of Gaussians. A concrete example of a mixture of multinomial (or categorical) distributions widely used in NLP is the single topic model discussed below.

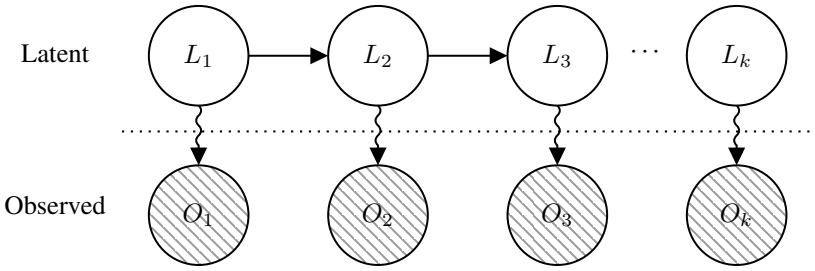

Figure 1: Hidden Markov model with k observations and k latent variables.

2.1 Single topic model

Let the topic be a discrete random variable taking values $1, .., k$ with probability $w_1, .., w_k$. Moreover, let the vocabulary $\mathcal{V} = \{v_1, .., v_d\}$ be a set of d words. To each value of the topic, we associate a distribution over the vocabulary $\boldsymbol{\mu}_h \in \mathbb{R}^d,\ h = 1, .., k$.

Definition 1 (Probability simplex)**.** *The $d-$dimensional probability simplex is defined to be the set:*

$$\Delta^d = \left\{ \mathbf{u} \in \mathbb{R}^d \middle| \sum_{i=1}^{d} \mathbf{u}_i = 1, \mathbf{u}_i \geq 0, i = 1, .., d \right\} \tag{1}$$

A vector $\mathbf{u} \in \mathbb{R}^d$ is said to lie on the probability simplex if $\mathbf{u} \in \Delta^d$. Since each $\boldsymbol{\mu}_h$ is a probability distribution by definition, it hence lies on a $d-$dimensional simplex.
Finally, a document of length l is a collection of l words $v_1, v_2, \cdots, v_l \in \mathcal{V}$. By assumption, every document can follow only one topic to enforce a simple latent variable model.
In order to sample a a document of length l, one proceeds as follows:

- Draw a random topic $h \sim \text{Categorical}(w_1, .., w_k)$, where $\mathbb{P}(h = i) = w_i$;
- Draw l words independently and identically distributed (iid) from the corresponding vocabulary distribution $\boldsymbol{\mu}_h$.

In the single topic model, the parameter set consists of the topic weights and vocabulary distributions, that is $\mathbf{w} = \{w_1, .., w_k\} \in \Delta^k$ and $\boldsymbol{\mu}_1, .., \boldsymbol{\mu}_k \in \Delta^d$. The documents are considered as observations, while the topic corresponding to each document is a latent variable. The challenge is to recover the topic corresponding to a given document.
It is important to observe here that the words in a document are not independent. However, conditioned on a given topic, they become independent: if x_1 and x_2 denote the first two words in the document, we have

$$\mathbb{P}[x_1 = i, x_2 = i] \neq \mathbb{P}[x_1 = i]\,\mathbb{P}[x_2 = i] \quad \text{but} \quad \mathbb{P}[x_1 = i, x_2 = i \mid \text{topic} = h] = \mathbb{P}[x_1 = i \mid \text{topic} = h]\,\mathbb{P}[x_2 = i \mid \text{topic} = h]$$

2.2 Mixture of spherical Gaussians

Assume, as in the single topic model, a discrete component random variable taking values from 1 to k with probability $w_1, .., w_k$. The probabilities are also known as mixing weights. We consider a mixture of k Gaussians with different means $\boldsymbol{\mu}_1, .., \boldsymbol{\mu}_k \in \mathbb{R}^d$ and identical covariance matrix $\sigma^2\mathbf{I} \in \mathbb{R}^{d\times d}$.
Drawing an observation from this model can be performed in two steps:

- Pick a Gaussian component $h \sim \text{Categorical}(w_1, .., w_k)$;
- Draw $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_h, \sigma^2\mathbf{I})$.

Note how in both settings, learning the model is equivalent to learning the mixing weights (in both cases, a categorical distribution) and the component parameters (which is a categorical with d parameters in the single topic and a Gaussian with two parameters in the GMM setting).

3 Method of Moments

Recall that statistical inference techniques can be classified in roughly three categories: maximum likelihood estimation (MLE), (generalized) method of momends (MoM) and Bayesian estimation of the posterior.
Below is a recall of the definition of maximum likelihood estimation for discrete random variables:

Definition 2 (Maximum Likelihood Estimation (MLE))**.** *Let X be a random variable with probability function[a] f_θ parameterized by a parameter vector θ. Assume that a collection of independent and identically distributed observations $S = \{x_1, .., x_N\}$ drawn from f_θ is observed. Then, the MLE estimate of θ can be found through maximization of the joint probability of S:*

$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ f_\theta[S] = \underset{\theta}{\text{argmax}} \prod_{i=1}^{N} f_\theta[x_i] \tag{2}$$

[a] f_θ is either the probability mass function of if X is a discrete random variable, or the probability density function if X is continuous.

MLE involves local searches, which may yield suboptimal solutions. However, the method of moments yields global solutions if the system of equations involved can be solved.

Definition 3 (Method of Moments[2]). *Let X be a random variable with a parameterized probability function f_θ. Let the parameter set be a k-dimensional vector θ, such that the k first moments of the distribution $\mu_1 = \mathbb{E}[X], \mu_2 = \mathbb{E}[X^2], ...$ can be computed from θ by solving the system of equations*

$$\begin{cases} \mu_1 = g_1(\theta) \\ \mu_2 = g_2(\theta) \\ \quad \vdots \\ \mu_k = g_k(\theta) \end{cases}$$

The method of moments (MoM) consists in approximating θ by computing the k first moments $\hat{\mu}_i$ from a sample of the data, and solving for $\hat{\theta}$ in the system of equations given by $\hat{\mu}_i = g_i(\hat{\theta})$ for $i = 1, ..., k$.

As an example, let us look at a Gaussian distribution, in which case we have $\theta = (\mu, \sigma^2)$. If $x \sim \mathcal{N}(\mu, \sigma^2)$, then

$$g_1(\mu, \sigma^2) \triangleq \mathbb{E}[x] = \mu$$
$$g_2(\mu, \sigma^2) \triangleq \mathbb{E}[x^2] = \mu^2 + \sigma^2$$

Using the method of moments and a training set $S = \{x_1, ..., x_n\}$ drawn i.i.d from $\mathcal{N}(\mu, \sigma^2)$, we find $\hat{\mu}, \hat{\sigma}^2$ by solving the system of equations:

$$\hat{\mathbb{E}}_S[x] = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$$
$$\hat{\mathbb{E}}_S[x^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$$

which yields $\hat{\mu} = \hat{\mathbb{E}}_s[x]$ and $\hat{\sigma}^2 = \hat{\mathbb{E}}_s[x^2] - \hat{\mu}^2$.
Remark: While method of moments is a parametric estimation technique, non-parametric extensions with the use of kernel density estimates have been studied [4].

3.1 Single topic model with Method of Moments

Within the scope of the problem, we let $\theta = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, w_1, ..., w_k\}$ be the parameters of a single topic model (which is parameterized by the vocabulary distributions for each topic and mixing probabilities). Given a document, we make use of the conditional independence of words in the document given a known topic h:

$$\begin{aligned} \mathbb{P}_\theta\left[\textit{words i, j} \mid \textit{topic h}\right] &= \mathbb{P}_\theta\left[\textit{word i} \mid \textit{topic h}\right] \mathbb{P}_\theta\left[\textit{word j} \mid \textit{topic h}\right] \\ &= (\boldsymbol{\mu}_h)_i (\boldsymbol{\mu}_h)_j \\ &= (\boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h)_{i,j} \end{aligned} \tag{3}$$

Thus,

$$\begin{aligned} \mathbb{P}_\theta[\textit{words i, j}] &= \sum_{h=1}^{k} \mathbb{P}_\theta\left[\textit{word i, j} \mid \textit{topic h}\right] \underbrace{\mathbb{P}_\theta[\textit{topic h}]}_{w_h} \\ &= \sum_{h=1}^{k} w_h (\boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h)_{i,j} \\ &\triangleq \mathbf{M}_{i,j}, \end{aligned} \tag{4}$$

where we marginalize over the unknown topic.
Similarly,

$$\begin{aligned} \mathbb{P}_\theta[\textit{words } i_1, i_2, i_3] &= \sum_{h=1}^{k} w_k (\boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h)_{i_1,i_2,i_3} \\ &\triangleq \boldsymbol{\mathcal{T}}_{i_1,i_2,i_3} \end{aligned} \tag{5}$$

We can estimate $\mathbf{M} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d\times d\times d}$ from data and then try to solve the system.

$$\begin{cases}\hat{\mathbf{M}} \approx \sum_{h=1}^{k} \hat{w}_h \, (\hat{\boldsymbol{\mu}}_h \circ \hat{\boldsymbol{\mu}}_h) \\ \hat{\boldsymbol{\mathcal{T}}} \approx \sum_{h=1}^{k} \hat{w}_h \, (\hat{\boldsymbol{\mu}}_h \circ \hat{\boldsymbol{\mu}}_h \circ \hat{\boldsymbol{\mu}}_h)\end{cases} \tag{6}$$

Note that if $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^d$ denote the one-hot encoding of the first 3 words in a document, we have $\mathbf{M} = \mathbb{E}[\mathbf{x}_1 \circ \mathbf{x}_2]$ and $\boldsymbol{\mathcal{T}} = \mathbb{E}[\mathbf{x}_1 \circ \mathbf{x}_2 \circ \mathbf{x}_3]$.

4 Tensor Method of Moments

Re-using the notation from the single topic example, we aim to solve the following system of equations

$$\mathbf{M}_{i,j} = \sum_{h=1}^{k} w_h (\boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h)_{i,j} \in \mathbb{R}^{d\times d} \tag{7}$$

$$\boldsymbol{\mathcal{T}}_{i_1,i_2,i_3} = \sum_{h=1}^{k} w_h (\boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h \circ \boldsymbol{\mu}_h)_{i_1,i_2,i_3} \in \mathbb{R}^{d\times d\times d} \tag{8}$$

for $(w_1, .., w_k) \in \Delta^k, \boldsymbol{\mu}_1, .., \boldsymbol{\mu}_k \in \mathbb{R}^d$.

4.1 Jennrich's algorithm

When the vectors $\boldsymbol{\mu}_i$ are linearly independent, this system of equations can be solved using Jennrich's algorithm by operating directly on the order 3 tensor $\boldsymbol{\mathcal{T}}$. In the classical derivation, $\boldsymbol{\mu}$ from (8) is assumed to be full column rank. First, write

$$\begin{aligned}\boldsymbol{\mathcal{T}} &= \sum_{h=1}^{k} (w_h^{1/3} \boldsymbol{\mu}_h) \circ (w_h^{1/3} \boldsymbol{\mu}_h) \circ (w_h^{1/3} \boldsymbol{\mu}_h) \\ &= [\![\mathbf{A}, \mathbf{A}, \mathbf{A}]\!],\end{aligned} \tag{9}$$

where

$$\mathbf{A} = \begin{bmatrix} w_1^{1/3} \boldsymbol{\mu}_1 & \cdots & w_k^{1/3} \boldsymbol{\mu}_k \end{bmatrix}$$

The idea behind Jennrich's algorithm is to take the inner product along some dimension with two random vectors (in order to guarantee uniqueness of the diagonal matrix), and then run an SVD on their product, which would recover the values of $\boldsymbol{\mu}$.
For instance, let $\mathbf{x}, \mathbf{y}$ be two independently sampled noise vectors. Then,

$$\begin{aligned}\boldsymbol{\mathcal{T}} \bullet_3 \mathbf{x} &= \mathbf{A} \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{A}^T \\ \boldsymbol{\mathcal{T}} \bullet_3 \mathbf{y} &= \mathbf{A} \boldsymbol{\Lambda}_{\mathbf{y}} \mathbf{A}^T,\end{aligned} \tag{10}$$

where

$$(\boldsymbol{\Lambda}_{\mathbf{x}})_{i,j} = \begin{cases} \langle \mathbf{x}_i, w_j^{1/3} \boldsymbol{\mu}_j \rangle & i = j \\ 0 & i \neq j \end{cases}$$

If we examine the quantity $\big(\boldsymbol{\mathcal{T}} \bullet_3 \mathbf{x}\big)\big(\boldsymbol{\mathcal{T}} \bullet_3 \mathbf{y}\big)^\dagger$, it simplifies to

$$\begin{aligned}\big(\boldsymbol{\mathcal{T}} \bullet_3 \mathbf{x}\big)\big(\boldsymbol{\mathcal{T}} \bullet_3 \mathbf{y}\big)^\dagger &= \mathbf{A} \boldsymbol{\Lambda}_{\mathbf{x}} \boldsymbol{\Lambda}_{\mathbf{y}}^\dagger \mathbf{A}^\dagger \\ &= \mathbf{A} \boldsymbol{\Lambda} \mathbf{A}^\dagger\end{aligned} \tag{11}$$

where we combine $\boldsymbol{\Lambda}_{\mathbf{x}} \boldsymbol{\Lambda}_{\mathbf{y}} = \boldsymbol{\Lambda}$. The eigenvectors of the matrix in (11) are in fact the columns of $\mathbf{A}$, allowing us to recover $\pm w_h^{1/3} \boldsymbol{\mu}_h$. In the single topic model case, we can stop at this point, since we know that each $\boldsymbol{\mu}_h$ lies on a simplex and hence separating them from w_h can be done through re-normalization. However, this approach does not allow to recover the $\boldsymbol{\mu}_i$ in the Gaussian mixture case. In the next sections, we will show how knowledge of the 2nd order moment matrix $\mathbf{M}$ can help us deal with this case.

4.2 Orthogonalization of $\mathcal{T}$

Remark: Let $\mathcal{T} = \sum_{h=1}^{k} \lambda_h \mathbf{a}_h \circ \mathbf{a}_h \circ \mathbf{a}_h$ such that $\langle \mathbf{a}_i, \mathbf{a}_j \rangle = 0$ for all $i \neq j$ and 1 otherwise (pairwise orthonormal). Then, it is possible to recover the weights λ_h and vectors $\mathbf{a}_h$ by a simple eigendecomposition of $\mathcal{T} \bullet_3 \mathbf{x}$ for some random vector $\mathbf{x}$. Indeed, in this case the eigenvectors of $\mathcal{T} \bullet_3 \mathbf{x}$ are the columns of $\mathbf{A}$.
Claim: Let $\mathbf{M}$ and $\mathcal{T}$ be defined as in (4) and (5), respectively. If we let $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be the eigendecomposition of $\mathbf{M}$ and $\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{U}^T$, then

$$\tilde{\mathcal{T}} = \mathcal{T} \times_1 \mathbf{W} \times_2 \mathbf{W} \times_3 \mathbf{W} = \sum_{h=1}^{k} \tilde{w}_h \tilde{\boldsymbol{\mu}}_h \circ \tilde{\boldsymbol{\mu}}_h \circ \tilde{\boldsymbol{\mu}}_h \quad \in \mathbb{R}^{k \times k \times k}, \tag{12}$$

where $\tilde{w}_h = \frac{1}{\sqrt{w_h}}$ and $\tilde{\boldsymbol{\mu}}_h = \sqrt{w_h}\, \mathbf{W}_{\boldsymbol{\mu}_h}$, **is an orthogonal decomposition**, i.e., the $\tilde{\boldsymbol{\mu}}_h$ are pairwise orthogonal (proof is outlined below).

Proof. Let $\tilde{\mathbf{U}} \in \mathbb{R}^{k \times k}$ be the matrix having the vectors $\tilde{\boldsymbol{\mu}}_h$ as columns. We have

$$\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \sum_{h=1}^{k} \tilde{\boldsymbol{\mu}}_h \tilde{\boldsymbol{\mu}}_h^\top = \sum_{h=1}^{k} \mathbf{W}\boldsymbol{\mu}\boldsymbol{\mu}^\top\mathbf{W}^\top = \mathbf{W}\mathbf{M}\mathbf{W}^T = \mathbf{I}$$

and since $\tilde{\mathbf{U}}$ is square we also have $\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}$ which shows the claim. □

Thus we can solve the system of equations (7)-(8) in the general case as follows:

1. use the second order moment matrix $\mathbf{M}$ to orthogonalize the tensor $\mathcal{T}$ to obtain $\tilde{\mathcal{T}}$
2. recover the vectors $\tilde{\boldsymbol{\mu}}_i$ and weights $\tilde{w}_i$ using an eigendecomposition of $\tilde{\mathcal{T}} \bullet_3 \mathbf{x}$ for some random vector $\mathbf{x}$
3. recover the original weights w_i and $\boldsymbol{\mu}_i$ from the $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{w}_i$

In practice, step (2) can be very unstable and sensitive to noise. Better robustness can be obtained by performing simultaneous diagonalization of several random projections [3] or by using the robust tensor power method proposed in [1].

References

[1] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[2] K. Bowman and L. Shenton. Estimation: Method of moments. *Encyclopedia of statistical sciences*, 3, 2004.

[3] V. Kuleshov, A. Chaganty, and P. Liang. Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pages 507–516, 2015.

[4] A. Lewbel. A local generalized method of moments estimator. *Economics Letters*, 94(1):124–128, 2007.