Method of Moments for Mixture of Gaussians

Yao Zhu*
Derivatives Pricing, Bloomberg LP

1 Mixture of Spherical Gaussians with the same variance parameter σ^2

For clarity of presenting the idea of method of moments, we consider the mixture of spherical Gaussians with the same covariance. Under this setting, a random variable $X \in R^d$ is generated as follows:

$$\begin{aligned} p(Z = i) &= w_i \\ X|Z = i &\sim \mathcal{N}(\mu_i, \sigma^2 I_d) \end{aligned} \tag{1}$$

for $i = 1, \dots, k$, where k is a known number and $k \leq d$, I_d is the $d \times d$ identity matrix, and σ^2 is the variance parameter common to all k components. We further assume the following non-degeneracy condition:

Non-degeneracy Condition: $\mu_i, i = 1, \dots, k$ are linearly independent, and $w_i > 0$ for all $i = 1, \dots, k$.

We target to recover the parameters $\{w_i, \mu_i\}_{i=1}^k$ and σ^2 from a sample of data points $\{x_n\}_{n=1}^N$ generated according to (1).

2 Raw Moments

2.1 Tensor product

In order to work with moments (especially moments of order higher than 2), we need to the notion of *tensor product*, denoted by $\otimes$. Let $X_1, X_2, X_3 \in R^d$, we define $X_1 \otimes X_2 \otimes X_3$ such that

$$(X_1 \otimes X_2 \otimes X_3)_{jlm} = (X_1)_j (X_2)_l (X_3)_m \tag{2}$$

for $j, l, m = 1, \dots, d$. In a similar style, we define $X_1 \otimes X_2 = X_1 X_2^T$.

*yzhu221@bloomberg.net

2.2 Moments of Mixture of Gaussians

Given (1), when $Z = i$ we can write and

$$\begin{aligned} X &= \mu_i + Y \\ Y &\sim \mathcal{N}(0, \sigma^2 I_d) \end{aligned} \tag{3}$$

We can compute the following raw moments up to order 3

$$\mathbb{E}[X] = \sum_{i=1}^{k} w_i \mu_i \tag{4}$$

We will denote $M_1 = \mathbb{E}[X]$ in the following.

$$\begin{aligned} \mathbb{E}[X \otimes X] &= \sum_{i=1}^{k} w_i \mathbb{E}[X \otimes X | Z = i] && (5) \\ &= \sum_{i=1}^{k} w_i \mathbb{E}[(\mu_i + Y) \otimes (\mu_i + Y)] && (6) \\ &= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i + \sigma^2 I_d && (7) \end{aligned}$$

Note we have used the fact that $\mathbb{E}[Y] = 0$, and $\mathbb{E}[Y \otimes Y] = \sigma^2 I_d$. Plus the fact that $\mathbb{E}[Y \otimes Y \otimes Y] = 0$, we compute the 3rd order moment

$$\begin{aligned} \mathbb{E}[X \otimes X \otimes X] &= \sum_{i=1}^{k} w_i \mathbb{E}[X \otimes X \otimes X | Z = i] && (8) \\ &= \sum_{i=1}^{k} w_i \mathbb{E}[(\mu_i + Y) \otimes (\mu_i + Y) \otimes (\mu_i + Y)] && (9) \\ &= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i + && (10) \\ &\sum_{i=1}^{k} w_i \left(\mathbb{E}[\mu_i \otimes Y \otimes Y] + \mathbb{E}[Y \otimes \mu_i \otimes Y] + \mathbb{E}[Y \otimes Y \otimes \mu_i] \right) \end{aligned}$$

Now let's take a closer look at the term

$$\begin{aligned} \sum_{i=1}^{k} w_i \mathbb{E}[Y \otimes \mu_i \otimes Y] &= \mathbb{E}[Y \otimes (\sum_{i=1}^{k} w_i \mu_i) \otimes Y] \\ &= \mathbb{E}[Y \otimes M_1 \otimes Y] \end{aligned}$$

In order to further simplify it, we look at a particular cell

$$\begin{aligned} \mathbb{E}[(Y \otimes M_1 \otimes Y)_{jlm}] &= (M_1)_l \mathbb{E}[Y_j Y_m] \\ &= (M_1)_l \sigma^2 \delta_{jm} \end{aligned}$$

where δ_{jm} is the Kronecker delta. Thus, in tensor form we have

$$\mathbb{E}[Y \otimes M_1 \otimes Y] = \sigma^2 \sum_{j=1}^{d} e_j \otimes M_1 \otimes e_j \tag{11}$$

where $\{e_1, \ldots, e_d\}$ is the canonical basis of d dimension. Similarly we have

$$\mathbb{E}[M_1 \otimes Y \otimes Y] = \sigma^2 \sum_{j=1}^{d} M_1 \otimes e_j \otimes e_j \tag{12}$$

$$\mathbb{E}[Y \otimes Y \otimes M_1] = \sigma^2 \sum_{j=1}^{d} e_j \otimes e_j \otimes M_1 \tag{13}$$

Thus, in summary we have

$$\begin{aligned}\mathbb{E}[X \otimes X \otimes X] = &\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i + \\ &\sigma^2 \sum_{j=1}^{d} (M_1 \otimes e_j \otimes e_j + e_j \otimes M_1 \otimes e_j + e_j \otimes e_j \otimes M_1)\end{aligned} \tag{14}$$

3 Parameter identification from the moments

From the data sample $\{x_n\}_{n=1}^{N}$, we can compute the empirical moments $\widetilde{\mathbb{E}}[X]$, $\widetilde{\mathbb{E}}[X \otimes X]$, and $\widetilde{\mathbb{E}}[X \otimes X \otimes X]$ as estimates of the theoretical moments. For this reason, we say $\mathbb{E}[X]$, $\mathbb{E}[X \otimes X]$, and $\mathbb{E}[X \otimes X \otimes X]$ are observable. We want to come up with a recipe to identify the parameters $\{w_i, \mu_i\}_{i=1}^{k}$ and σ^2 from these observable moments.

3.1 Identify σ^2

Let's compute the covariance matrix of X

$$\begin{aligned} cov(X) &= \mathbb{E}[(X - M_1) \otimes (X - M_1)] && (15) \\ &= \sum_{i=1}^{k} w_i \mathbb{E}[(X - M_1) \otimes (X - M_1) | Z = i] && (16) \\ &= \sum_{i=1}^{k} w_i \mathbb{E}[(\mu_i - M_1 + Y) \otimes (\mu_i - M_1 + Y)] && (17) \\ &= \sum_{i=1}^{k} w_i (\mu_i - M_1) \otimes (\mu_i - M_1) + \sigma^2 I_d && (18) \end{aligned}$$

Note that because $\sum_{i=1}^{k} w_i(\mu_i - M_1) = 0$, the k vectors $(\mu_i - M_1)$ for $i = 1, \ldots, k$ are linearly dependent. Thus, from (18) we know $\sigma^2 = \lambda_{\min}(cov(X))$, i.e., σ^2 is the smallest eigenvalue of $cov(X)$. Note because $cov(X) = \mathbb{E}[X \otimes X] - M_1 \otimes M_1$, $cov(X)$ is also observable.

3.2 Identify $\{w_i, \mu_i\}_{i=1}^k$

We define the following two purified moments

$$M_2 = \mathbb{E}[X \otimes X] - \sigma^2 I_d \tag{19}$$

$$= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \tag{20}$$

$$M_3 = \mathbb{E}[X \otimes X \otimes X] - \sigma^2 \sum_{j=1}^{d} (M_1 \otimes e_j \otimes e_j + e_j \otimes M_1 \otimes e_j + e_j \otimes e_j \otimes M_1) \tag{21}$$

$$= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \tag{22}$$

Once we have identified σ^2, M_2 is observable thanks to equation (19), and M_3 is observable thanks to equation (21). In the **Non-degeneracy Condition**, we only assume $\mu_i, i = 1, \ldots, k$ to be linearly independent, which is not strong enough for us to extract μ_i directly through the tensor decomposition in equation (22). We want to cook up another tensor that admits an *orthogonal tensor decomposition*, on which we can apply the *tensor power method*. From (20), we see that M_2 is a symmetric positive semidefinite matrix with rank k. Thus, it admits a thin eigendecomposition

$$M_2 = U \Lambda U^T \tag{23}$$

where $U = (u_1, \ldots, u_k) \in R^{d \times k}$, and $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$ is a diagonal matrix with $\lambda_i > 0$ for $i = 1, \ldots, k$. Now let's define whitening matrix

$$B = U \Lambda^{-1/2} \tag{24}$$

and the following whitened vectors

$$\widehat{\mu_i} = \sqrt{w_i} B^T \mu_i \tag{25}$$

Also note that because $\mu_i \in span(U)$, and $w_i > 0$, we recover μ_i by

$$\mu_i = \frac{1}{\sqrt{w_i}} (B^T)^\dagger \widehat{\mu_i} \tag{26}$$

where $B^T)^\dagger$ is the Moore-Penrose pseudoinverse on the right of B^T such that $B^T(B^T)^\dagger = I_k$. From the definition of (24), we have

$$
\begin{aligned}
I_k &= B^T M_2 B = M_2(B, B) &(27)\\
&= \sum_{i=1}^{k} w_i (B^T \mu_i) \otimes (B^T \mu_i) &(28)\\
&= \sum_{i=1}^{k} (\sqrt{w_i} B^T \mu_i) \otimes (\sqrt{w_i} B^T \mu_i) &(29)\\
&= \sum_{i=1}^{k} \widehat{\mu_i} \otimes \widehat{\mu_i} &(30)
\end{aligned}
$$

Thus, the vectors $\widehat{\mu_i}, i = 1, \ldots, k$ are orthogonal. Now we apply the whitening to M_3 as follows:

$$
\begin{aligned}
M_3(B, B, B) &= \sum_{i=1}^{k} w_i (B^T \mu_i) \otimes (B^T \mu_i) &(31)\\
&= \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \widehat{\mu_i} \otimes \widehat{\mu_i} \otimes \widehat{\mu_i} &(32)
\end{aligned}
$$

Thus, the whitened tensor $M_3(B, B, B)$ admits an orthogonal decomposition (32), which is the merit we need in order to apply the tensor power method for identifying $\widehat{\mu_i}$, for $i = 1, \ldots, k$. We denote $M_3(B, B, B) = \widehat{M_3}$. The tensor power method is given by the following iteration

$$
\theta_{t+1} \leftarrow \frac{\widehat{M_3}(:, \theta_t, \theta_t)}{\|\widehat{M_3}(:, \theta_t, \theta_t)\|} \tag{33}
$$

starting from an initial random vector θ_0 on the unit sphere $\mathcal{S}^k$. It can be proved that θ_t will converge to a certain eigenvector $\widehat{\mu_i}$ of $\widehat{M_3}$. Once we have an estimate of $\widehat{\mu_i}$, we can identify w_i by

$$
\frac{1}{\sqrt{w_i}} = \widehat{M_3}(\widehat{\mu_i}, \widehat{\mu_i}, \widehat{\mu_i}) \tag{34}
$$

Now we want to find another $\widehat{\mu_j}$ different from $\widehat{\mu_i}$. For this purpose, we need to *deflate* $\widehat{\mu_i}$ from $\widehat{M_3}$. Let $\mathcal{I}$ be the index set such that $i \in \mathcal{I}$ if and only if $(\widehat{\mu_i}, w_i)$ have been identified. The deflation is defined by

$$
\widehat{M_3} \leftarrow \widehat{M_3} - \sum_{i \in \mathcal{I}} \frac{1}{\sqrt{w_i}} \widehat{\mu_i} \otimes \widehat{\mu_i} \otimes \widehat{\mu_i} \tag{35}
$$

Please see [1] for the details of orthogonal tensor decomposition and the tensor power method.

3.3 Recipe

In summary, our recipe using the method of moments are as follows:

1. Compute the empirical moments explicitly $\widetilde{M_1} = \widetilde{\mathbb{E}}[X]$ and $\widetilde{\mathbb{E}}[X \otimes X]$.
2. Identify σ^2 by extracting the smallest eigenvalue of $\widetilde{\mathbb{E}}[X \otimes X] - \widetilde{M_1} \otimes \widetilde{M_1}$.
3. Form M_2 explicitly by (19), and do the thin eigendecomposition (23) to extract the whitening matrix B in (24).
4. Start with $\mathcal{I} = \emptyset$. For $i = 1, \ldots, k$, do the tensor power iteration (33) using the deflated version (35) until converge (or maximum number of iterations met). We can estimate w_i by (34). Let $\mathcal{I} = \mathcal{I} \cup i$.

   Note because in the tensor power iteration, only the action $\widehat{M_3}(:, \theta_t, \theta_t)$ is needed, we don't need to explicitly form $\widehat{M_3}$. Instead, from (21), we have

$$\begin{aligned}\widehat{M_3}(:, \theta_t, \theta_t) = \mathbb{E}[B^T X(\theta_t^T B^T X)^2] - \\ \sigma^2 \sum_{j=1}^{d} \left(B^T M_1(\theta_t^T B^T e_j)^2 + 2B^T e_j(\theta_t^T B^T e_j)(\theta_t^T B^T M_1)\right)\end{aligned} \tag{36}$$

5. Recover μ_i from $\widehat{\mu_i}$ by (26).

4 More general Gaussians

4.1 Differing σ_i^2 for $i = 1, \ldots, k$

The method presented above can be straightforwardly extended to the case where each mixture component has as different variance parameter σ_i^2, with some tweaks to the form of the observed moments M_2 and M_3. Please see [2] for the details.

4.2 General covariance matrices Σ_i

Intuitively, when we have general covariance matrices Σ_i for $i = 1, \ldots, k$, we have many more parameters to estimate (each Σ_i have $\frac{d(d+1)}{2}$ entries) than the case of spherical Gaussians. It turns out we need the 4th and 6th order moments in order to approximately recover Σ_i (with the assumption that $d = O(k^2)$). The algorithm is much more complicated and non-trivial to implement, please see [3] for the details.

References

[1] A. Anandkumar and R. Ge and D. Hsu and S. M. Kakade and M. Telgarsky, Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, Vol. 15, Issue 1, pp. 2773-2832, January 2014.

[2] D. Hsu and S. M. Kakade, Learning mixtures of spherical Gaussians: moment methods and spectral decompositions, *Proceedings of the fourth Innovations in Theoretical Computer Science*, pp. 11-20, January, 2013.

[3] R. Ge and Q. Huang and S. M. Kakade, Learning mixtures of Gaussians in high dimensions, *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 761-770, June, 2015.