۱)

برای محاسبه S_B روابط زیر را داریم:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \quad (۱)$$

پس ابتدا میانگین داده‌ها را بدست می‌آوریم:

$$\mu_1 = \frac{1}{5}\begin{bmatrix} 4+2+2+3+4 \\ 1+4+3+6+4 \end{bmatrix} , \mu_2 = \frac{1}{5}\begin{bmatrix} 9+6+9+8+10 \\ 10+8+5+7+8 \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 3 \\ 3.6 \end{bmatrix}, \mu_2 = \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}$$

$$\mu_1 - \mu_2 = \begin{bmatrix} -5.4 \\ -4 \end{bmatrix}$$

$$\longrightarrow S_B = \begin{bmatrix} -5.4 \\ -4 \end{bmatrix}\begin{bmatrix} -5.4 & -4 \end{bmatrix}$$

$$S_B = \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16 \end{bmatrix}$$

برای محاسبه S_W رابطه زیر را داریم:

$$S_W = S_1^2 + S_2^2 \quad (۲)$$

$$S_1^2 = \sum_{i=1}^{N}(x_i - \mu_1)(x_i - \mu_1)^T = (X1 - \mu_1)(X1 - \mu_1)^T$$

$$x_1 - \mu_1 = \{(1, -2.6), (-1, 0.4), (-1, -0.6), (0, 2.4), (1, 0.4)\}$$

$$S_1^2 = \begin{bmatrix} 1 & -1 & -1 & 0 & 1 \\ -2.6 & 0.4 & -0.6 & 2.4 & 0.4 \end{bmatrix} \begin{bmatrix} 1 & -2.6 \\ -1 & 0.4 \\ -1 & -0.6 \\ 0 & 2.4 \\ 1 & 0.4 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -2 & 13.2 \end{bmatrix}$$

$$x_2 - \mu_2 = \{(0.6, 2.4), (-2.4, 0.4), (0.6, -2.6), (-0.4, -0.6), (1.6, 0.4)\}$$

$$S_2^2 = \begin{bmatrix} 0.6 & -2.4 & 0.6 & -0.4 & 1.6 \\ 2.4 & 0.4 & -2.6 & -0.6 & 0.4 \end{bmatrix} \begin{bmatrix} 0.6 & 2.4 \\ -2.4 & 0.4 \\ 0.6 & -2.6 \\ -0.4 & -0.6 \\ 1.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{bmatrix}$$

$$S_W = \begin{bmatrix} 13.2 & -2.2 \\ -2.2 & 26.4 \end{bmatrix}$$

برای محاسبه مقادیر ویژه ماتریس زیر را باید ایجاد کنیم:

$$A = S_W^{-1} S_B$$

$$S_w^{-1} = \frac{1}{|S_w|} \begin{bmatrix} 26.4 & 2.2 \\ 2.2 & 13.2 \end{bmatrix} \; ; \; |S_w| = 343.64$$

$$S_W^{-1} = \begin{bmatrix} 0.0768 & 0.0064 \\ 0.0064 & 0.384 \end{bmatrix}$$

$$A = \begin{bmatrix} 0.0768 & 0.0064 \\ 0.0064 & 0.384 \end{bmatrix} \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16 \end{bmatrix} = \begin{bmatrix} 2.378 & 1.762 \\ 1.016 & 0.753 \end{bmatrix}$$

$$|\lambda I - A| = 0$$

$$\begin{vmatrix} \lambda - 2.378 & -1.762 \\ -1.016 & \lambda - 0.753 \end{vmatrix} = 0 \longrightarrow \lambda^2 - 3.131\lambda + 1.79 - 1.79 = 0$$

$$\longrightarrow \lambda_{1,2} = 0, \boxed{3.131}$$

Max λ_i

۲)

الف) **Model Selection**

فرآیندی که در طی آن از بین مدل‌های مختلف، بهترین مدل انتخاب می‌شود. هدف از انتخاب مدل، یافتن بهترین مدل است به طوریکه بهترین عملکرد را روی داده‌های تست داشته باشد.

Model Assessment

به فرآیند ارزیابی عملکرد مدل منتخب بر روی داده‌هایی که در فرآیند آموزش نبودند اشاره دارد. هدف از این فرآیند تعیین کیفیت و دقت مدل در پیش‌بینی داده‌های واقعی است.

تفاوت بین این دو رویکرد این است که Model Selection مربوط به یافتن بهترین مدل است درحالیکه Model Assessment مربوط به ارزیابی و تایید عملکرد مدل بر روی داده‌های تست است.

دلایل بکارگیری Model Selection

- ❖ یافتن بهترین مدل
- ❖ تنظیم هایپر پارامترها
- ❖ پیشگیری از overfitting

ب)

Probabilistic Method

روش‌های احتمالی برای انتخاب مدل از معیارهای آماری استفاده می‌کنند تا به طور مستقیم مدل‌ها را بر اساس احتمال یا اطلاعات آماری ارزیابی کنند. چند نمونه از این روش‌ها عبارتند از:

- ❖ AIC (Akaike Information Criterion)
- ❖ BIC (Bayesian Information Criterion)
- ❖ Bayes Factor

مزایای روش‌های احتمالی:

- ❖ کارایی محاسباتی بالا: معمولاً نیاز به محاسبات کمتری نسبت به روش‌های Resampling دارند.
- ❖ تفسیر آسان: معیارهای مشخصی مانند AIC و BIC برای انتخاب مدل ارائه می‌دهند.

معایب روش‌های احتمالی:

- ❖ محدودیت‌ها در فرضیات: معمولاً فرضیات خاصی مانند توزیع نرمال یا استقلال داده‌ها را نیاز دارند.
- ❖ پیش‌فرض‌های قوی: ممکن است در صورتی که داده‌ها یا مدل‌ها با این پیش‌فرض‌ها هماهنگ نباشند، نتایج دقیقی ارائه ندهند.

Resampling Method

روش‌های بازنمونه‌گیری بر اساس تقسیم داده‌ها به چندین مجموعه آموزش و اعتبارسنجی برای ارزیابی مدل‌ها عمل می‌کنند. برخی از روش‌های مشهور عبارتند از:

- ❖ Cross-Validation
- ❖ Bootstrap

مزایای روش‌های بازنمونه‌گیری:

- ❖ انعطاف‌پذیری بالا: می‌تواند برای هر نوع داده و مدل استفاده شود، بدون نیاز به پیش‌فرض‌های قوی.
- ❖ ارزیابی دقیق: با استفاده از بخش‌های مختلف داده‌ها، ارزیابی دقیق‌تری از عملکرد مدل‌ها ارائه می‌دهند.

معایب روش‌های بازنمونه‌گیری:

- ❖ هزینه محاسباتی بالا: به دلیل نیاز به آموزش و ارزیابی مدل‌ها بر روی چندین زیرمجموعه داده‌ها، محاسبات بیشتری نیاز دارند.
- ❖ پیچیدگی بیشتر: ممکن است نیاز به تنظیمات و پیکربندی بیشتری داشته باشند.

در مجموع روش‌های احتمالی سریع‌تر و ساده‌تر هستند اما ممکن است دقت کمتری داشته باشند، درحالیکه روش‌های بازنمونه‌گیری دقت بالاتر و هزینه محاسباتی بیشتری دارند.

ج) زمانی که تعداد داده‌ها کم باشد، انتخاب و ارزیابی مدل با چالش‌های خاصی مواجه می‌شود. این چالش‌ها عبارت‌اند از:

- ❖ Overfitting: با تعداد داده‌های کم، مدل ممکن است به خوبی بر روی داده‌های آموزشی عملکرد نشان دهد اما نتواند به خوبی روی داده‌های جدید تعمیم یابد.
- ❖ ارزیابی ناقص: کمبود داده‌ها می‌تواند منجر به عدم دقت کافی در ارزیابی مدل‌ها شود، زیرا نمونه‌های کمتری برای تست وجود دارند.
- ❖ تنوع کم در داده‌ها: ممکن است داده‌های محدود نتوانند تمام الگوهای موجود در مسئله را به خوبی پوشش دهند.
- ❖ کار نکردن برخی مدل‌ها: مدل‌هایی مثل Regression در صورت بیشتر بودن بعد داده‌ها از تعداد آن کار نمی‌کند.

راهکارهای انتخاب مدل با تعداد کم داده‌ها:

- ❖ استفاده از Cross-Validation با k بالا (Leave-One-Out Cross-Validation) [۱]
- ❖ استفاده از Bootstrap [۲]
- ❖ انتخاب مدل‌های ساده‌تر مثل رگرسیون خطی بجای مدل‌های پیچیده
- ❖ استفاده از روش‌های Regularization [۳]
- ❖ استفاده از روش‌های مبتنی بر پیش‌بینی بیزین [۴]

[۱] Leave-One-Out Cross-Validation (LOOCV): در این روش، هر داده به نوبت به عنوان مجموعه تست و بقیه به عنوان مجموعه آموزش استفاده می‌شوند. این روش تمام داده‌های موجود را در ارزیابی مدل درگیر می‌کند و از داده‌ها به صورت بهینه‌تری استفاده می‌شود.

[۲] Bootstrap با نمونه‌گیری مکرر با جایگزینی از داده‌های موجود، می‌تواند چندین مجموعه آموزشی و تست ایجاد کند. این روش به تخمین بهتر توزیع‌های آماری و ارزیابی مدل‌ها کمک می‌کند.

[۳] روش‌های regularization مانند L1 (Lasso) و L2 (Ridge) به کاهش overfitting کمک می‌کنند و می‌توانند عملکرد مدل‌ها را در شرایط کمبود داده بهبود بخشند.

[۴] روش‌های بیزین می‌توانند از اطلاعات پیشین استفاده کنند و با ترکیب این اطلاعات با داده‌های موجود، به تخمین‌های بهتری دست یابند. این روش‌ها به ویژه در شرایط کمبود داده مفید هستند.

$$\log P(D|\theta) = \log P(x_1, z_1, \dots, x_n, z_n | \alpha, \lambda_1, \lambda_2) \overset{i.i.d}{=\!=} \sum_{i=1}^{N} \log\left(P(x_i, z_i | \alpha, \lambda_1, \lambda_2)\right) \quad ①$$

$$P(x_i, z_i | \alpha, \lambda_1, \lambda_2) = \left(\alpha \lambda_1 e^{-\lambda_1 x_i}\right)^{z_i} \left((1-\alpha)\lambda_2 e^{-\lambda_2 x_i}\right)^{1-z_i}$$

$$\xrightarrow{①} = \sum_{i=1}^{N} z_i \left(\log\alpha + \log\lambda_1 - \lambda_1 x_i\right) + (1-z_i)\left(\log(1-\alpha) + \log\lambda_2 - \lambda_2 x_i\right) \longrightarrow \text{log - complete likelihood}$$

E - Step: $$E_z[L(\theta)] = \sum_{i=1}^{N} E[z_i]\left(\log\alpha + \log\lambda_1 - \lambda_1 x_i\right) + \left(1 - E[z_i]\right)\left(\log(1-\alpha) + \log\lambda_2 - \lambda_2 x_i\right)$$

$$E[z_i] = E[z_i | x_i] = P(z_i = 1 | x) \overset{\text{Bayse Rule}}{=\!=} \frac{P(x_i | z_i = 1) P(z_i = 1)}{P(x_i)} = \frac{\lambda_1^t e^{-\lambda_1^t x_i} \alpha^t}{\alpha^t \lambda_1^t e^{-\lambda_1^t x_i} + (1-\alpha^t)\lambda_2^t e^{-\lambda_2^t x_i}} = \gamma_i^t$$

$\alpha, \lambda_1, \lambda_2$

$$\longrightarrow Q(\theta) = E_z[L(\theta)] = \sum_{i=1}^{N} \gamma_i^t \left(\log\alpha + \log\lambda_1 - \lambda_1 x_i\right) + \left(1 - \gamma_i^t\right)\left(\log(1-\alpha) + \log\lambda_2 - \lambda_2 x_i\right)$$

M - Step: $$\frac{\partial Q}{\partial \alpha} = 0 \longrightarrow \frac{\sum_{i=1}^{n} \gamma_i^t}{\alpha} - \frac{\sum_{i=1}^{n} (1-\gamma_i^t)}{1-\alpha} = 0 \longrightarrow n\alpha - \alpha \cancel{\sum_{i=1}^{n} \gamma_i^t} = \sum_{i=1}^{n} \gamma_i^t - \alpha \cancel{\sum_{i=1}^{n} \gamma_i^t}$$

$$\longrightarrow \boxed{\alpha = \frac{\sum_{i=1}^{n} \gamma_i^t}{n}}$$

$$\frac{\partial Q}{\partial \lambda_1} = 0 \longrightarrow \frac{\sum_{i=1}^{n} \gamma_i^t}{\lambda_1} - \sum_{i=1}^{n} \gamma_i^t x_i = 0 \longrightarrow \boxed{\lambda_1 = \frac{\sum_{i=1}^{n} \gamma_i^t}{\sum_{i=1}^{n} \gamma_i^t x_i}}$$

$$\frac{\partial Q}{\partial \lambda_2} = 0 \longrightarrow \frac{\sum_{i=1}^{n} (1-\gamma_i^t)}{\lambda_2} - \sum_{i=1}^{n} (1-\gamma_i^t) x_i = 0 \longrightarrow \boxed{\lambda_2 = \frac{n - \sum_{i=1}^{n} \gamma_i^t}{\sum_{i=1}^{n} (1-\gamma_i^t) x_i}}$$

(٤

الف) برای تخمین پارامترهای توزیع مخلوط گوسی (GMM) با استفاده از یک شبکه عصبی چندلایه، باید خروجی شبکه به گونه‌ای طراحی شود که بتواند تمامی پارامترهای لازم برای تعریف یک GMM را ارائه دهد. توزیع مخلوط گوسی شامل چندین مولفه گوسی است و هر مولفه دارای پارامترهای مخصوص به خود است. در نتیجه خروجی شبکه شامل پارامترهای GMM خواهد بود.

ب) در این بخش پارامترهای یک GMM را بررسی کرده و با توجه به محدودیت هرکدام تابع فعالساز خروجی شبکه را پیشنهاد می‌کنیم.

- ❖ وزن‌های مخلوط (α_k): وزن‌های مخلوط باید مثبت باشند و مجموع آنها برابر با یک است. منظور از وزن مخلوط ضریب مشارکت هر توزیع گوسی در توزیع مخلوط است.
  برای این پارامتر بهتر است از تابع فعالساز softmax استفاده شود. چرا که این تابع اطمینان می‌دهد که مجموع خروجی‌ها برابر ۱ است و تمامی مقادیر بین ۰ و ۱ قرار دارند.
- ❖ میانگین‌های گوسی (μ_k): چون محدودیتی روی میانگین‌ها وجود ندارد می‌توان از تابع فعالساز خطی استفاده کرد یا می‌توان بدون تابع فعالسازی آنرا در نظر گرفت.
- ❖ کوواریانس‌های گوسی (Σ_k): ماتریس کوواریانس باید مثبت معین باشد.
  می‌توان از دو تابع فعالساز softplus یا exponential استفاده کرد چرا که خروجی این توابع همواره مثبت است.

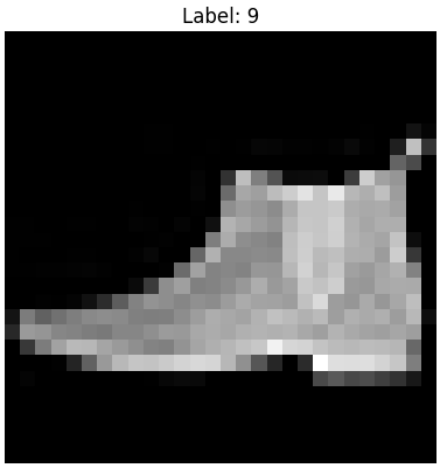
ج) برای یک شبکه عصبی که پارامترهای GMM را تخمین می‌زند، تابع هزینه باید به گونه‌ای تعریف شود که میزان خطا در تخمین پارامترها را به درستی اندازه‌گیری کند و بتواند مدل را به خوبی آموزش دهد. تابع هزینه مناسب برای این کار می‌تواند منفی لگاریتم درستنمایی (Negative Log-Likelihood یا NLL) باشد. این تابع هزینه اندازه‌گیری می‌کند که چقدر مدل ما به خوبی داده‌های مشاهده شده را توضیح می‌دهد.

$$p(x_i) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i \,|\, \mu_k,\, \Sigma_k)$$

$$NLL := -\sum_{i=1}^{N} \log p(x_i)$$

۵)

الف) در ابتدا بوسیله کتابخانه tensorflow داده‌ها را می‌خوانیم. سپس یکی از تصاویر را به شکل تصادفی رسم می‌کنیم.



شکل ۱: (تصویر برای نمایش بهتر سیاه و سفید شده) یکی از تصاویر مجموعه داده

ب) برای استانداردسازی میانگین و واریانس داده‌ها را بدست می‌آوریم و با رابطه زیر استانداردسازی را انجام می‌دهیم:

$$x_i^* = \frac{x_i - \mu}{\sigma}$$

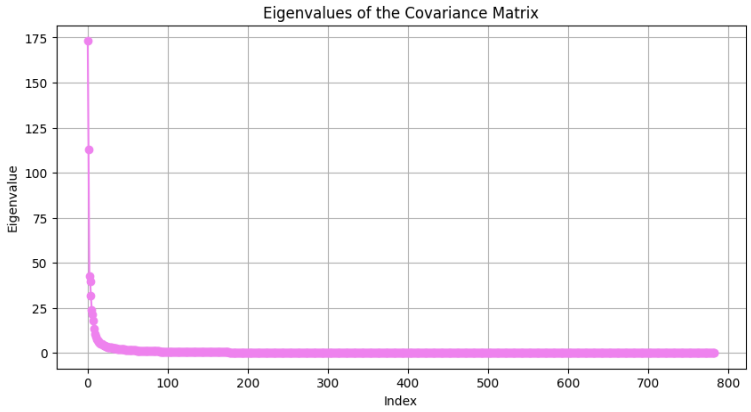
قبل از استانداردسازی لازم است تصاویر را به فرمت float32 ذخیره کنیم.

ج) برای محاسبه ماتریس کوواریانس لازم است داده‌های بدست آمده در بخش قبل که به شکل ماتریسی هستند، flat کنیم. اینکار باعث کاهش پیچیدگی محاسباتی می‌شود.
با توجه به اینکه ابعاد هر عکس ۲۸ × ۲۸ است، با flat کردن، ابعاد بردار بدست آمده ۷۸٤ × ۱ خواهد بود. در نتیجه انتظار داریم ابعاد ماتریس کوواریانس ۷۸٤ × ۷۸٤ شود.

```
Covariance matrix shape: (784, 784)
```

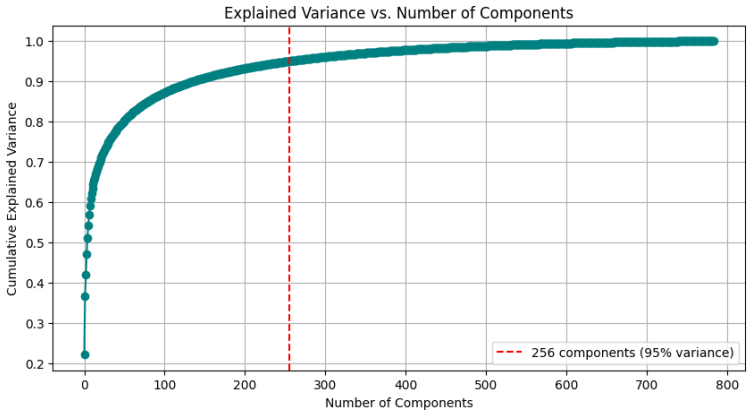
شکل ۲: ابعاد ماتریس کوواریانس

د) با محاسبه و مرتب‌سازی مقادیر ویژه و بردارهای ویژه داریم:



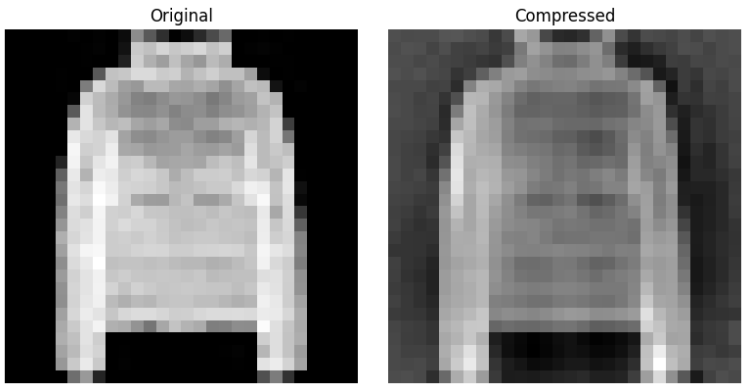
شکل ۳: نمودار مقادیر ویژه از بیشترین به کمترین

همانطور که مشاهده می‌کنیم، بیشترین مقدار ویژه در حدود ۱۷۵ بدست می‌آید.
برای پیداکردن تعداد کامپوننت مناسب در فرآیند فشرده‌سازی می‌توان از معیار Cumulative variance استفاده کرد. همانطور که می‌دانیم مقدار ویژه نمایانگر پراکندگی داده‌ها در جهت بردار ویژه متناظر با آن است. همچنین PCA به دنبال بیشینه کردن پراکندگی داده‌هاست و سعی می‌کند این پراکندگی را نزدیک به پراکندگی اصلی داده‌ها نگه دارد. این معیار بررسی می‌کند با انتخاب چند کامپوننت (بردار ویژه) با توجه به بیشترین مقادیر ویژه، واریانس داده‌ها نزدیک به واریانس اصلی می‌شود.
برای مثال این معیار می‌تواند برابر 95% واریانس اصلی داده‌ها باشد. در نتیجه در ابتدا با انتخاب بردار ویژه متناظر با بیشترین مقدار ویژه و در ادامه اضافه کردن ماکسیمم مقادیر ویژه از بین مقادیر ویژه انتخاب نشده، واریانس در هر مرحله محاسبه می‌شود و نسبت آن با واریانس اصلی داده‌ها، می‌توان تعداد کامپوننت‌های مناسب برای فشرده‌سازی را بدست آورد.
برای 95% thrshold تعداد کامپونت‌های مناسب ۲۵۶ بدست می‌آید.



شکل ٤: نمودار معیار Cumulative variance بر اساس تعداد بردارهای ویژه

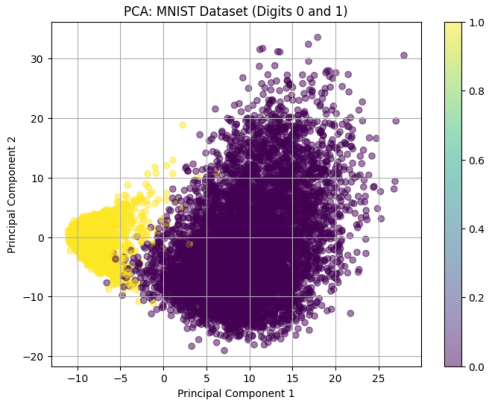
ه) برای فشرده‌سازی، ۲۵۶ بردار ویژه متناظر با ۲۵۶ تا از بیشترین مقدار ویژه انتخاب کرده و تصاویر flat شده را روی این مجموعه project می‌کنیم.
پس از فرآیند فشرده‌سازی برای بازگردانی تصاویر و visualize کردن داده‌ها ماتریس بدست آمده در transpose بردارهای ویژه ضرب می‌شود تا تصاویر به فضای اصلی (ابعاد ۲۸ × ۲۸) بازگردند.
با اعمال فشرده‌سازی مقایسه بین یک تصویر در حالت فشرده و در حالت معمولی به شکل زیر خواهد بود:



شکل ۵: مقایسه یک تصویر از مجموعه داده در حالت فشرده و غیرفشرده

۶)

الف) پس از خواندن دیتاست، کلاس ۰ و ۱ را جدا کرده و سپس آنها را ابتدا flat و بعد استاندارد سازی می‌کنیم. با اعمال PCA با دو Component بعد داده‌ها را از ۷۸۴ به ۲ کاهش می‌دهیم. سپس تابع مخلوط گوسی با دو جز را روی داده‌ها برازش می‌کنیم.

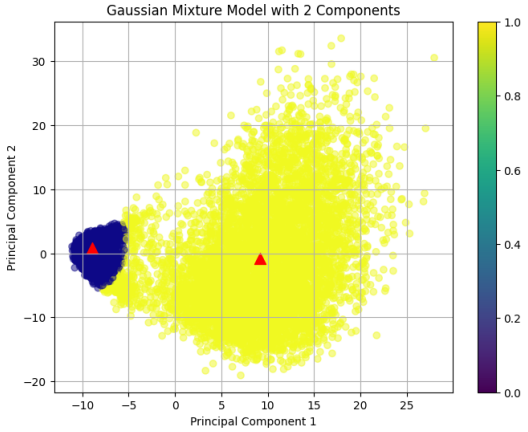


شکل ۶: داده‌ها پس از اعمال PCA

ب) نتیجه به شکل زیر خواهد بود:

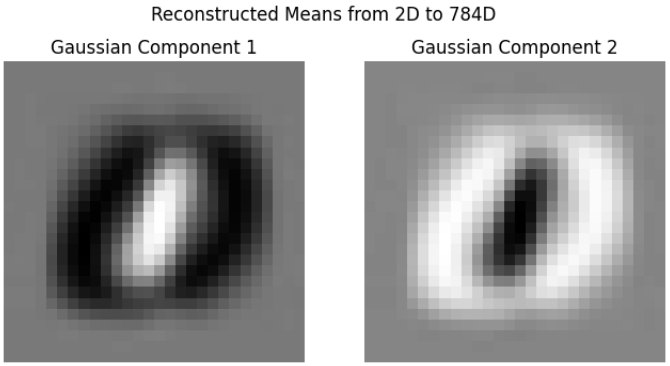
```
Euclidean distance between the means: 18.21587791824532
```

شکل ۷: فاصله اقلیدسی میانگین جز گوسی‌های تابع مخلوط گوسی



شکل ۸: داده‌ها پس از برازش GMM و میانگین هر کلاس

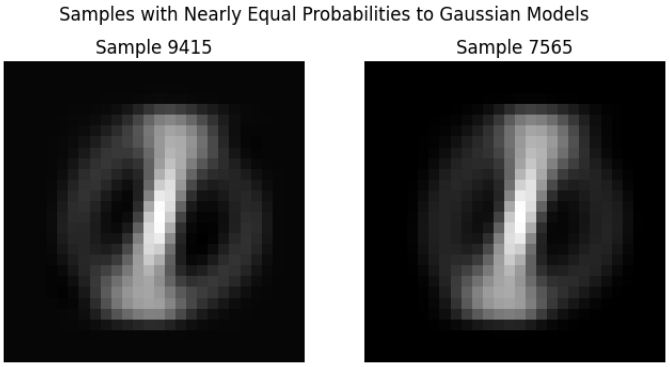
پ) با اعمال عکس PCA و تغییر ابعاد به منظور نمایش تصویر نتایج زیر را خواهیم داشت:



شکل ۹: تصویر میانگین کلاس‌ها از توزیع مخلوط گوسی

با توجه به شکل فوق، توزیع‌ها توانسته‌اند کلاس‌ها را مشخص کنند و تا حد خوبی از یکدیگر متمایز کنند. میانگین گوسی اول که در شکل ۸ روی کلاس ۰ برازش شده بود صفر است و میانگین گوسی دوم نیز یک است.

ت) ابتدا اختلاف احتمال تعلق به هر جز گوسی را برای تمام داده‌ها بدست می‌آوریم، سپس لیست بدست آمده را مرتب می‌کنیم و کوچکترین دو عضو آن را به عنوان نمونه‌هایی که احتمال تعلقشان به جزهای گوسی نزدیک بهم است، در نظر می‌گیریم.



شکل ۱۰: تصویر داده‌هایی که احتمال تعلقشان به جزهای گوسی نزدیک بهم است

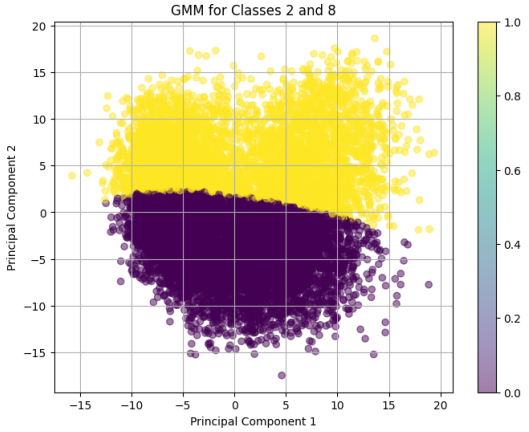
مشاهده می‌کنیم این داده‌ها در فرآیند فشرده‌سازی بگونه‌ای تغییر کرده‌اند که کلاس آنها قابل تشخیص نیست، به همین خاطر احتمال تعلق آنها به هر جز گوسی بسیار نزدیک بهم می‌باشد.

ث) فاصله هندسی میانگین کلاس‌های غیرهمسان را دو به دو محاسبه می‌کنیم. پس از محاسبه نتیجه به شکل زیر خواهد شد:

```
Pair with greatest difference between means: (0, 1) with distance 18.22
Pair with least difference between means: (2, 8) with distance 8.68
```

شکل ۱۱: بیشترین و کمترین فاصله هندسی میانگین کلاس‌ها

بیشترین فاصله بین کلاس ۰ و ۱ است. چرا که این دو عدد از نظر هندسی با هم تفاوت زیادی دارند در نتیجه فاصله توزیع‌های گوسی آنها از یکدیگر نیز زیاد است.
در مقابل کمترین فاصله مربوط به کلاس ۲ و ۸ است. این دو کلاس از نظر هندسی شبیه به یکدیگر می‌باشند، در نتیجه داده‌های این دو کلاس نزدیک بهم خواهد بود و در نتیجه توزیع‌های گوسی برازش شده نیز به یکدیگر نزدیک خواهند بود.



شکل ۱۲: پراکندگی داده‌های کلاس ۲ و ۸

مشاهده می‌کنیم داده‌ها بهم نزدیک هستند و با اعمال توزیع مخلوط گوسی میانگین این توزیع‌ها نزدیک بهم است.

۷)

الف) در مسئله خوشه‌بندی یک ابهام ذاتی وجود دارد که تعداد خوشه‌هاست. روش‌هایی برای حل این چالش وجود دارد:

۱. K-means و تحلیل ELBOW روی آن: K-means یکی از الگوریتم‌های خوشه‌بندی است. برای تعیین تعداد خوشه‌ها (K) از معیار ELBOW استفاده می‌شود. این معیار به این صورت در نظر گرفته می‌شود: ابتدا الگوریتم K-means با مقادیر مختلف K (مثلاً از ۱ تا ۱۰) اجرا می‌شود. سپس برای هر مقدار K ، مجموع مربعات فاصله‌های داده‌ها از مراکز خوشه‌ها (within-cluster sum of squares یا WCSS) محاسبه می‌شود. WCSS بر روی محور y و K بر روی محور x ترسیم می‌شود. نمودار حاصل به شکل یک آرنج (elbow) خواهد بود. نقطه‌ای که پس از آن کاهش WCSS کمتر می‌شود، تعداد بهینه خوشه‌ها را نشان می‌دهد.

۲. Silhouette Score: معیاری برای ارزیابی کیفیت خوشه‌بندی است که هر نقطه را در نظر می‌گیرد و میانگین فاصله از نقاط در خوشه خود (a) و کمینه میانگین فاصله از نقاط خوشه‌های دیگر (b) را محاسبه می‌کند. رابطه Silhouette Score بصورت زیر می‌باشد:

$$S = \frac{a - b}{\max(a, b)}$$

که S عددی بین ۱ و −۱ است. مقدار نزدیک به ۱ نشان‌دهنده خوشه‌بندی خوب، مقدار نزدیک به ۰ نشان‌دهنده نقاط در مرز خوشه‌ها و مقدار منفی نشان‌دهنده خوشه‌بندی ضعیف است.
برای تعیین تعداد بهینه خوشه‌ها، الگوریتم خوشه‌بندی با مقادیر مختلف K اجرا و Silhouette Score میانگین تمام نقاط محاسبه می‌شود. مقدار K با بالاترین Silhouette Score بهینه در نظر گرفته می‌شود.

۳. Davis-Bouldin Index: میانگین نسبت‌های فشردگی (داخل خوشه) به جدایی (بین خوشه) برای تمام خوشه‌ها است. رابطه آن به صورت زیر است:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} \left(\frac{d_i + d_j}{d_{ij}} \right)$$

K: تعداد خوشه‌ها
d_i: میانگین فاصله نقاط درون خوشه i از مرکز خوشه
d_{ij}: فاصله بین مراکز خوشه‌های i و j
مقدار پایین‌تر DBI نشان‌دهنده خوشه‌بندی بهتر است. برای تعیین تعداد بهینه خوشه‌ها، الگوریتم خوشه‌بندی با مقادیر مختلف K اجرا و DBI محاسبه می‌شود. کمترین مقدار DBI نشان‌دهنده تعداد بهینه خوشه‌ها است.

٤. Calinski-Harabasz Index: CHI نسبت واریانس بین‌خوشه‌ای به واریانس درون‌خوشه‌ای است. رابطه آن به صورت زیر است:

$$CHI = \frac{between - cluster\ dispersion\ sum}{within - cluster\ dispersion\ sum} \cdot \frac{N - K}{K - 1}$$

N: تعداد داده‌ها
K: تعداد خوشه‌ها
$between - cluster\ dispersion\ sum$: مجموع واریانس نقاط داده از مراکز خوشه‌ها
$within - cluster\ dispersion\ sum$: مجموع واریانس نقاط داده درون هر خوشه
مقدار بالاتر CHI نشان‌دهنده خوشه‌بندی بهتر است. برای تعیین تعداد بهینه خوشه‌ها، الگوریتم خوشه‌بندی با مقادیر مختلف K اجرا و CHI محاسبه می‌شود. بیشترین مقدار CHI نشان‌دهنده تعداد بهینه خوشه‌ها است.

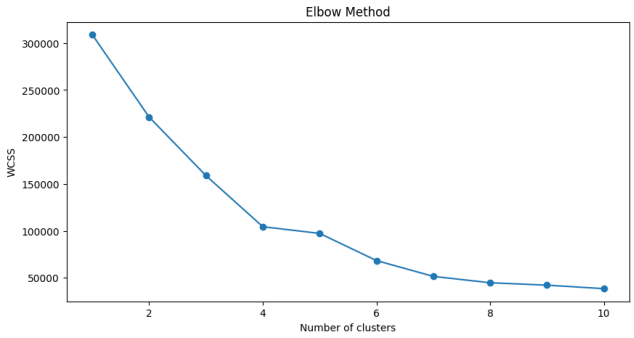
۵. Dunn Index: معیاری برای تعیین کیفیت خوشه‌بندی است که نسبت کمینه فاصله بین خوشه‌ها به بیشینه قطر خوشه‌ها را محاسبه می‌کند. فرمول آن به صورت زیر است:

$$D = \frac{\min_{1 \le i \le j \le K} d(c_i, c_j)}{\max_{1 \le k \le K} \delta_k}$$

$d(c_i, c_j)$: فاصله بین مراکز خوشه‌های i و j
δ_k: بیشینه فاصله بین نقاط درون خوشه k
مقدار بالاتر Dunn Index نشان‌دهنده خوشه‌بندی بهتر است. برای تعیین تعداد بهینه خوشه‌ها، الگوریتم خوشه‌بندی با مقادیر مختلف K اجرا و Dunn Index محاسبه می‌شود. بیشترین مقدار Dunn Index نشان‌دهنده تعداد بهینه خوشه‌ها است.

ب)

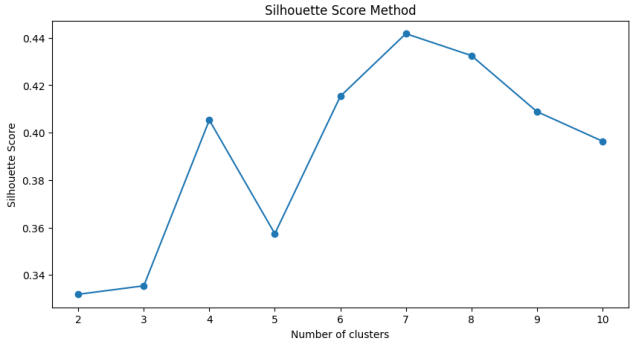
❖ تحلیل ELBOW



شکل ۱۳: نمودار معیار WCSS و تحلیل ELBOW به ازای تعداد خوشه‌های مختلف

با توجه به شکل فوق روند کاهش WCSS از تعداد ۷ خوشه به بعد کمتر می‌شود. در نتیجه می‌توان تعداد خوشه‌های مناسب با توجه به این تحلیل را ۷ در نظر گرفت.

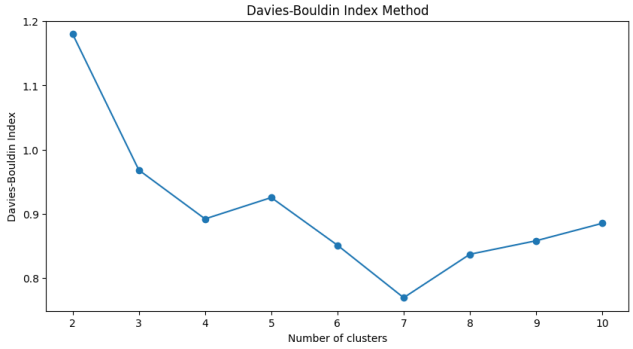
❖ معیار Silhouette Score



شکل ۱٤: نمودار معیار Silhouette Score به ازای تعداد خوشه‌های مختلف

مشاهده می‌کنیم که بیشترین مقدار این معیار به ازای ۷ خوشه اتفاق می‌افتد.

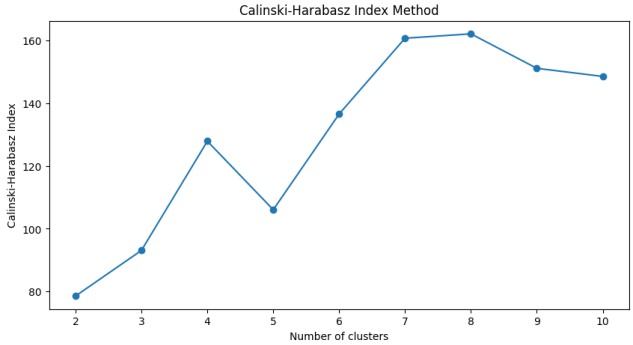
❖ معیار DBI



شکل ۱۵: نمودار معیار DBI به ازای تعداد خوشه‌های مختلف

با توجه به شکل کمترین میزان این معیار در تعداد خوشه ۷ اتفاق می‌افتد که با توجه به این معیار تعداد خوشه بهینه است.

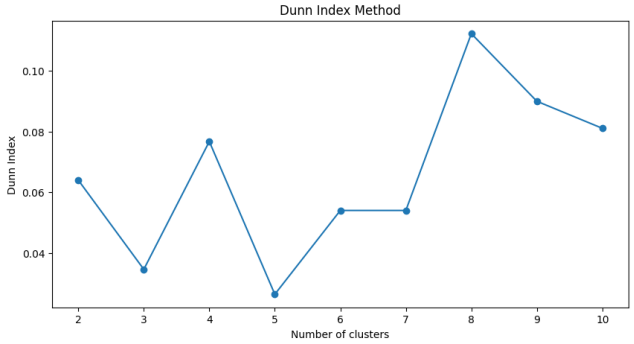
❖ معیار CHI



شکل ۱۶: نمودار معیار CHI به ازای تعداد خوشه‌های مختلف

مشاهده می‌کنیم بیشترین مقدار CHI در تعداد خوشه ۸ اتفاق می‌افتد.

❖ معیار Dunn Index



شکل ۱۷: نمودار معیار Dunn Index به ازای خوشه‌های مختلف

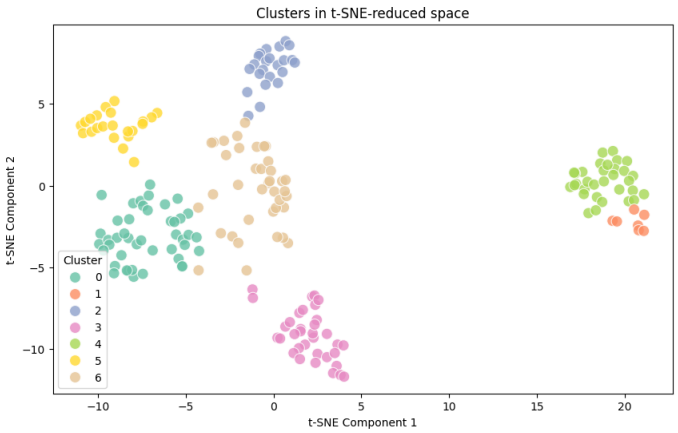
مشاهده می‌کنیم بیشترین مقدار این معیار در تعداد خوشه ۸ اتفاق می‌افتد.
در مجموع می‌توان گفت ۷ خوشه برای خوشه‌بندی این دیتاست مناسب است.

ج) برای نمایش داده‌ها با بعد بالاتر تکنیک‌هایی وجود دارد که در اینجا به برخی از آنها اشاره می‌کنیم:

❖ تکنیک‌های کاهش بعد: تکنیک‌هایی مانند PCA ، t-SNE و UMAP وجود دارند که بعد داده‌ها را به بعد قابل نمایش کاهش می‌دهند.

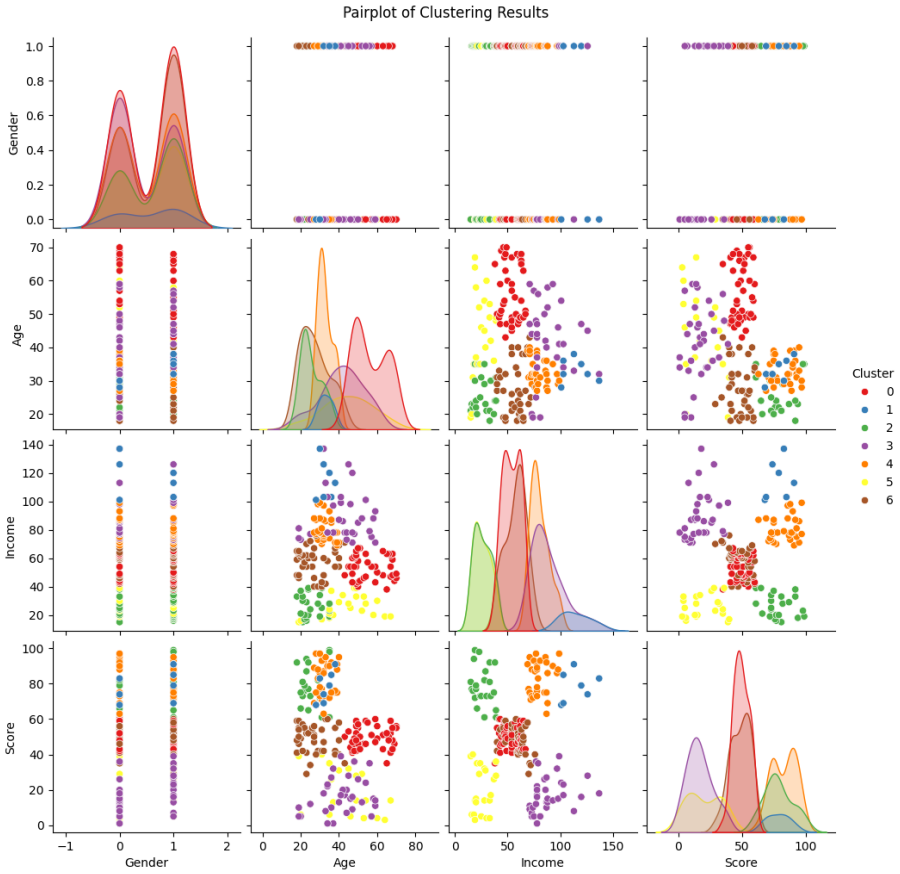
❖ تکنیک‌های نمایش خوشه‌ها: مثل رسم خوشه‌بندی دو به دو به ازای هر ویژگی که چندین نمودار خواهیم داشت. (Pairwise plots)

در این سوال از t-SNE و Pairwise plots استفاده می‌کنیم.



شکل ۱۸: نمایش خوشه‌بندی با استفاده از t-SNE

مشاهده می‌کنیم با ۷ خوشه داده‌ها به خوبی از یکدیگر تفکیک شده‌اند.



شکل ۱۹: نمایش خوشه‌بندی داده‌ها به شکل Pairwise

مشاهده می‌کنیم به ازای هر دو ویژگی یک نمودار داریم که در آن خوشه‌بندی به تعداد ۷ خوشه انجام شده است.

با رسم این نمودار می‌توان ویژگی‌هایی که در تفکیک‌پذیری داده‌ها تاثیر بیشتری دارند را نیز شناسایی کرد، برای مثال ترکیب ویژگی جنسیت با هیچ ویژگی دیگری نتوانسته داده‌ها را از یکدیگر به خوبی تفکیک کند. در مقابل ترکیب دو ویژگی Income و Score تا حد خوبی داده‌ها را از هم تفکیک کرده است.