Derivación de estimadores máximo verosímiles

Percepción - ETSInf

1. Verosimilitud y parámetros del modelo

La función verosimilitud para un conjunto de entrenamiento de N muestras independientes e idénticamente distribuidas (i.i.d.) extraídas aleatoriamente de C clases, $X = \{(x_n, c_n)\}_{n=1}^{N}$ se define como:

$$L(\mathbf{\Theta}; X) = p(X; \mathbf{\Theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n, c_n; \mathbf{\Theta})$$

siendo $\mathbf{\Theta}$ el vector de parámetros del modelo. Sin embargo, por comodidad en la derivación de los estimadores máximo versosímiles, se utiliza la función log-verosimilitud:

$$\mathcal{L}(\mathbf{\Theta}; X) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, c_n; \mathbf{\Theta})$$

En nuestro caso, el vector de parámetros se puede desglosar como sigue:

$$\mathbf{\Theta} = \{p_1, \cdots, p_c, \cdots, p_C, \mathbf{\Theta}_1, \cdots, \mathbf{\Theta}_c, \cdots, \mathbf{\Theta}_C\}$$

siendo p_c la probabilidad a priori de la clase c y $\mathbf{\Theta}_c$ el vector de parámetros que gobierna la distribución de probabilidad condicional de clase:

$$p(\mathbf{x}_n, c_n; \mathbf{\Theta}) = p_{c_n}\, p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c)$$

En nuestro caso, $p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c)$ se podrá instanciar a una distribución Bernoulli, multinomial o Gaussiana.

2. Derivación de la probabilidad a priori

Primero estimaremos el parámetro de la probabilidad a priori de la clase p_c sin necesidad de especificar el modelo que utilizaremos para la probabilidad condicional. La función de log-verosimilitud es:

$$\mathcal{L}(\mathbf{\Theta}; X) = \sum_{n=1}^{N} \log p_{c_n}\, p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) = \sum_{n=1}^{N} \log p_{c_n} + \log p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) \qquad (1)$$

Por lo tanto, el problema de optimización con la restricción de que las probabilidades a priori sumen 1 es

$$\hat{p_c} = \underset{p_c}{\text{argmax}} \sum_{n=1}^{N} \log p_{c_n} + \log p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) \quad \text{s.t.} \quad \sum_{c'} p_{c'} = 1$$

En este caso utilizaremos la técnica de multiplicadores de Lagrange para definir el problema de optimización con restricciones:

$$\begin{aligned}
\hat{p_c} &= \underset{p_c}{\text{argmax}} \max_{\lambda} \Lambda(\mathbf{\Theta}, \lambda) \\
&= \underset{p_c}{\text{argmax}} \max_{\lambda} \sum_{n=1}^{N} \log p_{c_n} + \log p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) - \lambda \left(\sum_{c'} p_{c'} - 1 \right)
\end{aligned}$$

El primer paso es calcular la derivada parcial de $\Lambda(\mathbf{\Theta}, \lambda)$ respecto del parámetro de interés p_c e igualar a cero$\left(\text{Nota: } \frac{d \log x}{dx} = \frac{dx}{x}\right)$:

$$\frac{\partial \Lambda(\mathbf{\Theta}, \lambda)}{\partial p_c} = \sum_{\substack{n=1: \\ c_n = c}}^{N} \frac{1}{p_c} - \lambda = 0$$

Despejamos p_c:

$$p_c = \frac{1}{\lambda} \sum_{\substack{n=1: \\ c_n = c}}^{N} 1 = \frac{1}{\lambda} N_c \tag{2}$$

Observa como el sumatorio suma 1 para cada muestra que sea de la clase c, por tanto, el sumatorio devolverá el número de muestras de la clase c en el conjunto de entrenamiento.

El segundo paso sería realizar la derivada parcial de $\Lambda(\mathbf{\Theta}, \lambda)$ respecto al multiplicador de Lagrange λ:

$$\frac{\partial \Lambda(\mathbf{\Theta}, \lambda)}{\partial \lambda} = - \left(\sum_{c'} p_{c'} - 1 \right) = 0$$

sustituimos el valor de p_c' por el obtenido en la Ec. 2

$$- \left(\sum_{c'} \frac{1}{\lambda} N_{c'} - 1 \right) = 0$$

y despejamos λ

$$\lambda = \sum_{c'} N_{c'} \tag{3}$$

Finalmente, substituimos Ec. 3 en Ec. 2 para obtener el estimador máximo verosímil de la probabilidad a priori es

$$p_c = \frac{N_c}{\sum_{c'} N_{c'}} = \frac{N_c}{N} \tag{4}$$

3. Derivación del parámetro Bernoulli

En este caso, la probabilidad condicional de la Ec. 1 se instancia en una distribución Bernoulli

$$p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) = \prod_{d=1}^{D} p_{c_n d}^{x_{nd}} \; (1 - p_{c_n d})^{(1-x_{nd})} \tag{5}$$

siendo $\mathbf{\Theta}_c$ en este caso el parámetro Bernoulli $\boldsymbol{p}_c = (p_{c1}, \cdots, p_{cd}, \cdots, p_{cD})$. Por tanto, el problema de optimización es

$$\begin{aligned}
\hat{\boldsymbol{p}}_c &= \operatorname*{argmax}_{\boldsymbol{p}_c} \mathcal{L}(\mathbf{\Theta}; X) \\
&= \operatorname*{argmax}_{\boldsymbol{p}_c} \sum_{n=1}^{N} \log p_{c_n} + \log \prod_{d=1}^{D} p_{c_n d}^{x_{nd}} \; (1 - p_{c_n d})^{(1-x_{nd})} \\
&= \operatorname*{argmax}_{\boldsymbol{p}_c} \sum_{n=1}^{N} \log p_{c_n} + \sum_{d=1}^{D} x_{nd} \log p_{c_n d} + (1 - x_{nd}) \log (1 - p_{c_n d})
\end{aligned}$$

Por simplicidad, calculamos la derivada parcial de $\mathcal{L}(\boldsymbol{\Theta}; X)$ respecto del parámetro de p_{cd} e igualamos a cero:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial p_{cd}} = \sum_{\substack{n=1:\\ c_n=c}}^{N} \frac{x_{nd}}{p_{cd}} - \frac{(1-x_{nd})}{(1-p_{cd})} = 0$$

$$\frac{1}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} - \frac{1}{(1-p_{cd})} \sum_{\substack{n=1:\\ c_n=c}}^{N} (1-x_{nd}) = 0$$

$$\frac{1}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} = \frac{1}{(1-p_{cd})} \sum_{\substack{n=1:\\ c_n=c}}^{N} (1-x_{nd})$$

$$\frac{(1-p_{cd})}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} = \sum_{\substack{n=1:\\ c_n=c}}^{N} (1-x_{nd})$$

$$\frac{1}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} - \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} = \sum_{\substack{n=1:\\ c_n=c}}^{N} (1-x_{nd})$$

$$\frac{1}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} = \sum_{\substack{n=1:\\ c_n=c}}^{N} (1-x_{nd}+x_{nd})$$

$$\frac{1}{p_{cd}} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd} = N_c$$

$$p_{cd} = \frac{1}{N_c} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_{nd}$$

Como generalización se podría realizar la derivación de la estimación del parámetro $\boldsymbol{p}_c$

$$\hat{\boldsymbol{p}}_c = \operatorname*{argmax}_{\boldsymbol{p}_c} \sum_{n=1}^{N} \log p_{c_n} + \mathbf{x}_n \log \boldsymbol{p}_{c_n} + (\mathbf{1}-\mathbf{x}_n) \log \left(\mathbf{1}-\boldsymbol{p}_{c_n}\right)$$

y la derivada parcial de $\mathcal{L}(\boldsymbol{\Theta}; X)$, que es un escalar, respecto del vector $\boldsymbol{p}_c$ es

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \boldsymbol{p}_c} = \begin{pmatrix} \frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial p_{c1}} \\ \cdots \\ \frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial p_{cd}} \\ \cdots \\ \frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial p_{cD}} \end{pmatrix} = \mathbf{0} \tag{6}$$

Es decir, cada componente del vector en la Ec. 6 es la derivada que habíamos calculado. Por tanto, la estimación del vector $\boldsymbol{p}_c$ sería

$$\boldsymbol{p}_c = \frac{1}{N_c} \sum_{\substack{n=1: \\ c_n=c}}^{N} \mathbf{x}_n$$

4. Derivación del parámetro multinomial

En este caso, la probabilidad condicional de la Ec. 1 se instancia en una distribución multinomial

$$p(\mathbf{x}_n \mid c_n; \boldsymbol{\Theta}_c) = \binom{\mathbf{x}_{n+}}{\mathbf{x}_n} \prod_{d=1}^{D} p_{c_n d}^{x_{nd}} \tag{7}$$

siendo $\boldsymbol{\Theta}_c$ en este caso el parámetro multinomial $\boldsymbol{p}_c = (p_{c1}, \cdots, p_{cd}, \cdots, p_{cD})$. Por tanto, el problema de optimización es

$$\begin{aligned}
\hat{\boldsymbol{p}}_c &= \operatorname*{argmax}_{\boldsymbol{p}_c} \mathcal{L}(\boldsymbol{\Theta}; X) \\
&= \operatorname*{argmax}_{\boldsymbol{p}_c} \sum_{n=1}^{N} \log p_{c_n} + \log \binom{\mathbf{x}_{n+}}{\mathbf{x}_n} \prod_{d=1}^{D} p_{c_n d}^{x_{nd}} \\
&= \operatorname*{argmax}_{\boldsymbol{p}_c} \sum_{n=1}^{N} \log p_{c_n} + \log \binom{\mathbf{x}_{n+}}{\mathbf{x}_n} + \sum_{d=1}^{D} x_{nd} \log p_{c_n d}^{x_{nd}}
\end{aligned}$$

sujeto a que las probabilidades del prototipo multinomial de cada clase c' sumen 1.

$$\sum_d p_{c'd} = 1 \quad \forall c'$$

Al igual que en la estimación máximo-verosimil de la probabilidad a priori, utilizaremos la técnica de multiplicadores de Lagrange para definir el problema de optimización con restricciones:

$$\begin{aligned}
\hat{p_c} &= \operatorname*{argmax}_{p_c} \max_{\lambda} \Lambda(\boldsymbol{\Theta}, \lambda) \\
&= \operatorname*{argmax}_{p_c} \max_{\lambda} \sum_{n=1}^{N} \log p_{c_n} + \log \binom{\mathbf{x}_{n+}}{\mathbf{x}_n} + \sum_{d=1}^{D} x_{nd} \log p_{c_n d} - \sum_{c'} \lambda_{c'} \left(\sum_{d'=1}^{D} p_{c'd'} - 1 \right)
\end{aligned}$$

Por simplicidad, primero calculamos la derivada parcial de $\mathcal{L}(\boldsymbol{\Theta}; X)$ respecto del parámetro de p_{cd} e igualamos a cero:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial p_{cd}} = \sum_{\substack{n=1: \\ c_n=c}}^{N} \frac{x_{nd}}{p_{cd}} - \lambda_c = 0 \tag{8}$$

Despejamos p_{cd}:

$$p_{cd} = \frac{1}{\lambda_c} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd} \tag{9}$$

El segundo paso sería realizar la derivada parcial de $\Lambda(\mathbf{\Theta}, \lambda)$ respecto al multiplicador de Lagrange λ_c:

$$\frac{\partial \Lambda(\mathbf{\Theta}, \lambda)}{\partial \lambda_c} = -\left(\sum_{d'=1}^{D} p_{cd'} - 1 \right) = 0$$

sustituimos el valor de $p_{cd'}$ por el obtenido en la Ec. 9

$$-\left(\sum_{d'=1}^{D} \frac{1}{\lambda_c} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd'} - 1 \right) = 0$$

y despejamos λ_c

$$\lambda_c = \sum_{d'=1}^{D} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd'} \tag{10}$$

Finalmente, substituimos Ec. 10 en Ec. 9 para obtener el estimador máximo verosímil del parámetro multinomial

$$p_{cd} = \frac{1}{\sum_{d'=1}^{D} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd'}} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd} \tag{11}$$

Como en el caso de la estimación del parámetro multinomial, se podría realizar la derivación de la estimación del parámetro $\boldsymbol{p}_c$

$$\hat{\boldsymbol{p}}_c = \underset{\boldsymbol{p}_c}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{c_n} + \log \binom{\mathbf{x}_{n_+}}{\mathbf{x}_n} + \mathbf{x}_n \log \boldsymbol{p}_c - \sum_{c'} \lambda_{c'} \left(\sum_{d'=1}^{D} p_{c'd'} - 1 \right)$$

y la derivada parcial de $\mathcal{L}(\mathbf{\Theta}; X)$, que es un escalar, respecto del vector $\boldsymbol{p}_c$ es la misma que en Ec. 6. Por tanto, la estimación del vector $\boldsymbol{p}_c$ sería

$$\boldsymbol{p}_c = \frac{1}{\sum_{d'=1}^{D} \sum_{\substack{n=1: \\ c_n=c}}^{N} x_{nd'}} \sum_{\substack{n=1: \\ c_n=c}}^{N} \mathbf{x}_n$$

5. Derivación de los parámetros gaussianos

En este caso, la probabilidad condicional de la Ec. 1 se instancia en una distribución gaussiana, que en el caso unidimensional resulta en

$$p(x_n \mid c_n; \mathbf{\Theta}_c) = \frac{1}{\sigma_{c_n}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_n - \mu_{c_n}}{\sigma_{c_n}}\right)^2\right) =$$

$$(2\pi\sigma_{c_n}^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\frac{x_n - \mu_{c_n}}{\sigma_{c_n}}\right)^2\right)$$

En el caso multidimensional con dimensión D, tendremos la equivalencia correspondiente como

$$p(\mathbf{x}_n \mid c_n; \mathbf{\Theta}_c) = (2\pi)^{-\frac{D}{2}} \left|\Sigma_{c_n}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})\right)$$

5.1. Caso unidimensional

Para el caso unidimensional, tendremos que $\mathbf{\Theta}_c$ son μ_c y σ_c, con lo que se establece el problema de optimización como sigue:

$$\hat{\mathbf{\Theta}}_c = (\hat{\mu}_c, \hat{\sigma}_c) = \operatorname*{argmax}_{\mu_c, \sigma_c} \mathcal{L}(\mathbf{\Theta}; X) =$$

$$\operatorname*{argmax}_{\mu_c, \sigma_c} \sum_{n=1}^{N} \log p_{c_n} + \log\left((2\pi\sigma_{c_n}^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\frac{x_n - \mu_{c_n}}{\sigma_{c_n}}\right)^2\right)\right) =$$

$$\operatorname*{argmax}_{\mu_c, \sigma_c} \sum_{n=1}^{N} \log p_{c_n} - \frac{1}{2}\log 2\pi\sigma_{c_n}^2 - \frac{1}{2}\left(\frac{x_n - \mu_{c_n}}{\sigma_{c_n}}\right)^2 =$$

$$\operatorname*{argmax}_{\mu_c, \sigma_c} \sum_{n=1}^{N} \log p_{c_n} - \frac{1}{2}\log 2\pi\sigma_{c_n}^2 - \frac{1}{2\sigma_{c_n}^2}(x_n - \mu_{c_n})^2$$

A partir de aquí, se plantea su derivada respecto a cada uno de los parámetros para optimizarlo.

Comenzando por μ_c:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \mu_c} = \sum_{\substack{n=1:\\ c_n=c}}^{N} \frac{1}{\sigma_c^2}\left(x_n - \mu_c\right) = \frac{1}{\sigma_c^2} \sum_{\substack{n=1:\\ c_n=c}}^{N} \left(x_n - \mu_c\right)$$

Igualando a cero y despejando:

$$\frac{1}{\sigma_c^2} \sum_{\substack{n=1:\\ c_n=c}}^{N} \left(x_n - \mu_c\right) = 0 \rightarrow \sum_{\substack{n=1:\\ c_n=c}}^{N} \left(x_n - \mu_c\right) = 0 \rightarrow$$

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} x_n - \sum_{\substack{n=1:\\ c_n=c}}^{N} \mu_c = 0 \rightarrow \sum_{\substack{n=1:\\ c_n=c}}^{N} x_n = \sum_{\substack{n=1:\\ c_n=c}}^{N} \mu_c$$

Si llamamos N_c al número de muestras de la clase c, esto queda como

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} x_n = N_c\, \mu_c \rightarrow \mu_c = \frac{1}{N_c} \sum_{\substack{n=1:\\ c_n=c}}^{N} x_n$$

Es decir, la μ_c estimada es la media de los datos de la clase c.

Respecto a σ_c^2, tendremos

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \sigma_c^2} = \sum_{\substack{n=1:\\ c_n=c}}^{N} -\frac{1}{2\sigma_c^2} + \frac{1}{2\left(\sigma_c^2\right)^2}(x_n - \mu_c)^2$$

Igualando a cero y despejando

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} -\frac{1}{2\sigma_c^2} + \frac{1}{2\left(\sigma_c^2\right)^2}(x - \mu_c)^2 = 0 \rightarrow \frac{1}{2\sigma_c^2} \sum_{\substack{n=1:\\ c_n=c}}^{N} -1 + \frac{1}{\sigma_c^2}(x_n - \mu_c)^2 = 0 \rightarrow$$

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} -1 + \frac{1}{\sigma_c^2}(x - \mu_c)^2 = 0 \rightarrow \sum_{\substack{n=1:\\ c_n=c}}^{N} \frac{1}{\sigma_c^2}(x - \mu_c)^2 = \sum_{\substack{n=1:\\ c_n=c}}^{N} 1$$

Usando de nuevo la definición de N_c, esto queda

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} \frac{1}{\sigma_c^2}(x_n - \mu_c)^2 = N_c \rightarrow \frac{1}{\sigma_c^2} \sum_{\substack{n=1:\\ c_n=c}}^{N} (x_n - \mu_c)^2 = N_c \rightarrow$$

$$\sigma_c^2 = \frac{1}{N_c} \sum_{\substack{n=1:\\ c_n=c}}^{N} (x_n - \mu_c)^2$$

El resultado es que σ_c^2 es la varianza de los datos de la clase c.

5.2. Caso multidimensional

Para el caso multidimensional (con dimensión D), tendremos que Θ_c son $\boldsymbol{\mu}_c$ y Σ_c, vector media y matriz de covarianzas respectivamente. El problema de optimización se establece entonces como

$$
\begin{aligned}
&\hat{\boldsymbol{\Theta}}_c = (\hat{\boldsymbol{\mu}}_c, \hat{\Sigma}_c) = \underset{\boldsymbol{\mu}_c, \Sigma_c}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\Theta}; X) \\
&= \underset{\boldsymbol{\mu}_c, \Sigma_c}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{c_n} + \log \left((2\pi)^{-\frac{D}{2}} |\Sigma_{c_n}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n}) \right) \right) \\
&= \underset{\boldsymbol{\mu}_c, \Sigma_c}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{c_n} - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_{c_n}| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})
\end{aligned}
$$

A partir de aquí, se plantea la derivación de cada uno de los parámetros para optimizarlos.

En el caso de $\boldsymbol{\mu}_c$:

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \boldsymbol{\mu}_c} = \sum_{n=1}^{N} & \frac{\partial \left(\log p_{c_n} \right)}{\partial \boldsymbol{\mu}_c} - \frac{\partial \left(\frac{D}{2} \log(2\pi) \right)}{\partial \boldsymbol{\mu}_c} - \frac{\partial \left(\frac{1}{2} \log |\Sigma_{c_n}| \right)}{\partial \boldsymbol{\mu}_c} \\
& - \frac{\partial \left(\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n}) \right)}{\partial \boldsymbol{\mu}_c}
\end{aligned}
$$

Como puede verse, los tres primeros términos son nulos al no variar respecto a $\boldsymbol{\mu}_c$, mientras que el cuarto término es la derivada respecto al vector $\boldsymbol{\mu}_c$ que es muy similar a la derivada del error de reconstrucción respecto al vector de proyección que se realiza en PCA*.

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \boldsymbol{\mu}_c} &= \sum_{n=1}^{N} -\frac{1}{2} \frac{\partial \left((\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n}) \right)}{\partial \boldsymbol{\mu}_c} \\
&= \sum_{\substack{n=1: \\ c_n = c}}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_c)^t \Sigma_c^{-1}
\end{aligned}
$$

*Páginas 5 y 6 del documento “Derivación del problema de optimización de PCA”

Para terminar la optimización se iguala al vector nulo transponiendo el resultado anterior:

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} \Sigma_c^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_c\right) = \mathbf{0} \rightarrow \Sigma_c^{-1} \sum_{\substack{n=1:\\ c_n=c}}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}_c\right) = \mathbf{0}$$

$$\sum_{\substack{n=1:\\ c_n=c}}^{N} \boldsymbol{\mu}_c = \sum_{\substack{n=1:\\ c_n=c}}^{N} \mathbf{x}_n \rightarrow N_c\, \boldsymbol{\mu}_c = \sum_{\substack{n=1:\\ c_n=c}}^{N} \mathbf{x}_n$$

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\substack{n=1:\\ c_n=c}}^{N} \mathbf{x}_n$$

Respecto a la matriz de covarianzas Σ_c, tenemos que la derivada es:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}; X)}{\partial \Sigma_c} = \sum_{\substack{n=1:\\ c_n=c}}^{N} \frac{\partial \left(\log p_{c_n}\right)}{\partial \Sigma_c} - \frac{\partial \left(\frac{D}{2} \log(2\pi)\right)}{\partial \Sigma_c} - \frac{\partial \left(\frac{1}{2} \log |\Sigma_{c_n}|\right)}{\partial \Sigma_c}$$
$$- \frac{\partial \left(\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})\right)}{\partial \Sigma_c}$$

Los dos primeros términos son nulos, con lo cual queda conseguir las derivadas del tercer y cuarto término. Respecto al tercer término:

$$\frac{\partial \left(\frac{1}{2} \log |\Sigma_{c_n}|\right)}{\partial \Sigma_c} = \frac{1}{2} \frac{\partial \left(\log |\Sigma_{c_n}|\right)}{\partial \Sigma_c}$$

Aplicando la regla $\frac{d}{dA} \log |A| = \operatorname{Tr}\left(A^{-1}\right)^{\dagger}$, queda finalmente:

$$\frac{\partial \left(\frac{1}{2} \log |\Sigma_{c_n}|\right)}{\partial \Sigma_c} = \frac{1}{2} \operatorname{Tr}\left(\Sigma_c^{-1}\right)$$

Respecto al cuarto término, se debe tener en cuenta la propiedad $\boldsymbol{v}^t A \boldsymbol{v} = \operatorname{Tr}(A\boldsymbol{v}\boldsymbol{v}^t)$, con $v \in \mathbb{R}^{D \times 1}$ y $A \in \mathbb{R}^{D \times D}$. Como también $d(\operatorname{Tr}(A)) = \operatorname{Tr} dA$, queda:

$$\frac{\partial \left(\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})\right)}{\partial \Sigma_c} = \frac{1}{2} \frac{\partial \left(\left(\mathbf{x}_n - \boldsymbol{\mu}_{c_n}\right)^t \Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})\right)}{\partial \Sigma_c} =$$

$$\frac{1}{2} \frac{\partial \left(\operatorname{Tr}\left(\Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})(\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t\right)\right)}{\partial \Sigma_c} = \frac{1}{2} \operatorname{Tr} \frac{\partial \left(\Sigma_{c_n}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{c_n})(\mathbf{x}_n - \boldsymbol{\mu}_{c_n})^t\right)}{\partial \Sigma_c}$$

Teniendo en cuenta ahora que la parte vectorial $(\mathbf{x}_n - \boldsymbol{\mu}_{c_n})$ es constante con respecto a Σ_c, queda:

† `https://tminka.github.io/papers/matrix/`

$$\frac{\partial\left(\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_{c_n})^t\Sigma_{c_n}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_{c_n})\right)}{\partial\Sigma_c}=\frac{1}{2}\operatorname{Tr}\frac{\partial\left(\Sigma_{c_n}^{-1}\right)}{\partial\Sigma_c}(\mathbf{x}_n-\boldsymbol{\mu}_{c_n})(\mathbf{x}_n-\boldsymbol{\mu}_{c_n})^t$$

Y aplicando que $\frac{d}{dA}A^{-1}=-A^{-1}A^{-1\dagger}$, este cuarto término queda como:

$$\frac{1}{2}\operatorname{Tr}\left(-\Sigma_c^{-1}\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)$$

En conclusión, el problema de optimización queda como:

$$\frac{\partial\mathcal{L}(\boldsymbol{\Theta};X)}{\partial\Sigma_c}=\sum_{\substack{n=1:\\c_n=c}}^{N}-\frac{1}{2}\operatorname{Tr}\left(\Sigma_c^{-1}\right)-\frac{1}{2}\operatorname{Tr}\left(-\Sigma_c^{-1}\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)$$

Como la traza presenta la propiedad $\operatorname{Tr}(A)+\operatorname{Tr}(B)=\operatorname{Tr}(A+B)$, el desarrollo sigue como:

$$\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\Theta};X)}{\partial\Sigma_c}&=-\frac{1}{2}\sum_{\substack{n=1:\\c_n=c}}^{N}\operatorname{Tr}\left(\Sigma_c^{-1}-\Sigma_c^{-1}\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)\\
&=-\frac{1}{2}\operatorname{Tr}\left(\sum_{\substack{n=1:\\c_n=c}}^{N}\Sigma_c^{-1}-\Sigma_c^{-1}\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)\\
&=-\frac{1}{2}\operatorname{Tr}\left(\Sigma_c^{-1}\sum_{\substack{n=1:\\c_n=c}}^{N}I-\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)
\end{aligned}$$

Igualando a cero, y considerando que igualar a cero una traza equivale (en términos de optimización) a igualar a la matriz nula la matriz sobre la que se opera, tendremos:

$$
\begin{aligned}
-\frac{1}{2}\,\mathrm{Tr}\left(\Sigma_c^{-1}\sum_{\substack{n=1:\\c_n=c}}^{N} I-\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t\right)&=0\\
\Sigma_c^{-1}\sum_{\substack{n=1:\\c_n=c}}^{N} I-\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t&=0\\
\sum_{\substack{n=1:\\c_n=c}}^{N} I-\Sigma_c^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t&=0\\
N_c\, I-\Sigma_c^{-1}\sum_{\substack{n=1:\\c_n=c}}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t&=0\\
\Sigma_c^{-1}\sum_{\substack{n=1:\\c_n=c}}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t&=N_c\, I\\
\Sigma_c=\frac{1}{N_c}\sum_{\substack{n=1:\\c_n=c}}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_c)(\mathbf{x}_n-\boldsymbol{\mu}_c)^t&
\end{aligned}
$$