25. Consider a non-singular transformation of the feature space: $y = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a d-by-d non-singular matrix.

(a) If we let $\tilde{\mathcal{D}}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}_i\}$ denote the data set transformed to the new space, then the scatter matrix in the transformed domain can be written as

$$\begin{aligned}
\mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{y} - \mathbf{m}_i^y)(\mathbf{y} - \mathbf{m}_i^y)^t \\
&= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)^t \\
&= \mathbf{A} \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t
\end{aligned}$$

where $\mathbf{A}^t = \mathbf{A}\mathbf{S}_W\mathbf{A}^t$. We also have the between-scatter matrix

$$\begin{aligned}
\mathbf{S}_B^y &= \sum_{i=1}^{c} n_i(\mathbf{m}_i^y - \mathbf{m}^y)(\mathbf{m}_i - \mathbf{m}^y)^t \\
&= \sum_{i=1}^{c} n_i(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})^t \\
&= \mathbf{A}\left[\sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t\right]\mathbf{A}^t \\
&= \mathbf{A}\mathbf{S}_B\mathbf{A}^t.
\end{aligned}$$

The product of the inverse matrices is

$$\begin{aligned}[\mathbf{S}_W^y]^{-1}[\mathbf{S}_B^y]^{-1} &= (\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_B\mathbf{A}^t)\\ &= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B\mathbf{A}^t\\ &= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t.\end{aligned}$$

We let λ_i for $i = 1, \ldots, d$ denote the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$. There exist vectors $\mathbf{z}_i, \ldots, \mathbf{z}_d$ such that

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{z}_i = \lambda_i\mathbf{z}_i,$$

for $i = 1, \ldots, d$, and this in turn implies

$$(\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t\mathbf{A}^t)^{-1}\mathbf{z}_i = \lambda_i(\mathbf{A}^t)^{-1}\mathbf{z}_i,$$

or

$$\mathbf{S}_W^{y-1}\mathbf{S}_B^y\mathbf{u}_i = \lambda_i\mathbf{u}_i,$$

where $\mathbf{u}_i = (\mathbf{A}^t)^{-1}\mathbf{z}_i$. This implies that $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{y-1}\mathbf{S}_B^y$, and finally that $\lambda_1, \ldots, \lambda_d$ are invariant to non-singular linear transformation of the data.

(b) Our total scatter matrix is $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, and thus

$$\begin{aligned}\mathbf{S}_T^{-1}\mathbf{S}_W &= (\mathbf{S}_B + \mathbf{S}_W)^{-1}\mathbf{S}_W\\ &= [\mathbf{S}_W^{-1}(\mathbf{S}_B + \mathbf{S}_W)]^{-1}\\ &= [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}.\end{aligned}$$

If $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and the $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the corresponding eigenvectors, then $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for $i = 1, \ldots, d$ and hence

$$\mathbf{u}_i + \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \mathbf{u}_i + \lambda_i\mathbf{u}_i.$$

This equation implies

$$[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]\mathbf{u}_i = (1 + \lambda_i)\mathbf{u}_i.$$

We multiply both sides of the equation by $(1 + \lambda_i)^{-1}[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}$ and find

$$(1 + \lambda_i)^{-1}\mathbf{u}_i = [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}\mathbf{u}_i$$

and this implies $\nu_i = 1/(1 + \lambda_i)$ are eigenvalues of $\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B$.

(c) We use our result from part (a) and find

$$J_d = \frac{|\mathbf{S}_W|}{|\mathbf{S}_T|} = |\mathbf{S}_T^{-1}\mathbf{S}_W| = \prod_{i=1}^{d}\nu_i = \prod_{i=1}^{d}\frac{1}{1 + \lambda_i},$$

which is invariant to non-singular linear transformations described in part (a).

26. Consider a non-singular transformation of the feature space: $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a d-by-d non-singular matrix. We let $\mathcal{D}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}^x\}$. We have

$$\begin{aligned}
\mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y}\in\mathcal{D}_i} (\mathbf{y} - \mathbf{m}^y)(\mathbf{y} - \mathbf{m}^y)^t \\
&= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i^x)(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i^x)^t \\
&= \mathbf{A}\mathbf{S}_W^x\mathbf{A}^t.
\end{aligned}$$

In a similar way, we have

$$\begin{aligned}
\mathbf{S}_B^y &= \mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\
\mathbf{S}_t^y &= \mathbf{A}\mathbf{S}_t^x\mathbf{A}^t \\
(\mathbf{S}_t^y)^{-1}\mathbf{S}_W^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_W^x\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x\mathbf{A}^t \\
(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t.
\end{aligned}$$

(a) From problem 25 (b), we know that

$$\begin{aligned}
\mathrm{tr}[(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x] &= \sum_{i=1}^{d} \nu_i \\
&= \sum_{i=1}^{d} \frac{1}{1+\lambda_i}
\end{aligned}$$

as well as $\mathrm{tr}[\mathbf{B}^{-1}\mathbf{S}\mathbf{B}]$, because they have the same eigenvalues so long as $\mathbf{B}$ is non-singular. This is because if $\mathbf{S}\mathbf{x} = \nu_i\mathbf{x}$, then $\mathbf{S}\mathbf{B}\mathbf{B}^{-1}\mathbf{x} = \nu_i\mathbf{x}$, then also $\mathbf{B}^{-1}\mathbf{S}\mathbf{B}\mathbf{B}^{-1}\mathbf{x} = \mathbf{B}^{-1}\nu_i\mathbf{x} = \nu_i\mathbf{B}^{-1}\mathbf{x}$. Thus we have

$$\mathbf{B}^{-1}\mathbf{S}\mathbf{B}(\mathbf{B}^{-1}\mathbf{x}) = \nu_i(\mathbf{B}^{-1}\mathbf{x}).$$

We put this together and find

$$\begin{aligned}
\mathrm{tr}[(\mathbf{S}_t^y)^{-1}\mathbf{S}_W^y] &= \mathrm{tr}[(\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x\mathbf{A}^t] \\
&= \mathrm{tr}[(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x] \\
&= \sum_{i=1}^{d} \frac{1}{1+\lambda_i}.
\end{aligned}$$

(b) See Solution to Problem 25 part (c).

(c) Here we have the determinant

$$\begin{aligned}
|(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y| &= |(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t| \\
&= \prod \text{eigenvalues of } [(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t] \\
&= \prod \text{eigenvalues of } [(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x] \\
&= \prod_{i=1}^{d} \lambda_i.
\end{aligned}$$

(d) The typical value of the criterion is zero or close to zero. This is because $\mathbf{S}_B$ is often singular, even if samples are not from a subspace. Even when $\mathbf{S}_B$ is not singular, some λ_i is likely to be very small, and this makes the product small. Hence the criterion is not always useful.

27. Equation 68 in the text defines the criterion $J_d = |\mathbf{S}_W| = \left|\sum_{i=1}^{c} \mathbf{S}_i\right|$, where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

is the scatter matrix for category ω_i, defined in Eq. 61 in the text. We let $\mathbf{T}$ be a non-singular matrix and consider the change of variables $\mathbf{x}' = \mathbf{T}\mathbf{x}$.

(a) From the conditions stated, we have

$$\mathbf{m}_i' = \frac{1}{n_i} \sum_{\mathbf{x}'\in\mathcal{D}_{i'}} \mathbf{x}'$$

where n_i is the number of points in category ω_i. Thus we have the mean of the transformed data is

$$\mathbf{m}_i' = \frac{1}{n_i} \sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{T}\mathbf{x} = \mathbf{T}\mathbf{m}_i.$$

Furthermore, we have the transformed scatter matrix is

$$\begin{aligned}
\mathbf{S}_i' &= \sum_{\mathbf{x}'\in\mathcal{D}_i'} (\mathbf{x}' - \mathbf{m}_i')(\mathbf{x}' - \mathbf{m}_i')^t \\
&= \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{T}\mathbf{x} - \mathbf{T}\mathbf{m}_i)(\mathbf{T}\mathbf{x} - \mathbf{T}\mathbf{m}_i)^t \\
&= \mathbf{T}\left[\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\right]\mathbf{T}^t = \mathbf{T}\mathbf{S}_i\mathbf{T}^t.
\end{aligned}$$

(b) From the conditions stated by the problem, the criterion function of the transformed data must obey

$$\begin{aligned}
J_d' = |\mathbf{S}_W'| &= \left|\sum_{i=1}^{c} \mathbf{S}_i'\right| = \left|\sum_{i=1}^{c} \mathbf{T}\mathbf{S}_i\mathbf{T}^t\right| = \left|\mathbf{T}\left(\sum_{i=1}^{c} \mathbf{S}_i\right)\mathbf{T}^t\right| \\
&= |\mathbf{T}||\mathbf{T}^t|\left|\sum_{i=1}^{c} \mathbf{S}_i\right| \\
&= |\mathbf{T}|^2 J_d.
\end{aligned}$$

Therefore, J_d' differs from J_d only by an overall non-negative scale factor $|\mathbf{T}|^2$.

(c) Since J_d' differs from J_d only by a scale factor of $|\mathbf{T}|^2$ (which does not depend on the partitioning into clusters) J_d' and J_d will rank partitions in the same order. Hence the optimal clustering based on J_d is always the optimal clustering based on J_d'. Optimal clustering is invariant to non-singular linear transformations of the data .

28. Consider a non-singular transformation of the feature space $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{A}$ is a d-by-d non-singular matrix. We let $\mathcal{D}_i = \{\mathbf{Ax} : \mathbf{x} \in \mathcal{D}_i^x\}$ be the transformed data set. We then have the within scatter matrix as

$$\begin{aligned}
\mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathcal{D}_i} (\mathbf{y} - \mathbf{m}_i^y)(\mathbf{y} - \mathbf{m}_i^y)^t \\
&= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathcal{D}_i} (\mathbf{Ax} - \mathbf{Am}_i^x)(\mathbf{y} - \mathbf{Am}_i^x)^t \\
&= \mathbf{AS}_W^x \mathbf{A}^t.
\end{aligned}$$

In a similar way, we have

$$\begin{aligned}
\mathbf{S}_B^y &= \mathbf{AS}_B^x \mathbf{A}^t \\
\mathbf{S}_t^y &= \mathbf{AS}_t^x \mathbf{A}^t \\
(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{A}^{-1}\mathbf{AS}_B^x\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t.
\end{aligned}$$

If λ is an eigenvalue of $(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x$ with corresponding eigenvector $\mathbf{x}$, that is, $(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{x} = \lambda\mathbf{x}$, then we have

$$(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x \underbrace{(\mathbf{A}^t(\mathbf{A}^t)^{-1})}_{\mathbf{I}} \mathbf{x} = \lambda\mathbf{x},$$

which is equivalent to

$$(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t((\mathbf{A}^t)^{-1}\mathbf{x}) = \lambda\mathbf{x}.$$

We multiply both sides on the left by $(\mathbf{A}^t)^{-1}$ and find

$$(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t((\mathbf{A}^t)^{-1}\mathbf{x}) = (\mathbf{A}^t)^{-1}\lambda\mathbf{x},$$

which yields

$$(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y((\mathbf{A}^t)^{-1}\mathbf{x}) = \lambda((\mathbf{A}^t)^{-1}\mathbf{x}).$$

Thus we see that λ is an eigenvalue of $(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y$ with corresponding eigenvector $(\mathbf{A}^t)^{-1}\mathbf{x}$.