Consider the Fisher's LDA for two populations. The original data are:

$$\text{First population: } \mathbf{y}_{1i} = (y_{1i1}, \ldots, y_{1ip})', \ i = 1, \ldots, n_1, \text{ with mean vector } \bar{\mathbf{y}}_1;$$
$$\text{Second population: } \mathbf{y}_{2i} = (y_{2i1}, \ldots, y_{2ip})', \ i = 1, \ldots, n_2, \text{ with mean vector } \bar{\mathbf{y}}_2.$$

They have the common covariance matrix $\mathbf{\Sigma}$, which is estimated by

$$\mathbf{S}_{pl\cdot y} = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)(\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)'$$

Correspondingly, the common correlation matrix is denoted by $\mathbf{P}_y$, estimated by $\mathbf{R}_y$.

The Fisher's LDA yields discriminant coefficient vector

$$\mathbf{a} = (a_1, \ldots, a_p)' = \mathbf{S}_{pl\cdot y}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2),$$

and the discriminant function

$$z = \mathbf{a}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl\cdot y}^{-1}\mathbf{y}.$$

That is, for original observations $\mathbf{y}_{ki}$, Fisher's LDA provides projected new observations

$$z_{ki} = \mathbf{a}'\mathbf{y}_{ki} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl\cdot y}^{-1}\mathbf{y}_{ki}, \ k = 1, 2; \ i = 1, \ldots, n_k$$

to maximize the statistical distance between $\bar{z}_1$ and $\bar{z}_2$. The corresponding maximized statistical distance is

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl\cdot y}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

Remark: Each element a_j in $\mathbf{a}$ can be interpreted as the "contribution/relative importance" of the original variable Y_j in $\mathbf{y}$ to the discrimination. But this is reasonable only when $Y_j, \ j = 1, \ldots, p$ are commensurate.

If not commensurate, a_j is highly influenced by the variability/measurement unit of Y_j itself, rather than its "contribution to discrimination". Therefore, under this circumstance, we often standardize each dimension of $\mathbf{y}$ before conducting Fisher's LDA by taking the ratio between original observations and their individual standard deviations. That is, LDA is applied to the standardized observations

$$\text{First population: } \mathbf{w}_{1i} = (w_{1i1}, \ldots, w_{1ip})', \; i = 1, \ldots, n_1, \text{ with mean vector } \bar{\mathbf{w}}_1;$$
$$\text{Second population: } \mathbf{w}_{2i} = (w_{2i1}, \ldots, w_{2ip})', \; i = 1, \ldots, n_2, \text{ with mean vector } \bar{\mathbf{w}}_2,$$

where $w_{kij} = y_{kij}/s_{j\cdot y}$, and $s_{j\cdot y}$ is the square root of the jth diagonal element of $\mathbf{S}_{pl\cdot y}$, i.e., the common standard deviation of the jth original variable.

(1) Show that $\mathbf{w}_{ki} = \mathbf{D}_s^{-1}\mathbf{y}_{ki}$, where $\mathbf{D}_s = diag[\{diag(\mathbf{S}_{pl\cdot y})\}^{1/2}]$, the $p \times p$ diagonal matrix with each diagonal element $s_{j\cdot y}$. (The notation here is consistent with Chapter 2.)

(2) Verify that the estimated covariance matrix of $\mathbf{w}_{ki}$ is $\mathbf{R}_y$.

(3) Apply Fisher's LDA to the standardized observations. You might directly use the result of LDA. Show that the discriminant coefficient vector $\mathbf{b}$ is

$$\mathbf{b} = \mathbf{D}_s\mathbf{a}.$$

This implies the discriminant coefficient vector does change upon standardization - meaning the interpretation of "relative importance/contribution" of each variable to discrimination might change.

(4) Obtain the projected new observations. They should still be z_{ki}'s above. That is, the final result of discrimination remains nevertheless. So does the maximum distance.

In a word, individual standardization would change the coefficient vector and its interpretation, but would not affect the discrimination result. This is because we consider STATISTICAL distance anyways. If we do individual standardization first, it's like we are going through "two steps" - 1. individual standardization by $\mathbf{D}_s$; 2. the covariance matrix $\mathbf{S}$ used in the statistical distance degenerates to the correlation matrix $\mathbf{R}$. If we do not standardize first, we directly use $\mathbf{S}$ in the statistical distance.