Dylan Plecki
cs 498 HW 9.2

10.3 a) Let Σ represent covmat($\{\vec{x}\}$)
$\Rightarrow \vec{v} = \vec{x}_i - \text{mean}(\{\vec{x}\})$
And let $V = [\vec{v}_1 \dots \vec{v}_n]$ (from pg. 305)

So, $\Sigma = \frac{1}{N} V V^T$

Let U be the eigenvalue of Σ

$$U = [\vec{E}_1 \dots \vec{E}_p]$$

And $P = U^T V$ (projected data onto eigenvectors)
So,
$$\text{covmat}(\{\vec{p}\}) = \frac{1}{N} P P^T = \Lambda \quad (\vec{p}_i = \vec{u}_i \cdot \vec{v}_i)$$

As Σ has one eigenvalue, $\Lambda \in \mathbb{R}^{1\times 1} \Rightarrow \Lambda = \lambda_1$
We also know $\vec{p}_i \in \mathbb{R}^{1\times 1}$, so $\vec{p}_i = p_i$
Thus, $p_i = p_1 + t_i (p_2 - p_1)$ for some $t_i \in t$

Since $p_i = u^T \vec{v}_i$, and $\vec{v}_i = \vec{x}_i - \text{mean}(\{\vec{x}_i\})$,

$$\vec{u}^T \vec{v}_i = \vec{u}^T \vec{v}_1 + t_i (\vec{u}^T \vec{v}_2 - \vec{u}^T \vec{v}_1)$$
$$\vec{u}^T \vec{v}_i = \vec{u}^T (\vec{v}_1 + t_i (\vec{v}_2 - \vec{v}_1))$$
$$\vec{v}_i = \vec{v}_1 + t_i (\vec{v}_2 - \vec{v}_1)$$

And as $\vec{x}_i - \text{mean} = (\vec{x}_1 - \text{mean}) + t_i (\vec{x}_2 - \text{mean})$

So, $x_i = x_1 + t_i (x_2 - x_1)$

□

(continued)

10.3 b) We know that $t_i = \frac{P_i - P_1}{P_2 - P_1}$ (from part a)

We want $\text{std}(t_i) = \text{std}\left(\frac{P_i - P_1}{P_2 - P_1}\right)$

$\text{std}(t_i) \Rightarrow \frac{1}{P_2 - P_1} \text{std}(P_i)$

$\text{Var}(t_i) = \frac{1}{(P_2 - P_1)^2} \text{Var}(P_i)$

We also know (from def.):

$$\text{covmat}(\{P_i\}) = \frac{1}{N} P \cdot P^T = \Lambda \in \mathbb{R}^{p \times p}$$

And since we know that there is 1 eigenvalue,

$$\text{covmat}(\{P_i\}) = \lambda_1$$

So,

$$\text{Var}(t_i) = \frac{1}{(P_2 - P_1)^2} (\lambda_1)$$

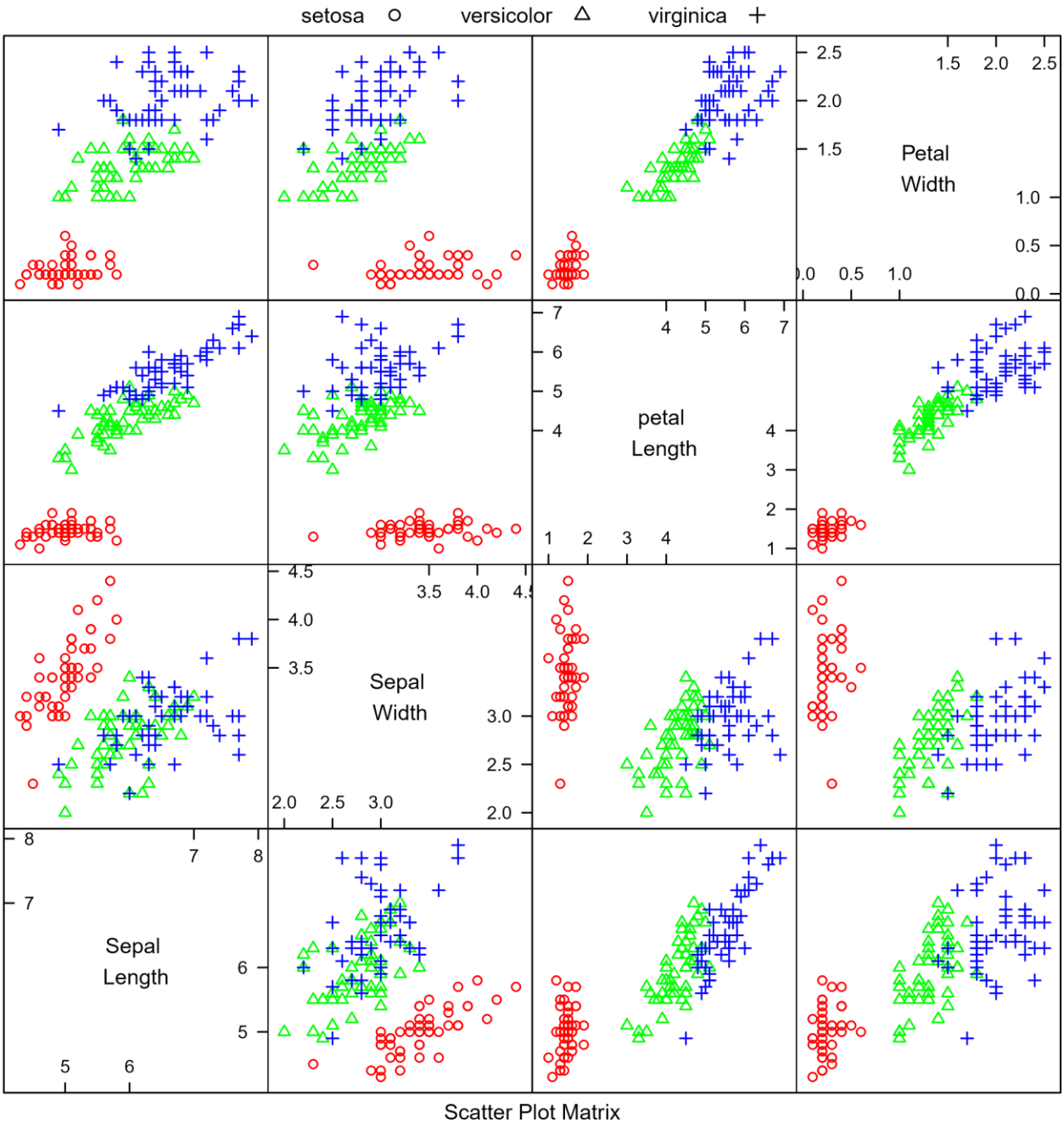
and,

$$\text{std}(t_i) = \frac{1}{|P_2 - P_1|} \cdot \sqrt{\lambda_1}$$

□

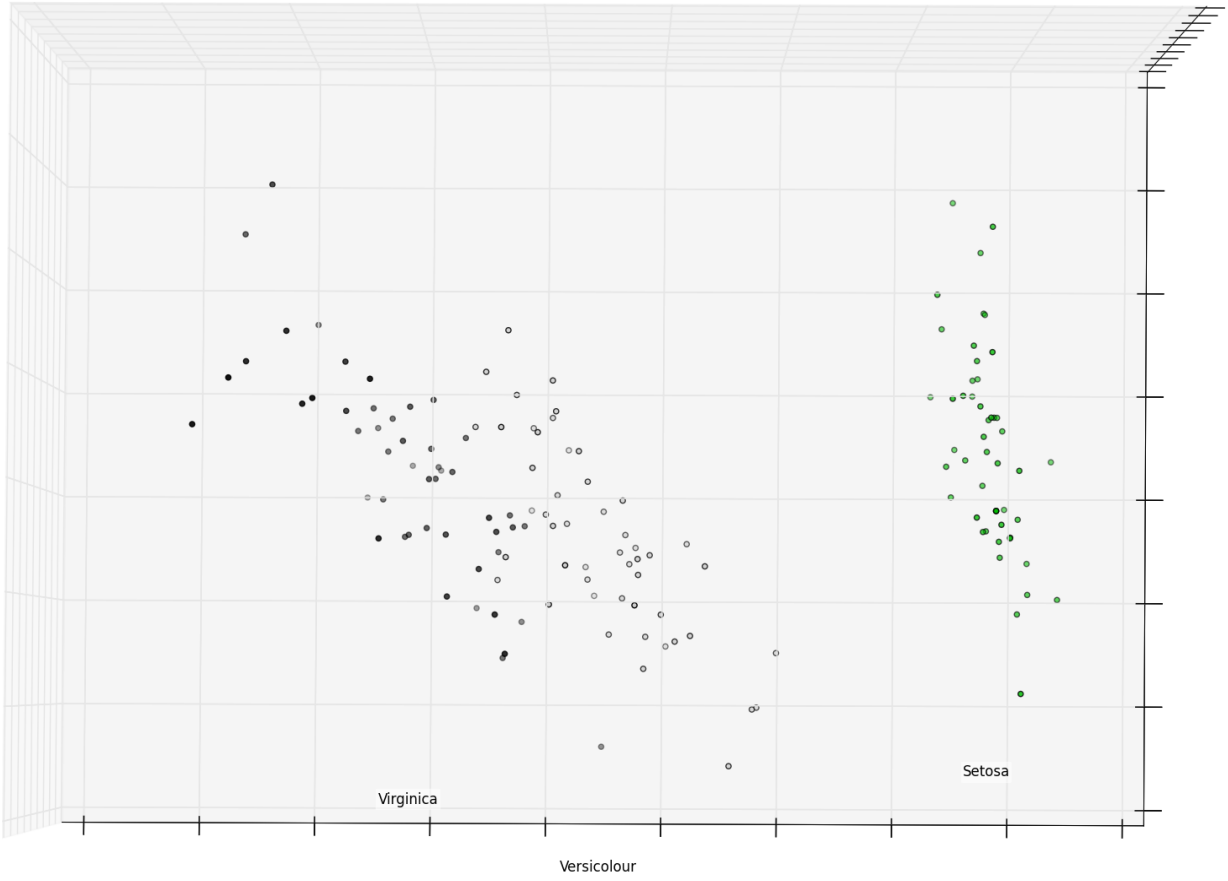
10.4)

a) Plot a scatterplot matrix of the dataset.



10.4) *Continued*

b) *Plot the data on those two principal components.*



b) *Continued*. Has this plot introduced significant distortions? Explain.

No. This new plot, in addition to the above scatter plot matrix of the data set, does not introduce any significant distortions. In fact, it almost identically mirrors the information provided by the scatter plot matrix (in relation to blobs). But, as this plot is only a plot of two components of the PCA, it is impossible to definitively conclude from *this plot alone* that there are in fact two distinct blobs in the dataset; these may be connected in another component, and may only show vary slight variations here. So although this plot has introduced no new distortions to the data, it alone is a distorted view of the data set in the fact that it only shows an incomplete picture of all of the components.