PCA Discussion

The following note is useful for this discussion: Note 17.

1. Conceptual PCA

(a) Consider a data matrix $A \in \mathbb{R}^{d \times n}$, where n is the number of data points and d is the dimensionality of each data point. Recall that PCA solves the problem of

$$\underset{W \in \mathbb{R}^{d \times \ell}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \vec{x}_i - WW^\top \vec{x}_i \right\|^2 \tag{1}$$

where $W^\top W = I_\ell$ is a rank ℓ matrix. For $\ell = 1$ (i.e., to find the first principal component), this can be rewritten as

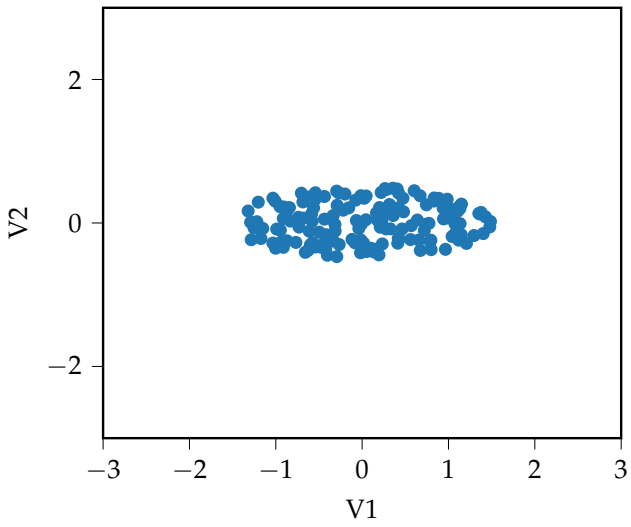
$$\underset{\vec{u} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \| \vec{x}_i - \langle \vec{x}_i, \vec{u} \rangle \vec{u} \|^2 \tag{2}$$

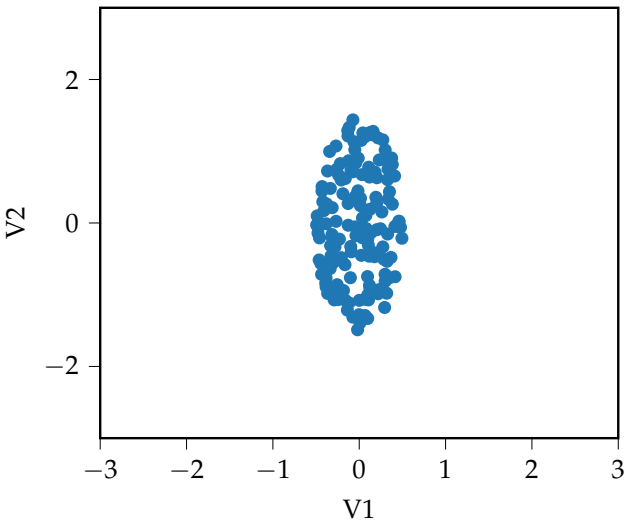
where $\|\vec{u}\| = 1$. **Show that finding the top principal component is equivalent to maximizing**

$$\sum_{i=1}^{n} \langle \vec{x}_i, \vec{u} \rangle^2 \tag{3}$$

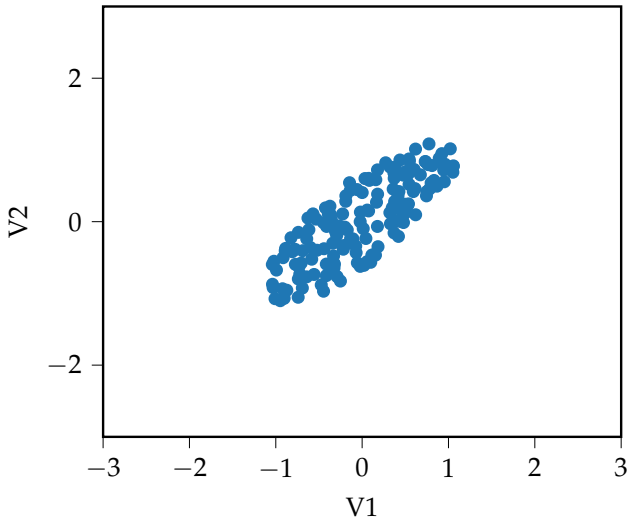
In each plot below, the data is projected onto two unit vectors. The x coordinate is the projection onto the first vector (written as "V1" or $\vec{v}_1$), and the y coordinate is the projection onto the second vector (written as "V2" or $\vec{v}_2$). We say that a plot is "valid" if the first vector would be the first principal component, and if the second vector would correspond to the second principal component. For each subpart, explain your answer.

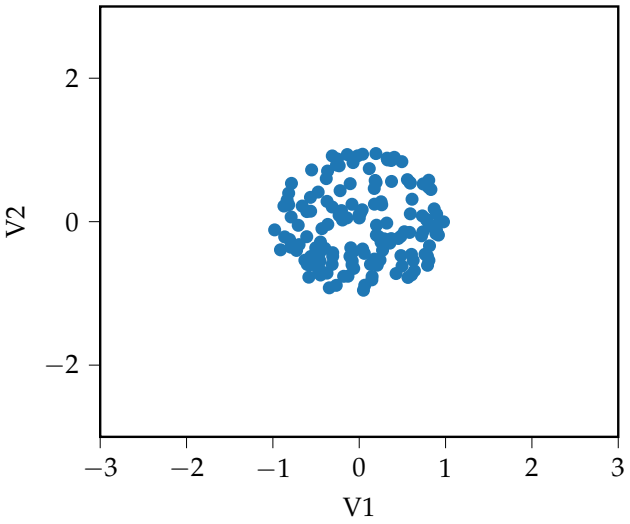
(b) **Which of these two plots is valid?**



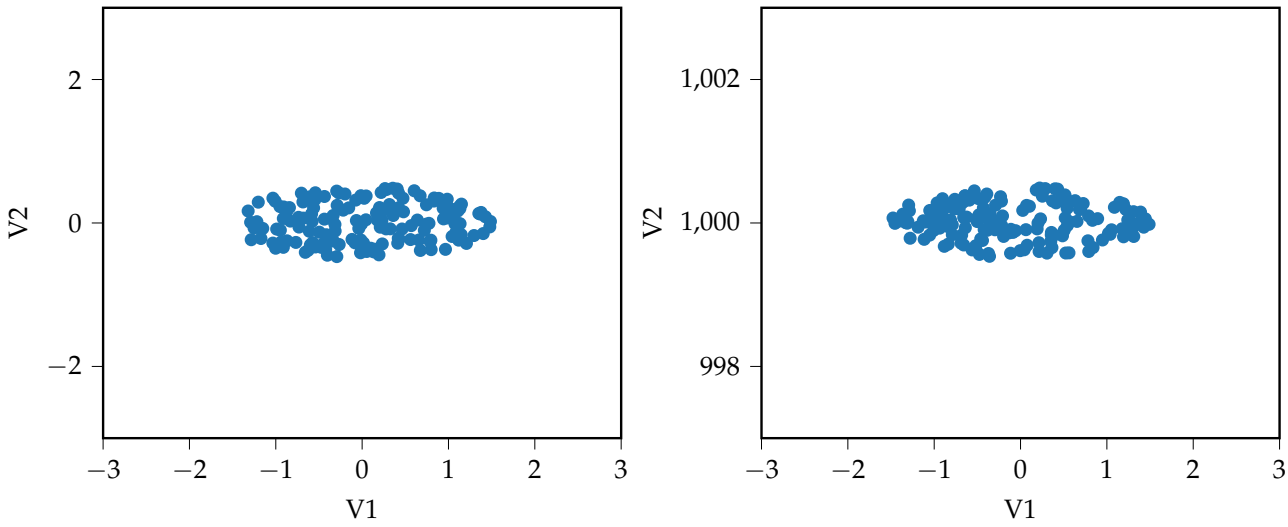


(c) **Which of these two plots is valid?**





(d) **Which of these two plots is valid?**



(HINT: Be careful of the axis scale on the right plot.)

(e) **(OPTIONAL)** As you may have noticed in parts **1.b** and **1.c**, the direction of greatest spread in the data tends to be the direction of the first principal component. Further note that all of the data in those examples are centered at the origin. This is no coincidence. **Why do we see the behavior that we do in part 1.d?**