EECS 16B Designing Information Systems and Devices II UC Berkeley Spring 2022

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

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

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

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

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

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

Solution: We can begin by writing out the norm as an inner product in eq. (2):

$$\begin{aligned}
&\sum_{i=1}^{n} \left\| \vec{x}_i - \langle \vec{x}_i, \vec{u} \rangle \vec{u} \right\|^2 && (4)\\
&= \sum_{i=1}^{n} \left(\vec{x}_i - \langle \vec{x}_i, \vec{u} \rangle \vec{u} \right)^\top \left(\vec{x}_i - \langle \vec{x}_i, \vec{u} \rangle \vec{u} \right) && (5)\\
&= \sum_{i=1}^{n} \langle \vec{x}_i, \vec{x}_i \rangle - 2 \langle \vec{x}_i, \vec{u} \rangle^2 + \langle \vec{x}_i, \vec{u} \rangle^2 \underbrace{\langle \vec{u}, \vec{u} \rangle}_{1} && (6)\\
&= \sum_{i=1}^{n} \|\vec{x}_i\|^2 - \langle \vec{x}_i, \vec{u} \rangle^2 && (7)
\end{aligned}$$

We cannot change $\|\vec{x}_i\|^2$ (the data points are fixed), so minimizing this expression is the same as minimizing

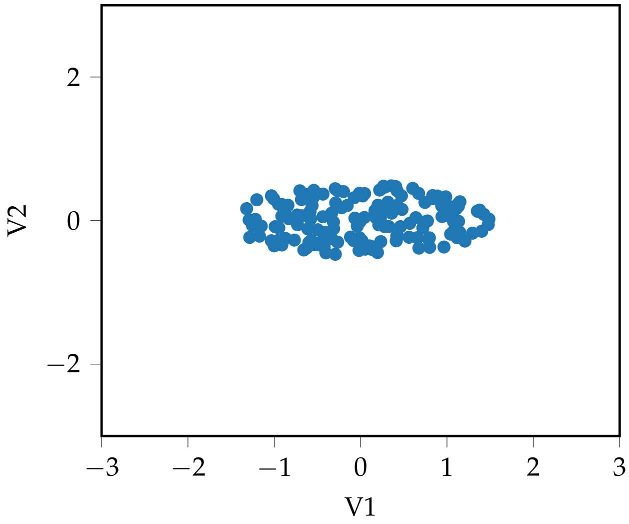
$$\sum_{i=1}^{n} \left(- \langle \vec{x}_i, \vec{u} \rangle^2 \right) = - \sum_{i=1}^{n} \langle \vec{x}_i, \vec{u} \rangle^2 \tag{8}$$

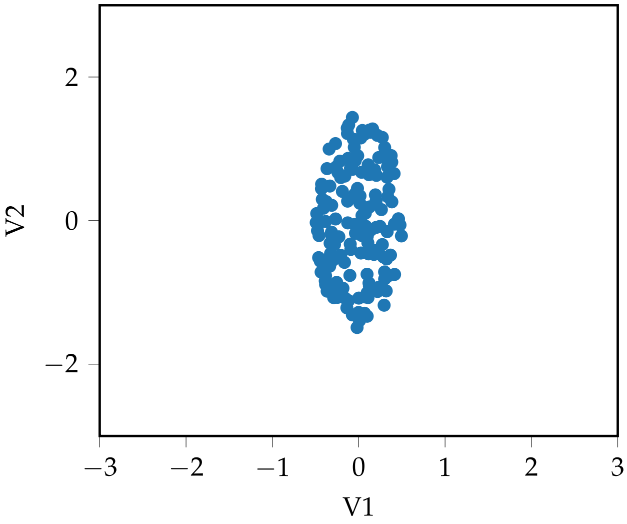
which is the same as maximizing

$$\sum_{i=1}^{n} \langle \vec{x}_i, \vec{u} \rangle^2 \tag{9}$$

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?**





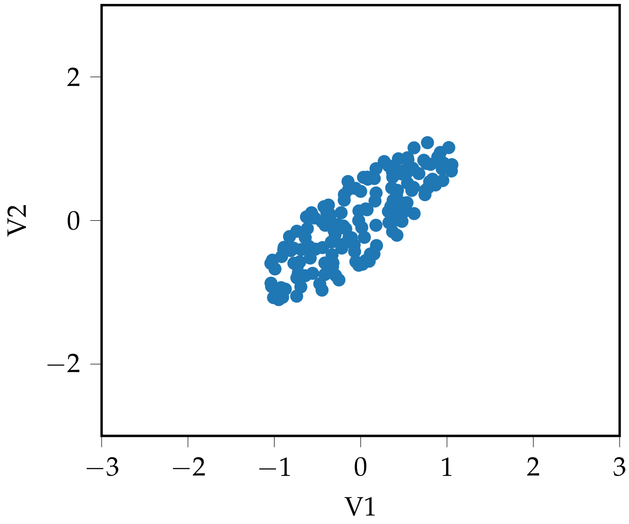
Solution: To check if a plot is valid, we can see whether the $\vec{v}_1$ satisfies the definition of the first principal component, namely that it should maximize

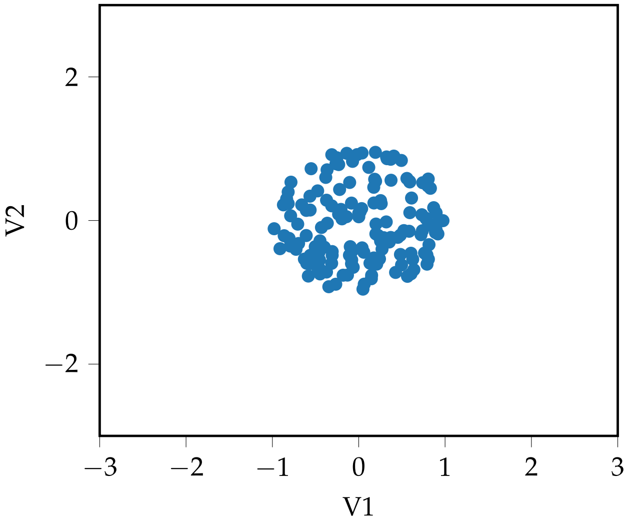
$$\sum_{i=1}^{n} \langle \vec{x}_i,\ \vec{v}_1 \rangle^2 \tag{10}$$

In other words, this is the same as maximizing the sum of squares of the x-coordinates in the plots above.

In this case, the plot on the left has the largest range of values, from -2 to 2. This is larger than the spread along the y-axis, so the sum of squares of x-coordinates is maximized. Hence, the left plot is valid. For the right plot, there exists another direction (i.e. the direction of $\vec{v}_2$) that maximizes the sum of squares, so $\vec{v}_1$ cannot be the first principal component.

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



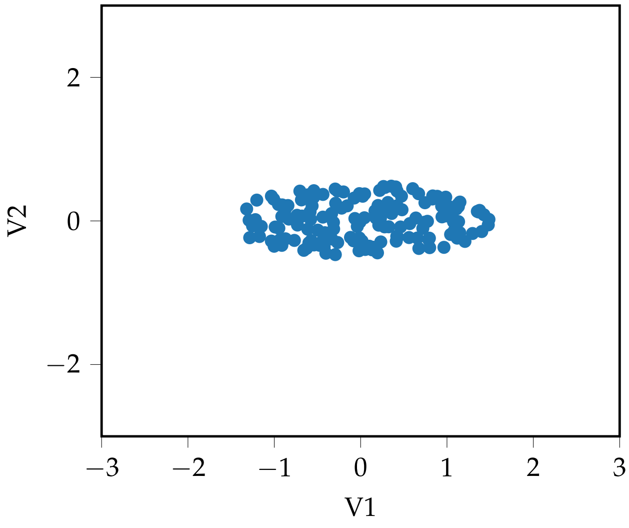


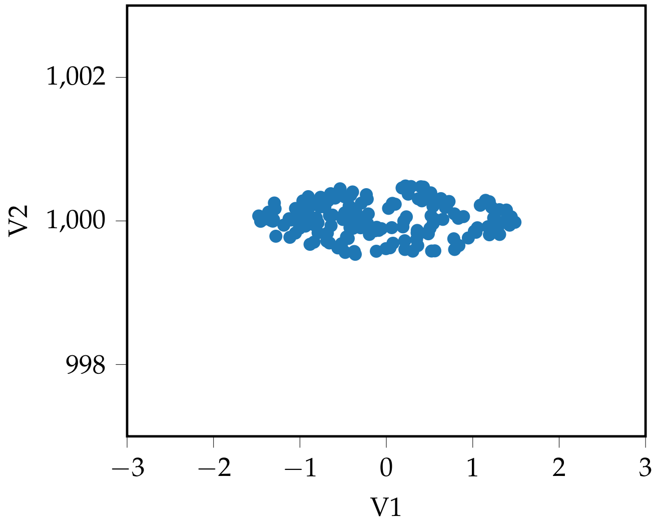
Solution: The plot on the left has maximum spread along the direction spanned by $\frac{1}{\sqrt{2}}\vec{v}_1 + \frac{1}{\sqrt{2}}\vec{v}_2$, which is not the same as $\vec{v}_1$. In other words,

$$\sum_{i=1}^{n} \langle \vec{x}_i, \vec{v}_1 \rangle^2 \neq \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^d} \sum_{i=1}^{n} \langle \vec{x}_i, \vec{u} \rangle^2 \tag{11}$$

Hence, $\vec{v}_1$ is not the first principal component and the plot is invalid. For the second plot, the sum of squares of x-coordinates appears to be the same as the sum of squares along any other axis, so it is entirely plausible for $\vec{v}_1$ to be the first principal component.

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





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

Solution: The plot on the left is certainly valid, as we have discussed earlier. Note that the plot on the right is centered at $(0, 1000)$. The sum of squares of y-coordinates is larger than the sum of squares of x-coordinates, since the y coordinates appear to be between 999 and 1001. Hence, it is not possible for $\vec{v}_1$ to be the first principal component.

(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?**

Solution: Notice that the two plots in part 1.d are the same, except that the one on the right is centered at $(0, 1000)$. This means that we cannot claim that the spread of the data is related to the sum of squares along any axis. Hence, we need the data to be centered at $(0, 0)$ for us to make such a claim. The mathematical reasoning behind this has connections in statistics, so it is out of scope for this course.