Discussion 14B

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

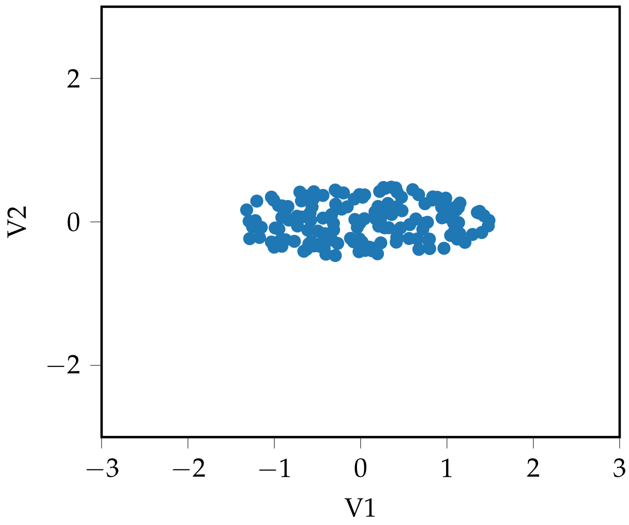
$$\underset{\vec{u} \in \mathbb{R}^{d}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \vec{x}_i - \langle \vec{x}_i,\, \vec{u} \rangle \, \vec{u} \right\|^2 \tag{2}$$

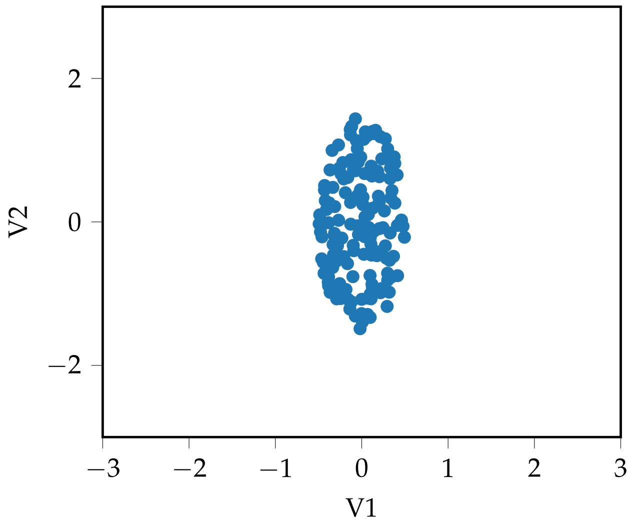
where $\|\vec{u}\| = 1$. **Show that finding the top principal component is equivalent to maximizing**

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

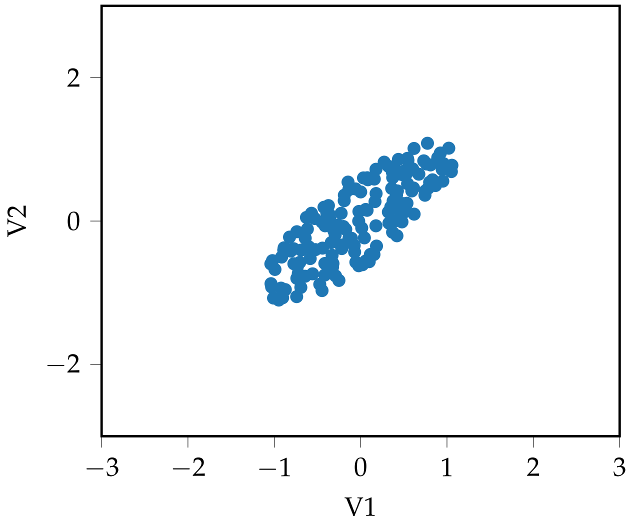
In each plot below, the data is projected onto two unit vectors. The x coordinate is the projection onto the first vector (written as "V1" or $\vec{v}_1$), and the y coordinate is the projection onto the second vector (written as "V2" or $\vec{v}_2$). We say that a plot is "valid" if the first vector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

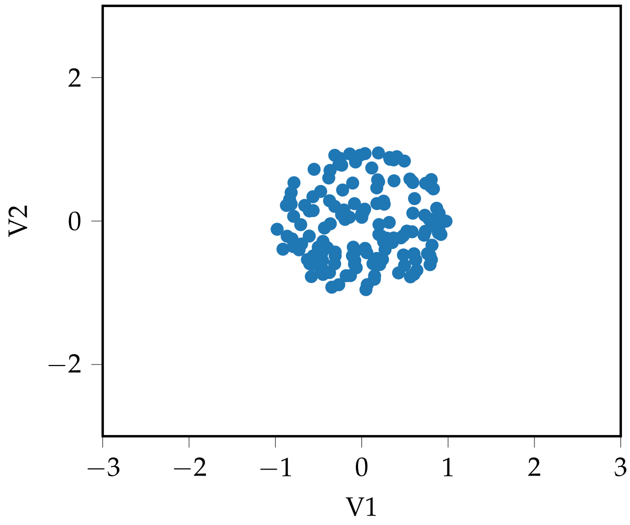
(b) **Which of these two plots is valid?**





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





2. I bet Cal will win this year (PCA Problem from Spring 2021 Final)

As huge fans of the Big Game, you and your friend want to bet on whether Cal or Stanford will win this year. You want to predict this year's result by analyzing historical records. Therefore, you decide to model this as a binary classification problem and do PCA for dimension reduction on the data you collected. The "$+1$" class represents victories of Cal and "-1" represents victories of Stanford.

After some research, you obtained a data matrix $A \in \mathbb{R}^{d\times n}$,

$$A = \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_n \end{bmatrix} \tag{4}$$

where each of the n columns $\vec{x}_i$ denotes a game and each of the d rows of A contains information of a possibly relevant factor of the games (weather, location, date, air quality, etc).

(a) Let the full SVD of $A = U\Sigma V^\top$, where A is given in eq. (4). You project your data along $\vec{u}_1$ and $\vec{u}_2$ (the first two principal components), and for comparison you also project your data along two randomly chosen directions $\vec{w}_1$ and $\vec{w}_2$ as well. You get the two pictures in Figure 3, but you forgot to label the axes. Of the two figures below, which one is the projection onto the principal components and which one is the projection onto the random directions? **Match axes (i), (ii), (iii), (iv) to $\vec{w}_1, \vec{w}_2, \vec{u}_1$, and $\vec{u}_2$, and justify your answer.**

Note that there may be multiple correct matchings; you only need to find and justify one of them.

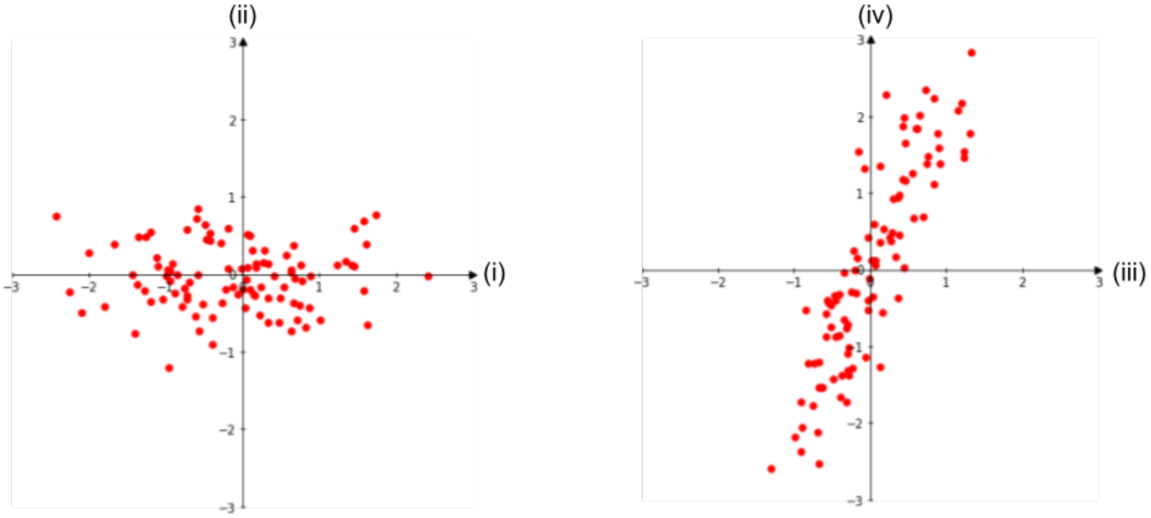


Figure 3: Projected datasets.

(b) In order to reduce the dimension of the data, we would like to project the data onto the first k principal components of A, where k is less than the original data dimension d. **Show how to find the new vector $\vec{z}_i \in \mathbb{R}^k$ which is the k-dimesional, compressed version of $\vec{x}_i$.** You may use the SVD of A.

(c) Given a new set of projection coefficients denoted $\vec{z}_{\text{new}} \in \mathbb{R}^k$, we can define a classifier that will predict $+1$ (i.e., that Cal wins) if $\vec{w}_\star^\top \vec{z}_{\text{new}} > 0$ and -1 (i.e., that Stanford wins) otherwise.

Assume $d = 6$, $k = 4$, and $\vec{w}_\star = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}^\top$. Let $A = U\Sigma V^\top$ for A defined in eq. (4), and you find that U is given by the identity matrix, i.e. $U = I_d$. Now suppose the data point for this year's big game $\vec{x}_{2021} = \begin{bmatrix} 3 & 6 & 4 & 1 & 9 & 6 \end{bmatrix}^\top$. **Would you bet on Cal or Stanford to win? Justify your answer.**

(HINT: Don't forget to project your data onto the principal components.)