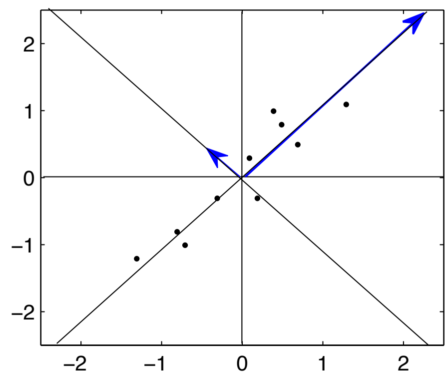
1 Principal components and least squares fitting



Suppose that $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and $\mathbf{y} = (y_1, y_2, \dots, y_p)$ are the x and y coordinates of some data. Suppose that the data is centered, so that $\bar{x} = \frac{1}{p}\sum_{j=1}^{p} x_j = 0$ and $\bar{y} = \frac{1}{p}\sum_{j=1}^{p} y_j = 0$.

- Show that the square of the distance between a point (x_j, y_j) and a fixed line $y = ax$ is

  $$d^2 = \left(\frac{1}{1+a^2}\right)^2 (y_j - ax_j)^2.$$

  (Recall that the distance between a point $\mathbf{x}$ and a line L is the shortest distance between $\mathbf{x}$ and any point on L.)

- The sum of squared distances between a fixed line $y = ax$ and the data $(x_1, y_1), \dots, (x_p, y_p)$ is

  $$D(a) = \left(\frac{1}{1+a^2}\right)^2 \sum_{j=1}^{p} (y_j - ax_j)^2.$$

- Argue that if $\mathbf{v}_1 = (1, u)$ and $\mathbf{v}_2 = (1, v)$ are distinct principal components of the covariance matrix

  $$\mathcal{C} = \begin{pmatrix} cov(\mathbf{x}, \mathbf{x}) & cov(\mathbf{x}, \mathbf{y}) \\ cov(\mathbf{x}, \mathbf{y}) & cov(\mathbf{y}, \mathbf{y}) \end{pmatrix},$$

  then u and v are the critical points of D (that is, $D'(v) = D'(u) = 0$).

- Which critical point corresponds to a least squares distance? Which corresponds to a maximal least squares distance? Why does this make sense geometrically?