Question 0

First consider the polynomial transformations (A) - (D):

- We can easily see the separation is not linear, ruling out (A) and (B).
- We can see a circular pattern in the data, suggesting we need a quadratic term.
- A cubic term would not be necessary and would lead to a less generalisable model, as well as a higher computational cost.

Hence out of the best polynomial transform would be option (B).

While (E) might also provide a good fit, the computational cost of finding estimates for 50 parameters means that option (B) is a far better all-around solution.

Question 1

1.1 We expect the vector $\mathbf{w}_{FDA}$ to be pointing in a direction approximately perpendicular to the decision boundary, as the embedded points from each class would generally fall on each side of the decision boundary, creating a good separation of the embedded points. This means we can rule out (a) and (b), and so (c) or (d) are likely, as the vector $\mathbf{w}_{FDA}$ is unique up to reversal, so (c) and (d) are equally good vectors to embed our points onto.

1.2 Assuming $\mathbf{x}_i \sim_{IID} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, a dataset D has likelihood:

$$p(D|\boldsymbol{\mu}, \Sigma) = \prod_{i \in D} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

We can solve this to get the maximumum likelihood estimate Σ_{ML}, which is symmetric and positive definite. Hence we get the eigen decomposition:

$$\Sigma_{ML} = [\mathbf{u}_1, \mathbf{u}_2] \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} [\mathbf{u}_1, \mathbf{u}_2]^T$$

where $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$ are egeinvectors of Σ_{ML} and $D_1 > D_2$ are the corresponding eigenvalues.

In the covariance matrix, the eigenvectors represent the directions of largest variance and the corresponding eigenvalues give the magnitude of this variance. Since we know $D_1 > D_2$ we know $\mathbf{u}_1$ will point in the direction of greatest variance, which we can see from the plot is likely to be either (a) or (b).

Question 3

We can rewrite the classic soft-margin SVM problem to incorporate a loss function

$$\min_{\mathbf{w}'} \|\mathbf{w}'\| + \sum_{i \in D} L(y_i f(\mathbf{x}_i))$$

where

$$L(t) = \begin{cases} 0 & t \geq 1 \\ 1 - t & t < 1 \end{cases}$$

Hence to adapt this to our context, we need to create a loss function which penalises a false negative 1000 times more than a false positive. We define the function

$$L'(y, f) = \begin{cases} 0 & yf \geq 1 \\ 1 - yf & yf < 1, y = -1 \\ 1000(1 - yf) & yf < 1, y = +1 \end{cases}$$

and so with this loss function, a false negative has a loss 1000 times greater than the loss of a false positive. Thus our new objective function is

$$\min_{\mathbf{w}'} \|\mathbf{w}'\| + \lambda \sum_{i \in D} L'(y_i, f(\mathbf{x}_i))$$

where λ is a scaling term that allows us to determine how important the loss function should be in our decision.

Question 4

4.1 The adjacency matrix of P indicates the non-zero entries of the inverse covariance matrix of P. Hence we expect P to appear as the inverse of a sparse matrix, which is satisfied by either (B) or (C). However cov$P$ (and by extension $(\text{cov}P)^{-1}$) must be symmetric, and so we choose (B).

4.2 The joint probability is

$$p(y, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) = p(y)p(x^{(1)}|y)p(x^{(2)}|x^{(1)}, y)p(x^{(3)}|x^{(2)})p(x^{(4)}|x^{(1)})$$

We can also list the conditional independences:

$$x^{(4)} \perp x^{(3)} \mid x^{(1)}, x^{(2)} \qquad x^{(4)} \perp x^{(2)} \mid x^{(1)} \qquad x^{(4)} \perp y \mid x^{(1)}$$
$$x^{(3)} \perp x^{(1)} \mid x^{(2)} \qquad x^{(3)} \perp y \mid x^{(2)}$$

Hence we conclude we only need to include $x^{(1)}$ and $x^{(2)}$ in our analysis. The independence of $x^{(3)}, x^{(4)}$ and y conditional on $x^{(1)}$ and $x^{(2)}$ means $x^{(3)}$ and $x^{(4)}$ do not provided any further infomation on y.