Solutions to Take-Home Quiz One
ECE 271B - Winter 2022

Department of Electrical and Computer Engineering
University of California, San Diego

Problem 1.

a) The distance from a point $\mathbf{x}$ to the origin is $||\mathbf{x}||$. Hence, the problem reduces to

$$\min_{\mathbf{x}} ||\mathbf{x}||^2 \quad \text{such that} \quad \mathbf{w}^T\mathbf{x} + b = 0.$$

The Lagrangian is

$$L(\mathbf{x}, \lambda) = ||\mathbf{x}||^2 - \lambda(\mathbf{w}^T\mathbf{x} + b)$$

and the optimality conditions are

$$\begin{aligned} \nabla_{\mathbf{x}} L &= 2\mathbf{x} - \lambda\mathbf{w} = 0 \\ \nabla_{\lambda} L &= \mathbf{w}^T\mathbf{x} + b = 0. \end{aligned}$$

From the first, it follows that, for any $\mathbf{w} \neq 0$,

$$\lambda = 2\left(\frac{\mathbf{w}}{||\mathbf{w}||^2}\right)^T \mathbf{x}$$

and, using the second,

$$\lambda = -2\,\frac{b}{||\mathbf{w}||^2}.$$

Plugging in on the first, we finally get

$$2\mathbf{x} + 2\,\frac{b\mathbf{w}}{||\mathbf{w}||^2} = 0$$

or

$$\mathbf{x}_0 = -b\,\frac{\mathbf{w}}{||\mathbf{w}||^2}.$$

We next consider the Hessian of $L(\mathbf{x}, \lambda)$. This is the matrix

$$\nabla_{\mathbf{xx}} L = 2\mathbf{I},$$

which is clearly positive definite. Hence, we have a minimum.

b) When $||\mathbf{w}|| = 1$, it follows that $||\mathbf{x}_0|| = |b|$. This means that $|b|$ is the minimum distance of the plane to the origin. Furthermore,

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = \mathbf{w}^T\mathbf{x} + b = 0.$$

This implies that, for any point $\mathbf{x}$ on the plane, $\mathbf{x} - \mathbf{x}_0$ is orthogonal to $\mathbf{w}$. But $\delta(\mathbf{x}) = \mathbf{x} - \mathbf{x}_0$ is a vector that lies entirely on the plane. Hence, $\mathbf{w}$ is orthogonal to any vector that has origin in $\mathbf{x}_0$ and lies entirely on the plane. Since any point $\mathbf{x}$ can be reached in this way, this implies that $\mathbf{w}$ is orthogonal to all the vectors that lie on the plane, and is thus orthogonal to the plane itself. For this reason, $\mathbf{w}$ is known as the *normal vector* of the plane.

Problem 2. In this case, the optimization problem is

$$\min_{p(x)} \int p(x) \log p(x) dx$$

with constraints

$$\begin{aligned} \int x p(x) dx &= \mu \\ \int (x-\mu)^2 p(x) dx &= \sigma^2 \\ \int p(x) dx &= 1. \end{aligned}$$

This has Lagrangian

$$L = \int p(x) \log p(x) dx - \lambda_1 \left(\int x p(x) dx - \mu \right) - \lambda_2 \left(\int (x-\mu)^2 p(x) dx - \sigma^2 \right) - \lambda_3 \left(\int p(x) dx - 1 \right).$$

Setting derivatives to zero

$$\begin{aligned} \frac{\partial L}{\partial p(v)} &= \log p(v) + 1 - \lambda_1 v - \lambda_2 (v-\mu)^2 - \lambda_3 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= \int x p(x) dx - \mu = 0 \\ \frac{\partial L}{\partial \lambda_2} &= \int (x-\mu)^2 p(x) dx - \sigma^2 = 0 \\ \frac{\partial L}{\partial \lambda_3} &= \int p(x) dx - 1 = 0 \end{aligned}$$

or

$$p(v) = \exp\{\lambda_1 v + \lambda_2 (v-\mu)^2 + \lambda_3 - 1\} \quad (1)$$

$$\mu = \int x p(x) dx \quad (2)$$

$$\sigma^2 = \int (x-\mu)^2 p(x) dx \quad (3)$$

$$1 = \int p(x) dx. \quad (4)$$

From (1), we see that $p(x)$ has the form

$$\begin{aligned} p(x) &= \exp\left\{ \lambda_2 \left[x^2 - 2\left(\mu - \frac{\lambda_1}{2\lambda_2} \right) x + \mu^2 \right] + \lambda_3 - 1 \right\} \\ &= \exp\left\{ \lambda_2 \left[x^2 - 2\left(\mu - \frac{\lambda_1}{2\lambda_2} \right) x + \left(\mu - \frac{\lambda_1}{2\lambda_2} \right)^2 - \left(\mu - \frac{\lambda_1}{2\lambda_2} \right)^2 + \mu^2 \right] + \lambda_3 - 1 \right\} \\ &= K \exp\left\{ \lambda_2 \left[x - \left(\mu - \frac{\lambda_1}{2\lambda_2} \right) \right]^2 \right\}, \end{aligned}$$

where

$$K = \exp\left\{ -\lambda_2 \left(\mu - \frac{\lambda_1}{2\lambda_2} \right)^2 + \lambda_2 \mu^2 + \lambda_3 - 1 \right\}.$$

It follows that $p(x)$ is a Gaussian of mean ξ and variance τ, with

$$\begin{aligned}\xi &= \mu - \frac{\lambda_1}{2\lambda_2} \\ \tau &= -\frac{1}{2\lambda_2}.\end{aligned}$$

But from the constraints, we know that $\xi = \mu$ and $\tau = \sigma^2$, from which it follows that

$$\begin{aligned}\lambda_1 &= 0 \\ \lambda_2 &= -\frac{1}{2\sigma^2}\end{aligned}$$

It follows that

$$\lambda_3 = 1 + \log K = 1 - \log\sqrt{2\pi\sigma^2}.$$

Problem 3.

a) We have seen in class that the principal components of $\mathbf{X}$ are the eigenvectors of the covariance matrix

$$\mathbf{\Sigma}_x = E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T].$$

If $\mathbf{Z} = \mathbf{X} - \boldsymbol{\mu}_x$, then clearly $\boldsymbol{\mu}_z = E[\mathbf{Z}] = \mathbf{0}$ and

$$\begin{aligned}\mathbf{\Sigma}_z &= E[(\mathbf{Z} - \boldsymbol{\mu}_z)(\mathbf{Z} - \boldsymbol{\mu}_z)^T] \\ &= E[\mathbf{Z}\mathbf{Z}^T] \\ &= E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T] \\ &= \mathbf{\Sigma}_x.\end{aligned}$$

This is really not surprising since subtracting the mean is equivalent to changing the coordinate system, i.e. shifting the probability mass of $\mathbf{X}$ so that it is centered at the origin. Hence, it does not change the covariance structure of the data.

b) If $\mathbf{X}$ is zero mean then

$$\mathbf{\Sigma}_x = E[\mathbf{X}\mathbf{X}^T],$$

and, when $\rho_{i,j} = 1, \forall i, j$,

$$\begin{aligned}(\mathbf{\Sigma}_x)_{i,j} &= E[X_i X_j] \\ &= \sqrt{E[X_i^2]E[X_j^2]} \\ &= \sqrt{\sigma^2\sigma^2} \\ &= \sigma^2,\end{aligned}$$

where we have also used the fact that the X_i are identically distributed. Hence

$$\mathbf{\Sigma}_x = \mathbf{1}d\sigma^2\mathbf{1}^T,$$

which is equivalent to

$$\boldsymbol{\Sigma}_x = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T,$$

where $\boldsymbol{\Lambda} = diag(d\sigma^2, 0, \ldots, 0)$ and $\boldsymbol{\Phi}$ is an orthonormal matrix whose first column is the vector $\mathbf{1}$. It follows that $\mathbf{1}$ is an eigenvector of $\boldsymbol{\Sigma}_x$.

c) Since the eigenvector matrix $\boldsymbol{\Phi}$ is orthonormal,

$$\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j = 0, \, \forall i \neq j.$$

Hence, for all $j \neq 1$,

$$\begin{aligned} \boldsymbol{\phi}_j^T \boldsymbol{\phi}_1 &= \boldsymbol{\phi}_j^T \mathbf{1} \\ &= \frac{1}{\sqrt{d}} \sum_k (\boldsymbol{\phi}_j)_k = 0. \end{aligned}$$

It follows that

$$\begin{aligned} E[z_i] &= E[\boldsymbol{\phi}_i^T \mathbf{x}] \\ &= E\left[\sum_k (\boldsymbol{\phi}_j)_k x_k\right] \\ &= \sum_k (\boldsymbol{\phi}_j)_k E[x_k] \\ &= E[\mathbf{x}] \sum_k (\boldsymbol{\phi}_j)_k \\ &= 0, \end{aligned}$$

where we have, once again, used the fact that the X_i are identically distributed. This shows that all coefficients, other than the DC, must have zero mean.

Problem 4.

a) To solve this problem in the most general form, we consider that $P_Y(1) = p$ and $P_Y(2) = 1 - p$. We then note that, for any function $f(\mathbf{x})$,

$$\begin{aligned} E_{\mathbf{X}}[f(\mathbf{x})] &= \int f(\mathbf{x})\, p\, P_{\mathbf{X}|Y}(\mathbf{x}|1) d\mathbf{x} + \int f(\mathbf{x})\, (1-p)\, P_{\mathbf{X}|Y}(\mathbf{x}|2) d\mathbf{x} \\ &= p E_{\mathbf{X}|Y}[f(\mathbf{x})|Y=1] + (1-p) E_{\mathbf{X}|Y}[f(\mathbf{x})|Y=2]. \end{aligned}$$

To compute the mean of $\mathbf{X}$, it suffices to apply this result with $f(\mathbf{x}) = \mathbf{x}$, which leads to

$$\boldsymbol{\mu}_x = p\boldsymbol{\mu}_1 + (1-p)\boldsymbol{\mu}_2.$$

For the covariance, we use $f(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T$ leading to

$$\begin{aligned} \boldsymbol{\Sigma}_x &= E_{\mathbf{X}}[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T] \\ &= p\, E_{\mathbf{X}|Y}[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T | Y=1] + (1-p)\, E_{\mathbf{X}|Y}[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T | Y=2]. \end{aligned}$$

We next note that

$$
\begin{aligned}
&E_{\mathbf{X}|Y}[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T|Y=i] = E_{\mathbf{X}|Y}[(\mathbf{x}-\boldsymbol{\mu}_i+\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_i+\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)^T|Y=i] \\
&= \boldsymbol{\Sigma}_i + E_{\mathbf{X}|Y}[(\mathbf{x}-\boldsymbol{\mu}_i)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)^T|Y=i] + E_{\mathbf{X}|Y}[(\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_i)^T|Y=i] + (\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)^T \\
&= \boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_x)^T,
\end{aligned}
$$

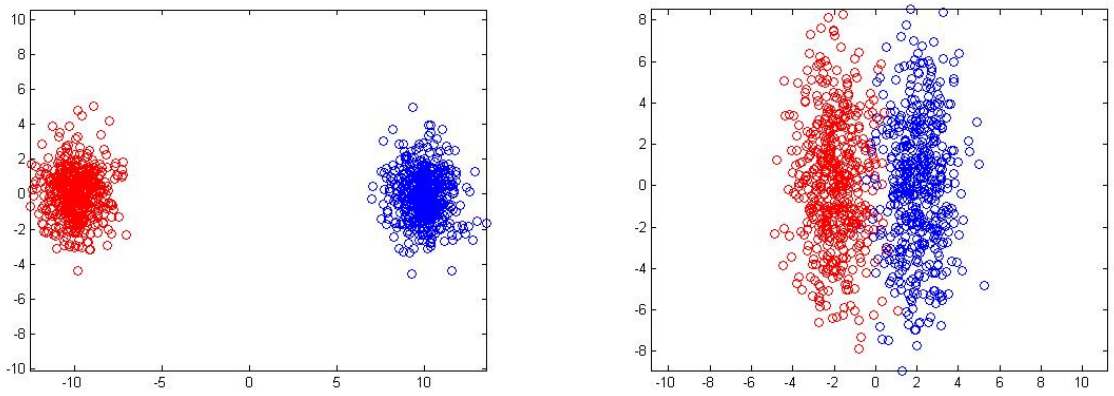
from which it follows that

$$
\begin{aligned}
\boldsymbol{\Sigma}_x &= p\boldsymbol{\Sigma}_1 + (1-p)\boldsymbol{\Sigma}_2 + p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)^T + (1-p)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)^T \\
&= p\boldsymbol{\Sigma}_1 + (1-p)\boldsymbol{\Sigma}_2 + p(1-p)^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T + (1-p)p^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \\
&= p\boldsymbol{\Sigma}_1 + (1-p)\boldsymbol{\Sigma}_2 + [p(1-p)^2 + (1-p)p^2](\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \\
&= p\boldsymbol{\Sigma}_1 + (1-p)\boldsymbol{\Sigma}_2 + p(1-p)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T.
\end{aligned}
$$

Hence, when $p = 1/2$,

$$
\begin{aligned}
\boldsymbol{\mu}_x &= E[\mathbf{X}] = \frac{1}{2}[\boldsymbol{\mu}_1+\boldsymbol{\mu}_2] \\
\boldsymbol{\Sigma}_x &= E[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T] = \frac{1}{2}\left[\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right] + \frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T.
\end{aligned}
$$

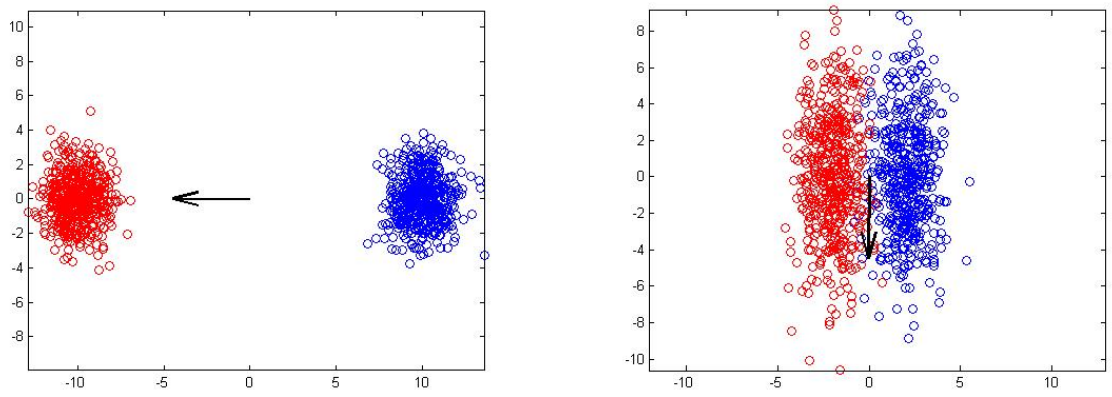
b) The two plots are shown below. Points from class 1 are shown in blue and points from class 2 in red.



(a) (b)

Figure 1: (a) Condition A and (b) condition B.

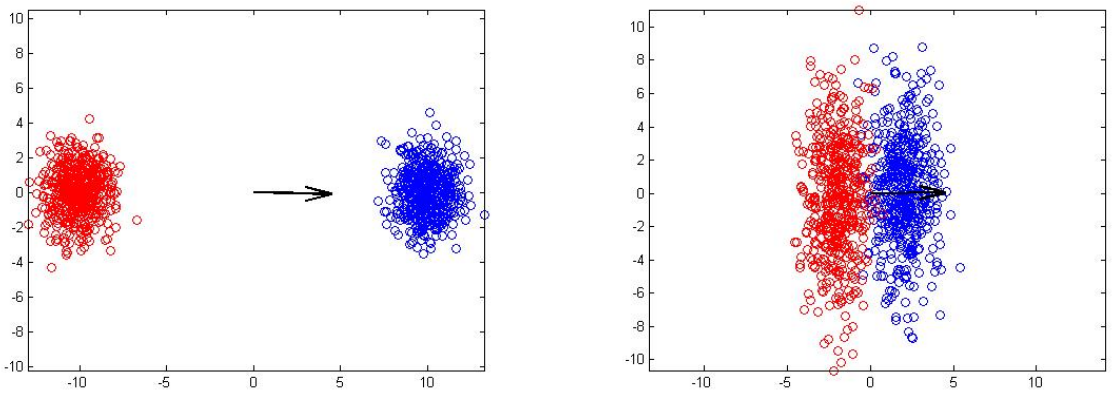
c) The PCA direction switches from the direction of the horizontal axis, under condition A, to the direction of the vertical axis, under condition B.



(a) (b)

Figure 2: (a) Condition A and (b) condition B.

d) The LDA direction is close to the direction of the horizontal axis under the two conditions.



a b

Figure 3: (a) Condition A and (b) condition B.

e) PCA is not always a good approach to reduce dimensionality. The direction of largest variance is not always discriminant. Note that, under condition B, PCA chooses the vertical dimension, along which the classes have similar projection. This is the worse possible direction for discrimination between the two classes. On the other hand, LDA always seems to select the most discriminant direction. Hence, LDA is a better dimensionality reduction approach for classification problems.

f) Regarding PCA, we start by computing the mean and covariance of $\mathbf{X}$. Using the results of **a)**, we

have $\boldsymbol{\mu}_x = 0$ and

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 1+\alpha^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}.$$

Since this is a diagonal matrix, its eigenvectors are just the coordinate axis. Thus, PCA will choose the eigenvector $\mathbf{e}_1$, when $1+\alpha^2 > \sigma^2$, and the eigenvector $\mathbf{e}_2 = (0,1)^T$, otherwise. Hence, we have two possible decision rules.

i) When $\alpha > \sqrt{\sigma^2 - 1}$, we have $z = \mathbf{e}_1^T \mathbf{x} = x_1$ and

$$\begin{aligned} P_{Z|Y}(z|1) &= P_{X_1|Y}(z|1) = \mathcal{G}(z, \alpha, 1) \\ P_{Z|Y}(z|2) &= P_{X_1|Y}(z|2) = \mathcal{G}(z, -\alpha, 1). \end{aligned}$$

This is the direction along which the Gaussians have greatest separation. Hence, the classifier obtained with PCA is optimal.

ii) When $\alpha < \sqrt{\sigma^2 - 1}$, we have $z = \mathbf{e}_2^T \mathbf{x} = x_2$ and

$$\begin{aligned} P_{Z|Y}(z|1) &= P_{X_2|Y}(z|1) = \mathcal{G}(z, 0, \sigma^2) \\ P_{Z|Y}(z|2) &= P_{X_2|Y}(z|2) = \mathcal{G}(z, 0, \sigma^2). \end{aligned}$$

This is a terrible selection of direction to project on since both classes have the same marginal along this projection. Therefore, the classification error is going to be the worst possible (0.5, i.e. the same as random guessing). In this case, PCA fails miserably.

Regrading LDA, since the covariance is diagonal, the linear discriminant $\boldsymbol{\phi}'$ is always in the direction of the line $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ that connects the two means. Since this is always the vector $(2\alpha, 0)$, the discriminant direction is always $\mathbf{e}_1$ and $y = \boldsymbol{\phi}'\mathbf{x} = x_1$. Hence, LDA always chooses the right direction to project the data on.

Overall, the conclusion is that one has to be very careful with the use of PCA for classification. There are situations in which it is optimal, but others in which it leads to the worst possible choice of features. Furthermore, a slight perturbation of the covariance matrix can lead to the switch from one situation to the other. Hence, in the context of classification problems, PCA is not necessarily an optimal technique for dimensionality reduction. LDA is optimal for Gaussian classes. It can also be sub-optimal for non-Gaussian data, but this is beyond the scope of the problem.

Problem 5.

a) The 16 principal components of the face dataset are shown in the figure below. Note that, because the faces are not precisely aligned, these principal components do not show as much face details as those presented in the lecture. Nevertheless, one can see a face-like structure in all of them.

Figure 4: 16 principal components of largest variance in the face dataset.

b) The 15 linear discriminants of the face dataset are shown in the figure below. Note that because the discriminants capture information about how the classes differ, they tend not to have such a clear face shape. Instead, they emphasize the areas of detail that differ from person to person.

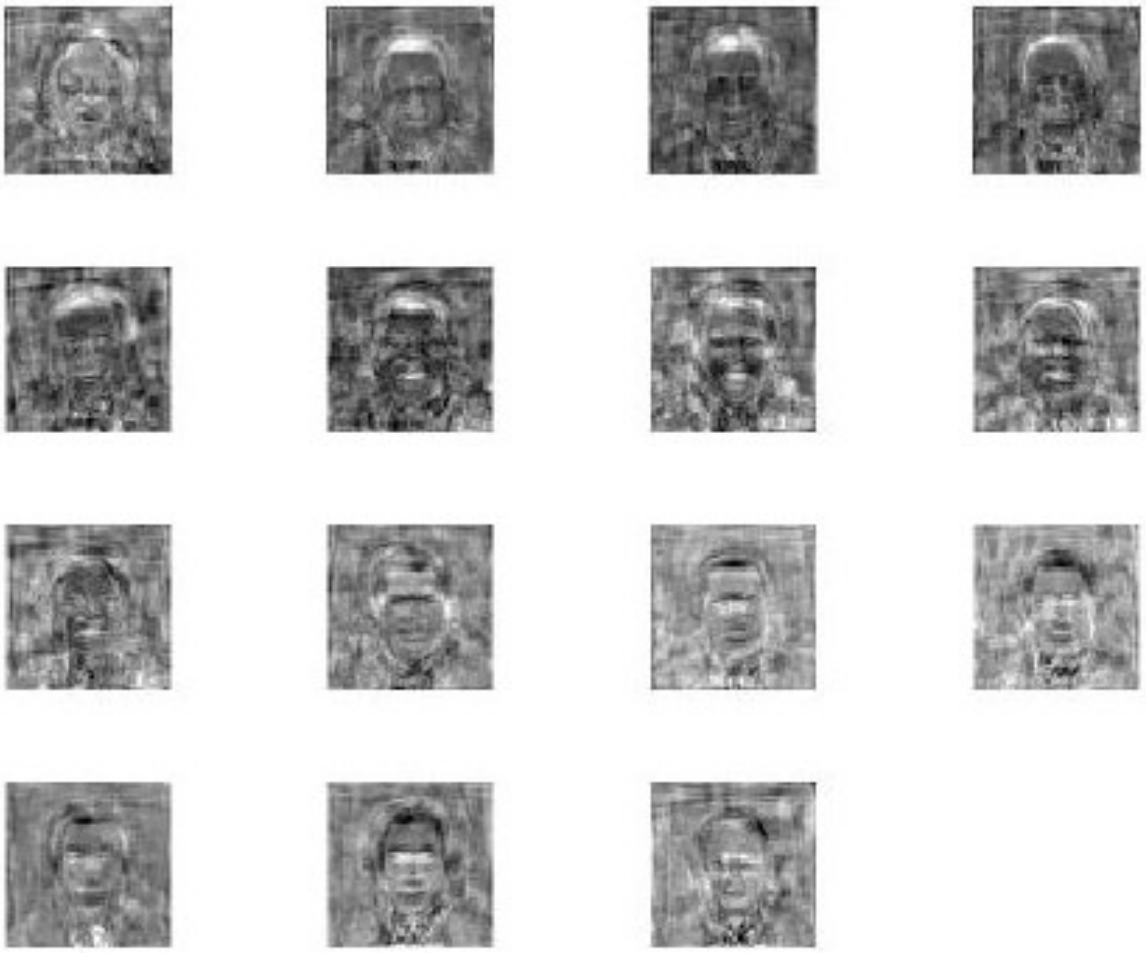


Figure 5: 15 linear discriminants of the face dataset.

c) The table below shows the error rates produced by the Gaussian classifier on the PCA space.

	person 1	person 2	person 3	person 4	person 5	person 6	overall
number of errors	2	2	5	5	1	6	21
probability of error	20%	20%	50%	50%	10%	60%	35%

d) The table below shows the error rates produced by the Gaussian classifier on the RDA space.

	person 1	person 2	person 3	person 4	person 5	person 6	overall
number of errors	1	2	3	2	0	3	11
probability of error	10%	20%	30%	20%	0%	30%	18.3%

e) The table below shows the error rates produced by the Gaussian classifier on the PCA+LDA space.

	person1	person2	person3	person4	person5	person6	overall
mismatched number	3	2	3	2	1	3	14
error of probability	30%	20%	30%	20%	10%	30%	23.3%