1) (a) (i) The inverse of a symmetric matrix is itself symmetric.
True.

We want to show $A^{-1} = A^{-T}$ given $A = A^T$

Proof: $I = AA^{-1} = A^T A^{-1}$

multiply A^{-T} on both sides (left multiply)

$$A^{-T} = A^{-T} A^T A^{-1}$$

$$A^{-T} = A^{-1} \quad \#$$

(ii) All 2x2 orthogonal matrices have the following form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \text{ or } \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$ True

let $P \in \mathbb{R}^{2\times2}$, $P = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

if P is orthogonal, then $P^T P = I$

$$\Rightarrow \begin{pmatrix} a^2+c^2 & ab+cd \\ ab+cd & b^2+d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Rightarrow a^2+c^2 = 1, \; b^2+d^2 = 1, \; ab+cd = 0$$

Without loss of generality, let $a = \cos\theta$, $c = \sin\theta$ and $b = -c$, $a = d$

$b = \sin\theta$ $d = \cos\theta$

Which are $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ or $\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$ or $b = c$, $a = -d$

(iii) Assume $A = CC^T$, then we have

$x^T A x = x^T C \cdot C^T x = \|C^T x\|_2^2 \geq 0 \Rightarrow A$ is positive semi-definite.

Then $A = U \Sigma U^T$ with $\Sigma_{ii} \geq 0$ which means all eigenvalues of A is non-negative.

Also $tr(A) = \sum_i \lambda_i$ λ_i $i=1,2,\cdots$ are the eigenvalues.

$tr(A) = -8-5-2 = -15 \Rightarrow$ Not all eigenvalues are non-negative.

Therefore, A cannot be written as $A = CC^T$

2. (a) (i) $E_Y[E_X(X|Y)] = E[X]$

proof: $E_Y[E_X(Y|X)] = \int_{R_y} E[X|y]\,P(y)\,dy = \int_{R_y}\int_{R_x} x\,P(x|y)\,dx\;P(y)\,dy$

$$= \int_{R_x}\int_{R_y} x\,P(x,y)\,dy\,dx$$

$$= \int_{R_x} x\,P(x)\,dx = E[X]$$

R_x, R_y are the region of X and Y respectively.

(ii) $E[I[X\in C]] = P(X\in C)$

Proof: $E[I[X\in C]] = \int_R P(x)\,I[x\in C]\,dx = \int_C P(x)\,dx = P(X\in C)$

(iii) $\mathrm{var}[X] = E_Y[\mathrm{var}_X[X|Y]] + \mathrm{var}_Y[E_X[X|Y]]$

Proof: $\mathrm{var}_X[X|Y] = E_X[(X-E[X|Y])^2|Y]$

$$E[\mathrm{var}(X|Y)] = E\left[E[(X-E[X|Y])^2|Y]\right]$$

$$= E[(X-E[X|Y])^2] \qquad ①$$

$$\mathrm{var}(E_X[X|Y]) = E[(E[X|Y]-E(E[X|Y]))^2]$$

$$= E[(E[X|Y]-E[X])^2] \qquad ②$$

$$\mathrm{var}(X) = E[(X-E[X])^2] = E[(X-E[X|Y]+E[X|Y]-E[X])^2]$$

$$= E[(X-E[X|Y])^2] + E[(E[X|Y]-E[X])^2]$$

$$+ 2E[(X-E[X|Y])\cdot(E[X|Y]-E[X])] \qquad ③$$

$$E[(X-E[X|Y])\cdot(E[X|Y]-E[X])]$$

$$=E[X\cdot E[X|Y]-X\cdot E[X]-E[X|Y]\cdot E[X|Y]+E[X]\cdot E[X|Y]]$$

$$=E[X\cdot E[X|Y]]-E^2[X]-E[E^2[X|Y]]+E[X]\cdot E(E[X|Y])$$

$$=E\{E[X\cdot E[X|Y]\,|\,Y]\}-E[E^2[X|Y]]-E^2[X]+E^2[X]$$

$$=E\{E[X|Y]\cdot E[X|Y]\}-E[E^2(X|Y)]=0 \qquad (4)$$

Combining (1) (2) (3) (4), we get

$$\mathrm{var}(X)=E[\mathrm{var}(X|Y)]+\mathrm{var}(E[X|Y])$$

(iv) $E[XY]=\iint_{-\infty}^{\infty} xy\, P_{XY}(x,y)\,dx\,dy$ ~~if X, Y~~

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, P_X(x)P_Y(y)\,dx\,dy \quad \text{by independence}$$

$$=\int_{-\infty}^{\infty} x P_X(x)\,dx\cdot\int_{-\infty}^{\infty} y P_Y(y)\,dy$$

$$=E[X]\cdot E[Y]$$

(v) $E[X]=0\cdot P(X=0)+1\cdot P(X=1)=P(X=1)$

$E[Y]=P(Y=1)$

$E[XY]=\sum_{x=0}^{1}\sum_{y=0}^{1} xy\, P(X=x,Y=y)$

$=P(X=1,Y=1)$

$E[XY]=E[X]\cdot E[Y]$

$\Rightarrow P(X=1,Y=1)=P(X=1)\cdot P(Y=1)$

$\Rightarrow X, Y$ are independent.

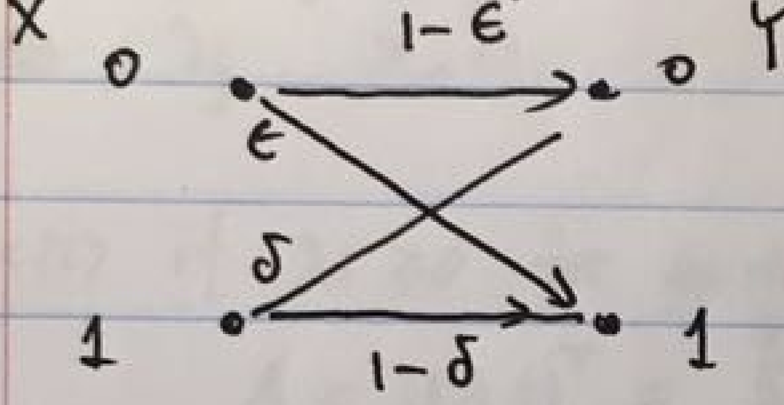
(b) (i) $P(H=h, D=d) \leq P(H=h)$

$$P(H=h) = \sum_{x} P(H=h, D=x) > P(H=h, D=d)$$

(ii) $P(H=h \mid D=d)$ depends $P(H=h)$

$$P(H=h) = \sum_{x} P(H=h \mid D=x) P(D=x)$$

For example, consider Binary Channel Model in Communication system



$P(Y=0 \mid X=0) = 1-\epsilon$

$P(Y=1 \mid X=0) = \epsilon$

$P(Y=0 \mid X=1) = \delta$

$P(Y=1 \mid X=1) = 1-\delta$

$P(X=0) = 1/2$

$P(X=1) = 1/2$

Thus $P(Y=0) = P(Y=0 \mid X=0) P(X=0) + P(Y=0 \mid X=1) P(X=1)$

$= (1-\epsilon) \times 1/2 + \delta \times \frac{1}{2} = \frac{1}{2}(1-\epsilon+\delta)$

$P(Y=0 \mid X=1) = \delta$

The relationship between $P(Y=0)$ and $P(Y=0 \mid X=1)$ depends on the choice of ϵ, δ.

(iii) $P(H=h \mid D=d) \geq P(D=d \mid H=h) P(H=h)$

According Bayes Rule,

$$P(H=h \mid D=d) = \frac{P(D=d \mid H=h) P(H=h)}{P(D=d)}$$

Since $P(D=d) \leq 1$, thus

$$P(H=h \mid D=d) \geq P(D=d \mid H=h) P(H=h)$$

3) (a) Let the jth eigenvector u_j, then

<i> $u_j^T A u_j \geq 0$ since A is PSD.

$\Leftrightarrow u_j^T U \Lambda U^T u_j \geq 0$

$\Leftrightarrow u_j^T \left(\sum_{i=1}^{d} \lambda_i u_i u_i^T \right) u_j \geq 0$

$\Leftrightarrow u_j^T \left(\sum_{i=1}^{d} \lambda_i u_i u_i^T u_j \right) \geq 0$

$\Leftrightarrow u_j^T \lambda_j u_j \geq 0$ since U is orthogonal

$\Leftrightarrow \lambda_j \geq 0$ for all $j = 1, 2, \cdots, d$

<ii> if $\lambda_i \geq 0$ for each i.

$$A = U \Lambda U^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T$$

let $x \in \mathbb{R}^d$

$$x^T A x = x^T \sum_{i=1}^{d} \lambda_i u_i u_i^T x$$

$$= \sum_{i=1}^{d} \lambda_i \| u_i^T x \|_2^2$$

$$\geq 0 \quad \text{for all } x$$

Thus A is PSD.

Combine (i) and (ii), we have A is PSD iff $\lambda_i \geq 0$ for each i.

(b) (i) If A is PD. By the same argument before, we have

$u_j^T A u_j > 0$ $\quad$ u_j is the jth eigenvector

$\Leftrightarrow u_j^T \sum_{i=1}^{d} \lambda_i u_i u_i^T u_j > 0$

$\Leftrightarrow u_j^T \lambda_j u_j > 0$

$\Leftrightarrow \lambda_j > 0$ for all $j = 1, 2, \cdots, d$

(ii) If $\lambda_i > 0$ for each i.

let $x \in \mathbb{R}^d$,

$$x^T A x = x^T \sum_{i=1}^{d} \lambda_i u_i u_i^T x$$

$$= \sum_{i=1}^{d} \lambda_i \|u_i^T x\|_2^2$$

$$> 0 \qquad \text{for all } x \neq 0$$

Thus A is PD.

Combine (i) and (ii) A is PD iff $\lambda_i > 0$ for each i.

4) (a) $f(t\underline{x}+(1-t)y) = a^T(t\cdot\underline{x}+(1-t)\cdot y)+b$

$$= t\cdot a^T\cdot\underline{x} + (1-t)\cdot a^T\cdot y + t\cdot b + (1-t)\cdot b$$

$$= t(a^T\cdot\underline{x}+b) + (1-t)(a^Ty+b)$$

$$= t\,f(\underline{x}) + (1-t)\,f(y) \quad ①$$

Therefore $f(x) = a^T\underline{x}+b$ is convex.

By the same reasoning, we can also have ~~that~~

$$-f(t\underline{x}+(1-t)y) = -t\,f(\underline{x})+(1-t)f(y) \quad ②.$$

implies $-f(x)$ is convex, thus $f(x)$ is concave.

From ① ②, the equality holds, so $f(x)$ is not strictly convex.

(b) Suppose there exists more than one global minimizer x^*, y^* where $x^* \neq y^*$

we have $f(x^*) = f(y^*)$

Since $f(x)$ is strictly convex on dom (f), we have

$$f(t\cdot x^* + (1-t)y^*) < t\,f(x^*) + (1-t)\,f(y^*)$$

$$\Leftrightarrow f(t(x^*-y^*)+y^*) < f(y^*) \quad (\text{or } f(x^*))$$

$\Rightarrow$ there exists $t(x^*-y^*)+y^* \in \text{dom}(f)$ such that

$$f(t(x^*-y^*)+y^*) < f(y^*)$$

which contradicts that y^* is the global minimizer.

Therefore f has at most one global minimizer.

(c) from (a), we have $f(x) = f(y) + \nabla f^T(y)(x-y) + \frac{1}{2}(x-y)^T \nabla^2 f(y+t(x-y))(x-y)$

let $y = x^*$ and $t=0$, we get

$$f(x) = f(x^*) + \nabla f^T(x^*)(x-x^*) + \frac{1}{2}(x-x^*)^T \nabla^2 f(x^*)(x-x^*)$$

$\Rightarrow f(x) - f(x^*) = \frac{1}{2}(x-x^*) \nabla^2 f(x^*)(x-x^*)$ since $\nabla f(x^*) = 0$

$\Rightarrow (x-x^*) \nabla^2 f(x^*)(x-x^*) \geq 0$ since x^* is local minimum.

$\Rightarrow \nabla^2 f(x^*) \succeq 0$

(d) From (a), we get

$f(x) = f(y) + \nabla f^T(y)(x-y) + \frac{1}{2}(x-y)^T \cdot \nabla^2 f(y+t(x-y)) \cdot (x-y)$ for all $x, y \in \mathbb{R}^d$

① Since f is convex, ~~we~~ have

$f(x) \geq f(y) + \nabla f^T(y)(x-y)$ 1st order condition.

Therefore, $\frac{1}{2}(x-y)^T \cdot \nabla^2 f(y+t(x-y))(x-y)$

$= \frac{1}{2}(x-y)^T \cdot \nabla^2 f(tx+(1-t)y)(x-y) \geq 0$

let $t=1$

$\Rightarrow \nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$

② If $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$, then

$a^T \cdot \nabla^2 f(x) \cdot a \geq 0$

$\Rightarrow a^T \cdot \nabla^2 f(y+t(x-y)) \cdot a \geq 0$ since $y+t(x-y) \in \mathbb{R}^d$

$\Rightarrow (x-y)^T \nabla^2 f(y+t(x-y))(x-y) \geq 0$

$\Rightarrow f(x) \geq f(y) + \nabla f^T(y)(x-y)$

Therefore f is convex.

Combining ① and ② if f is twice differentiable, then f is convex iff $\nabla^2 f(x)$ is PSD $\forall x \in \mathbb{R}^d$

(e) $\frac{\partial f}{\partial x_\ell \partial x_k}(x) = \frac{\partial}{\partial x_\ell}\left[\frac{\partial f(x)}{\partial x_k}\right] = \frac{\partial}{\partial x_\ell}\frac{\partial}{\partial x_k}\left[\frac{1}{2}x^T A x + b^T x + c\right]$

$$= \frac{\partial}{\partial x_\ell}\frac{\partial}{\partial x_k}\left[\frac{1}{2}\sum_{i=1}^{d} x_i \sum_{j=1}^{d} A_{ij} x_j + \sum_{i=1}^{d} b_i x_i + c\right]$$

$$= \frac{\partial}{\partial x_\ell}\left[\sum_{j=1}^{d} A_{kj} x_j + b_k\right]$$

$$= A_{k\ell} = A_{\ell k} \quad \text{since } A \text{ is symmetric.}$$

Therefore $\nabla_x^2 f(x) = A$

According to (d), we have

f is convex if and only if A is positive semi-definite on $\mathbb{R}^d$

Also, by the same argument followed by (d), we conclude

f is strictly convex if and only if A is positive definite on $\mathbb{R}^d$

In [13]:

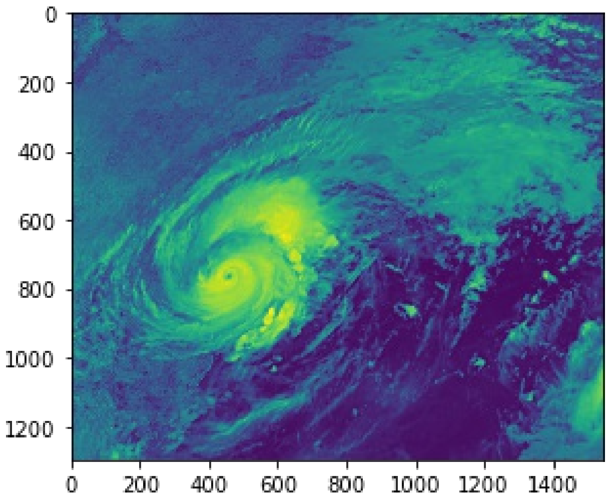
```
# Load the Libraries
import math
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

import cv2
```

In [6]:

```
# Load the color image in grayscale
img = cv2.imread('harvey-saturday-goes7am.jpg')
print("The size of the color image is " + str(img.shape))
gray = cv2.cvtColor(img, cv2.COLOR_RGB2GRAY)
print("The size of the grayscale image is " + str(gray.shape))
plt.imshow(gray)
plt.show()
```

```
The size of the color image is (1296, 1548, 3)
The size of the grayscale image is (1296, 1548)
```



In [25]:

```
# SVD
U, Sigma, V_trans = np.linalg.svd(gray, full_matrices=1)
print(U.shape)
print(Sigma.shape)
print(V_trans.shape)
```

```
(1296, 1296)
(1296,)
(1548, 1548)
```

In [29]:

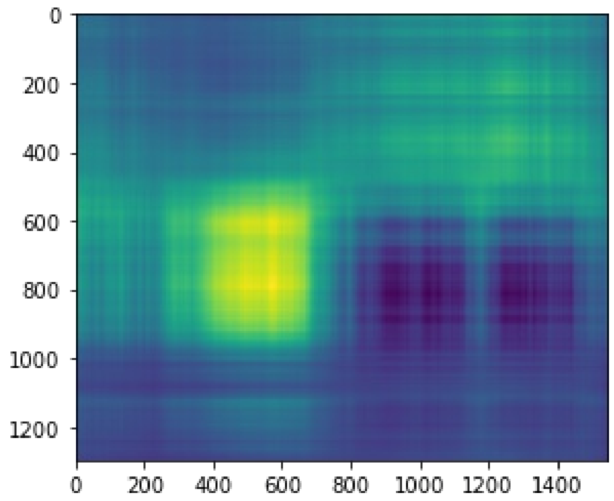
```
def F_norm(gray):
    """
    This function calculate the F-norm given a input matrix
    """
    tr_AA_transpose = np.trace(np.matmul(gray,gray.transpose()))
    norm = math.sqrt(tr_AA_transpose)
```

```
    norm = math.sqrt(tr_AA_transpose)
    return norm
print(F_norm(gray))
```

```
411.85191513455413
```

In [36]:

```
k = 2
X_bar = 0
for i in range(k):
    X_bar += Sigma[i] * np.outer(U.T[i], V_trans[i])
plt.imshow(X_bar)
plt.show()
```



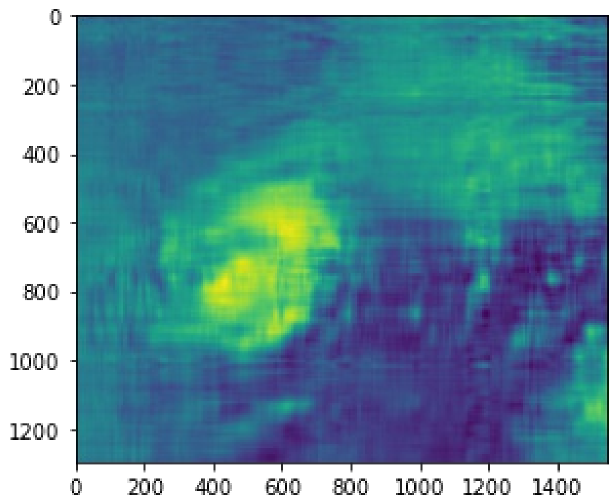
In [39]:

```
dif = np.subtract(gray,X_bar)
print(F_norm(dif)/F_norm(gray))
```

```
102.4922553316057
```

In [40]:

```
k = 10
X_bar = 0
for i in range(k):
    X_bar += Sigma[i] * np.outer(U.T[i], V_trans[i])
plt.imshow(X_bar)
plt.show()
```



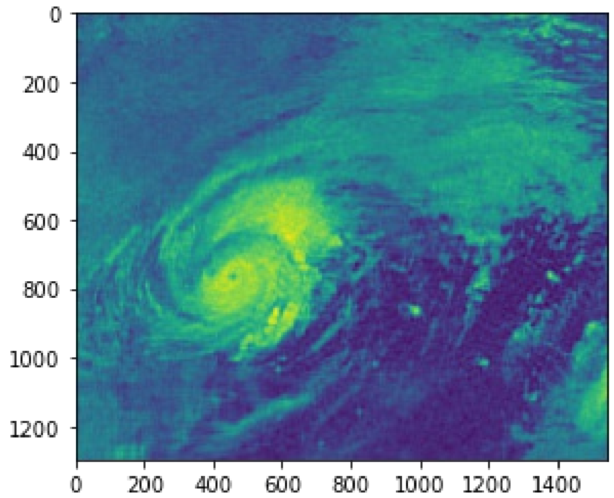
In [41]:

```
dif = np.subtract(gray,X_bar)
print(F_norm(dif)/F_norm(gray))
```

58.93770105042957

In [42]:

```
k = 40
X_bar = 0
for i in range(k):
    X_bar += Sigma[i] * np.outer(U.T[i], V_trans[i])
plt.imshow(X_bar)
plt.show()
```



In [43]:

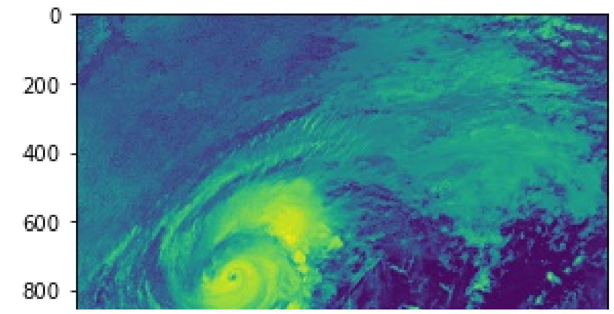
```
dif = np.subtract(gray,X_bar)
print(F_norm(dif)/F_norm(gray))
```

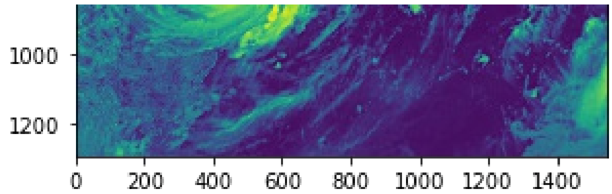
31.300100215056247

In [44]:

```
# this is only for test
k = 1296
X_bar = 0
for i in range(k):
    X_bar += Sigma[i] * np.outer(U.T[i], V_trans[i])
plt.imshow(X_bar)
plt.show()

dif = np.subtract(gray,X_bar)
print(F_norm(dif)/F_norm(gray))
```





(b) According to the fundamental theorem,(1296 $\times$ k) + (k $\times$ k) + (1548 $\times$ k) = 2844 $\times$ k + k^2
When k = 2, we need (1296 $\times$ 2) + (2 $\times$ 2) + (1548 $\times$ 2) =5692
When k = 10, we need (1296 $\times$ 10) + (10 $\times$ 10) + (1548 $\times$ 10) = 28540
When k = 40, we need (1296 $\times$ 40) + (40 $\times$ 40) + (1548 $\times$ 40) = 115360