Review

- ML: $\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(o|\theta)$
- MAP: $\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(\theta|o)$
- Bayesian: $\hat{x} = \sum_\theta p(\theta|o) \sum_x x p(x|\theta)$
- For zero-mean $\mathbf{X}$, $\Sigma_X = E[\mathbf{X}\mathbf{X}^T]$ and say we have $P^T \Sigma_X P = D$. The transformed $\mathbf{Y} = P^T \mathbf{X}$ are independent to each other
    - Note that the transform is just principal component analysis
- Marginalization of a normal distribution is still a normal distribution
- (a) $\Sigma_{\mathbf{XX}}^{-1} = \Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}} \Lambda_{\mathbf{YY}}^{-1} \Lambda_{\mathbf{YX}}$
- (b) $\det(a\Sigma) = \det(a\Sigma_{\mathbf{YY}}) \det(a\Lambda_{\mathbf{XX}}^{-1})$ for any constant a

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T \Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}})]$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})])$

[1] $tr(AB) = \sum_i \sum_j a_{i,j}b_{j,i} = \sum_j \sum_i b_{j,i}a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$
    $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$
    $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T])$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$ $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]) = \sum_{i=k+1}^n \lambda_i$
  - Similarly, if we "reconstruct" $\mathbf{X}$ as $\hat{\mathbf{X}} = P\hat{\mathbf{Y}}$. The mse of $\hat{\mathbf{X}} = E[(\mathbf{X}-\hat{\mathbf{X}})^T(\mathbf{X}-\hat{\mathbf{X}})] = tr(E[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T])$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$ $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]) = \sum_{i=k+1}^n \lambda_i$
  - Similarly, if we "reconstruct" $\mathbf{X}$ as $\hat{\mathbf{X}} = P\hat{\mathbf{Y}}$. The mse of $\hat{\mathbf{X}} = E[(\mathbf{X}-\hat{\mathbf{X}})^T(\mathbf{X}-\hat{\mathbf{X}})] = tr(E[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T])=$ $tr(PE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]P^T)$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$ $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]) = \sum_{i=k+1}^n \lambda_i$
  - Similarly, if we "reconstruct" $\mathbf{X}$ as $\hat{\mathbf{X}} = P\hat{\mathbf{Y}}$. The mse of $\hat{\mathbf{X}} = E[(\mathbf{X}-\hat{\mathbf{X}})^T(\mathbf{X}-\hat{\mathbf{X}})] = tr(E[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T])=$ $tr(PE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]P^T)= tr(P^TPE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T])$

[1] $tr(AB) = \sum_i\sum_j a_{i,j}b_{j,i} = \sum_j\sum_i b_{j,i}a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T \Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$ $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]) = \sum_{i=k+1}^n \lambda_i$
  - Similarly, if we "reconstruct" $\mathbf{X}$ as $\hat{\mathbf{X}} = P\hat{\mathbf{Y}}$. The mse of $\hat{\mathbf{X}} = E[(\mathbf{X}-\hat{\mathbf{X}})^T(\mathbf{X}-\hat{\mathbf{X}})] = tr(E[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T])=$ $tr(PE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]P^T)= tr(P^TPE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T])$ $= \sum_{i=k+1}^n \lambda_i$

[1] $tr(AB) = \sum_i \sum_j a_{i,j} b_{j,i} = \sum_j \sum_i b_{j,i} a_{i,j} = tr(BA)$

Principal component analysis (PCA)

- Recall that $\Sigma = E[\mathbf{X}\mathbf{X}^T]$ (assume $\mathbf{X}$ is zero-mean) and $\mathbf{Y} = P^T\mathbf{X}$ with $E[\mathbf{Y}\mathbf{Y}^T] = P^T\Sigma P = D$
- Assume that the diagonal of D (note that those are eigenvalues) are arranged in descending order that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
  - Generate an approximate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ by setting all components except first k as 0
  - The mean square error (mse) of[1] $\hat{\mathbf{Y}} = E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]$ $= tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})]) = E[tr((\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}))]$ $= E[tr((\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T)] = tr(E[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]) = \sum_{i=k+1}^n \lambda_i$
  - Similarly, if we "reconstruct" $\mathbf{X}$ as $\hat{\mathbf{X}} = P\hat{\mathbf{Y}}$. The mse of $\hat{\mathbf{X}} = E[(\mathbf{X}-\hat{\mathbf{X}})^T(\mathbf{X}-\hat{\mathbf{X}})] = tr(E[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T])=$ $tr(PE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T]P^T)= tr(P^TPE[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^T])$ $= \sum_{i=k+1}^n \lambda_i$
  - Note that the eigenvectors of Σ (columns of P) are known as the principal components

[1] $tr(AB) = \sum_i \sum_j a_{i,j}b_{j,i} = \sum_j \sum_i b_{j,i}a_{i,j} = tr(BA)$

Practical PCA

In practice, we typically are given a dataset with samples of **X** instead of the distribution or covariance matrix of **X**. Denote the data as $\mathcal{X}$ with each row is a data point and a total of m data points. Thus $\mathcal{X}$ is an m by n matrix

[2] I used the matlab notations for *ones*$(\cdot)$ and *mean*$(\cdot)$ here

[3] Note that $\hat{\Sigma}$ won't be full rank and positive definite as one would hope

Practical PCA

In practice, we typically are given a dataset with samples of $\mathbf{X}$ instead of the distribution or covariance matrix of $\mathbf{X}$. Denote the data as $\mathcal{X}$ with each row is a data point and a total of m data points. Thus $\mathcal{X}$ is an m by n matrix

- Data are rarely zero-mean to begin with, but we can easily preprocess it by subtracting the mean. That is[2] $\mathcal{X} \leftarrow \mathcal{X} - ones(m,1)mean(\mathcal{X})$

[2] I used the matlab notations for $ones(\cdot)$ and $mean(\cdot)$ here

[3] Note that $\hat{\Sigma}$ won't be full rank and positive definite as one would hope

Practical PCA

In practice, we typically are given a dataset with samples of $\mathbf{X}$ instead of the distribution or covariance matrix of $\mathbf{X}$. Denote the data as $\mathcal{X}$ with each row is a data point and a total of m data points. Thus $\mathcal{X}$ is an m by n matrix

- Data are rarely zero-mean to begin with, but we can easily preprocess it by subtracting the mean. That is[2] $\mathcal{X} \leftarrow \mathcal{X} - ones(m,1)mean(\mathcal{X})$
- Note that $\hat{\Sigma} \approx \frac{1}{m}\mathcal{X}^T\mathcal{X}$. We could directly compute the eigenvectors and eigenvalues of $\hat{\Sigma}$ as discussed previously. But in many cases, $m < n$ making $\hat{\Sigma}$ a bad approximate[3]

[2] I used the matlab notations for $ones(\cdot)$ and $mean(\cdot)$ here

[3] Note that $\hat{\Sigma}$ won't be full rank and positive definite as one would hope

Practical PCA

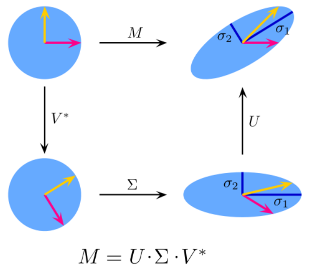
In practice, we typically are given a dataset with samples of **X** instead of the distribution or covariance matrix of **X**. Denote the data as $\mathcal{X}$ with each row is a data point and a total of m data points. Thus $\mathcal{X}$ is an m by n matrix

- Data are rarely zero-mean to begin with, but we can easily preprocess it by subtracting the mean. That is[2] $\mathcal{X} \leftarrow \mathcal{X} - ones(m,1)mean(\mathcal{X})$
- Note that $\hat{\Sigma} \approx \frac{1}{m}\mathcal{X}^T\mathcal{X}$. We could directly compute the eigenvectors and eigenvalues of $\hat{\Sigma}$ as discussed previously. But in many cases, $m < n$ making $\hat{\Sigma}$ a bad approximate[3]
    - A more common approach is to decompose $\mathcal{X}$ with singular value decomposition (SVD) instead

[2] I used the matlab notations for $ones(\cdot)$ and $mean(\cdot)$ here

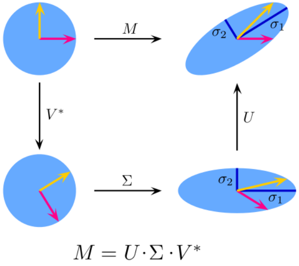
[3] Note that $\hat{\Sigma}$ won't be full rank and positive definite as one would hope

Singular value decomposition (SVD)



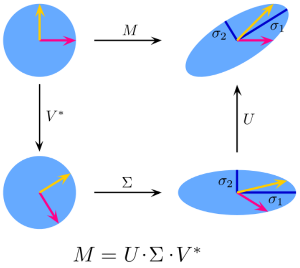
- Every matrix M can be decomposed as $M = UDV^\dagger$, where D is diagonal and U, V are unitary. The diagonal terms in Σ are known to be the singular values

Singular value decomposition (SVD)



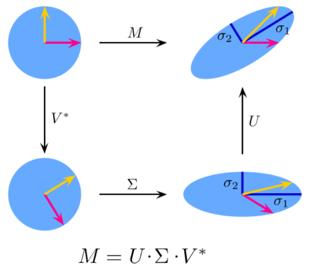
- Every matrix M can be decomposed as $M = UDV^\dagger$, where D is diagonal and U, V are unitary. The diagonal terms in Σ are known to be the singular values
- For real matrix M, we can write $M = UDV^T$ instead. U, V are now "real unitary" or orthogonal

Singular value decomposition (SVD)



- Every matrix M can be decomposed as $M = UDV^\dagger$, where D is diagonal and U, V are unitary. The diagonal terms in Σ are known to be the singular values
- For real matrix M, we can write $M = UDV^T$ instead. U, V are now "real unitary" or orthogonal
    - Note that $M^TM = VD^TU^TUDV^T = VD^2V^T$. Therefore, V are really eigenvectors of M^TM with eigenvalues equal to the square of the singular values

Singular value decomposition (SVD)



- Every matrix M can be decomposed as $M = UDV^\dagger$, where D is diagonal and U, V are unitary. The diagonal terms in Σ are known to be the singular values
- For real matrix M, we can write $M = UDV^T$ instead. U, V are now "real unitary" or orthogonal
  - Note that $M^T M = VD^T U^T UDV^T = VD^2 V^T$. Therefore, V are really eigenvectors of $M^T M$ with eigenvalues equal to the square of the singular values
  - Similar, we have $MM^T = UD^2 U^T$

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

- Estimate mean from data and subtract mean from that

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

- Estimate mean from data and subtract mean from that
- Decomposed the mean subtracted data with SVD. We get $\mathcal{X} = UDV^T$

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

- Estimate mean from data and subtract mean from that
- Decomposed the mean subtracted data with SVD. We get $\mathcal{X} = UDV^T$
- Note that column of V are now the principal components, and we can transform a data column as $V^T x$. The entire data set can be transformed as $\mathcal{Y} = \mathcal{X} V$

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

- Estimate mean from data and subtract mean from that
- Decomposed the mean subtracted data with SVD. We get $\mathcal{X} = UDV^T$
- Note that column of V are now the principal components, and we can transform a data column as $V^T x$. The entire data set can be transformed as $\mathcal{Y} = \mathcal{X} V$
  - The first few columns of $\mathcal{Y}$ will contain most "information" regarding the original $\mathcal{X}$

PCA with SVD

So from previous slides, instead of first estimating the covariance matrix and then diagonalize it. We should directly decompose the data $\mathcal{X}$ with SVD instead. The process is summarized below

- Estimate mean from data and subtract mean from that
- Decomposed the mean subtracted data with SVD. We get $\mathcal{X} = UDV^T$
- Note that column of V are now the principal components, and we can transform a data column as $V^T x$. The entire data set can be transformed as $\mathcal{Y} = \mathcal{X}V$
  - The first few columns of $\mathcal{Y}$ will contain most "information" regarding the original $\mathcal{X}$
  - For example, they can be taken as features for recognition or one can omit other columns besides the first few for "compression" as discussed earlier

Review

- ML: $\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(o|\theta)$
- MAP: $\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(\theta|o)$
- Bayesian: $\hat{x} = \sum_\theta p(\theta|o) \sum_x xp(x|\theta)$
- For zero-mean $\mathbf{X}$, $\Sigma_X = E[\mathbf{X}\mathbf{X}^T]$ and say we have $P^T \Sigma_X P = D$. The transformed $\mathbf{Y} = P^T \mathbf{X}$ are independent to each other
  - Note that the transform is just principal component analysis
- Marginalization of a normal distribution is still a normal distribution
- (a) $\Sigma_{\mathbf{XX}}^{-1} = \Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}} \Lambda_{\mathbf{YY}}^{-1} \Lambda_{\mathbf{YX}}$
- (b) $\det(a\Sigma) = \det(a\Sigma_{\mathbf{YY}}) \det(a\Lambda_{\mathbf{XX}}^{-1})$ for any constant a

Conditioning of normal distribution

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?

Conditioning of normal distribution

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?
- Basically, we want to find $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$

Conditioning of normal distribution

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{Z}, \Sigma_\mathbf{Z})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?
- Basically, we want to find $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$
- From previous result, we have $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_\mathbf{Y}, \Sigma_{\mathbf{YY}})$. Therefore,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}\left[\begin{pmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{pmatrix} - \tilde{\mathbf{y}}^T \Sigma_{\mathbf{YY}}^{-1} \tilde{\mathbf{y}}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}[\tilde{\mathbf{x}}^T \Lambda_{\mathbf{XX}} \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T \Lambda_{\mathbf{XY}} \tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}}]\right), \end{aligned}$$

where we use $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ as shorthands of $\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}$ and $\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}$ as before

Conditioning of normal distribution

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T\Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)$$
$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \mu_{\mathbf{Y}}))^T\Lambda_{\mathbf{XX}} (\mathbf{x} - \mu_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \mu_{\mathbf{Y}}))\right)$$

Conditioning of normal distribution

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T \Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)$$
$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \mu_{\mathbf{Y}}))^T \Lambda_{\mathbf{XX}} (\mathbf{x} - \mu_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \mu_{\mathbf{Y}}))\right)$$

- Therefore $\mathbf{X}|\mathbf{y}$ is Gaussian distributed with mean $\mu_{\mathbf{X}} - \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \mu_{\mathbf{Y}})$ and covariance $\Lambda_{\mathbf{XX}}^{-1}$

Conditioning of normal distribution

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T\Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)$$
$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))^T\Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))\right)$$

- Therefore $\mathbf{X}|\mathbf{y}$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\mathbf{X}} - \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})$ and covariance $\Lambda_{\mathbf{XX}}^{-1}$
- Note that since $\Lambda_{\mathbf{XX}}\Sigma_{\mathbf{XY}} + \Lambda_{\mathbf{XY}}\Sigma_{\mathbf{YY}} = 0 \Rightarrow \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}} = -\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}$ and from (a), we have

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

Interpretation of conditioning

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \mu_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of **Y** is exactly the mean, the conditioned mean does not change

Interpretation of conditioning

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \mu_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of **Y** is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of **Y** for the 1-D case.
  - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of **Y** to the variation of **X**

Interpretation of conditioning

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\mu_\mathbf{X} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \mu_\mathbf{Y}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of **Y** is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of **Y** for the 1-D case.
    - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of **Y** to the variation of **X**
    - In particular, if **X** and **Y** are negatively correlated, the sign of the adjustment will be reversed

Interpretation of conditioning

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \mu_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of **Y** is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of **Y** for the 1-D case.
  - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of **Y** to the variation of **X**
  - In particular, if **X** and **Y** are negatively correlated, the sign of the adjustment will be reversed
- As for the variance of the conditioned variable, it always decreases and the decrease is larger if $\Sigma_{\mathbf{YY}}$ is smaller and $\Sigma_{\mathbf{XY}}$ is larger (**X** and **Y** are more correlated)

$X \perp\!\!\!\perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Corollary

Given multivariate Gaussian variables X, Y and Z, we have X and Y are conditionally independent given Z if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$, where $\rho_{XZ} = \frac{E[(X-E(X))(Z-E(Z))]}{\sqrt{E[(X-E(X))^2]E[(Z-E(Z))^2]}}$ is the correlation coefficent between X and Z. Similarly, ρ_{YZ} and ρ_{XY} are the correlation coefficients between Y and Z, and X and Y, respectively.

$X \perp\!\!\!\perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- Without loss of generality, we can assume the variables with mean 0 and variance 1. Thus, $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}$

$X \perp\!\!\!\perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- Without loss of generality, we can assume the variables with mean 0 and variance 1. Thus, $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}$
- Then from the conditioning result, we have

$$\Sigma_{\left(\begin{smallmatrix} X \\ Y \end{smallmatrix}\right)|Z} = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{XZ} & \rho_{YZ} \end{pmatrix} \sigma_{YY}^{-1} \begin{pmatrix} \rho_{XZ} \\ \rho_{YZ} \end{pmatrix}$$
$$= \begin{pmatrix} 1 - \rho_{XZ}^2 & \rho_{XY} - \rho_{XZ}\rho_{YZ} \\ \rho_{XY} - \rho_{XZ}\rho_{YZ} & 1 - \rho_{YZ}^2 \end{pmatrix}$$

$X \perp\!\!\!\perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- Without loss of generality, we can assume the variables with mean 0 and variance 1. Thus, $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}$
- Then from the conditioning result, we have

$$\Sigma_{\left(\begin{smallmatrix} X \\ Y \end{smallmatrix}\right)\big|Z} = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{XZ} & \rho_{YZ} \end{pmatrix} \sigma_{YY}^{-1} \begin{pmatrix} \rho_{XZ} \\ \rho_{YZ} \end{pmatrix}$$
$$= \begin{pmatrix} 1-\rho_{XZ}^2 & \rho_{XY} - \rho_{XZ}\rho_{YZ} \\ \rho_{XY} - \rho_{XZ}\rho_{YZ} & 1-\rho_{YZ}^2 \end{pmatrix}$$

- Therefore, X and Y are uncorrelated given Z when $\sigma_{XY|Z} = \rho_{XY} - \rho_{XZ}\rho_{YZ} = 0$ or $\rho_{XY} = \rho_{XZ}\rho_{YZ}$. Since for Gaussian variables, uncorrelatedness implies independence. This concludes the proof.

□

Product of normal distributions

- Assume that we tries to recover some vector parameter $\mathbf{x}$, which is subject to multivariate Gaussian noise

Product of normal distributions

- Assume that we tries to recover some vector parameter $\mathbf{x}$, which is subject to multivariate Gaussian noise
- Say we made two measurements $\mathbf{y}_1$ and $\mathbf{y}_2$, where $\mathbf{Y}_1 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_1})$ and $\mathbf{Y}_2 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_2})$. Note that even though both measurements have mean $\mathbf{x}$, they have different covariance
  - This variation, for instance, can be due to environment change between the two measurements

Product of normal distributions

- Assume that we tries to recover some vector parameter $\mathbf{x}$, which is subject to multivariate Gaussian noise
- Say we made two measurements $\mathbf{y}_1$ and $\mathbf{y}_2$, where $\mathbf{Y}_1 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_1})$ and $\mathbf{Y}_2 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_2})$. Note that even though both measurements have mean $\mathbf{x}$, they have different covariance
  - This variation, for instance, can be due to environment change between the two measurements
- Now, if we want to compute the overall likelihood, $p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x})$. Assuming that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are conditionally independent given $\mathbf{X}$, we have

$$\begin{aligned} p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x}) &= p(\mathbf{y}_1|\mathbf{x})p(\mathbf{y}_2|\mathbf{x}) \\ &= \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}). \end{aligned}$$

Product of normal distributions

- Assume that we tries to recover some vector parameter $\mathbf{x}$, which is subject to multivariate Gaussian noise
- Say we made two measurements $\mathbf{y}_1$ and $\mathbf{y}_2$, where $\mathbf{Y}_1 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_1})$ and $\mathbf{Y}_2 \sim \mathcal{N}(\mathbf{x}, \Sigma_{\mathbf{Y}_2})$. Note that even though both measurements have mean $\mathbf{x}$, they have different covariance
  - This variation, for instance, can be due to environment change between the two measurements
- Now, if we want to compute the overall likelihood, $p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x})$. Assuming that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are conditionally independent given $\mathbf{X}$, we have

$$\begin{aligned} p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x}) &= p(\mathbf{y}_1|\mathbf{x})p(\mathbf{y}_2|\mathbf{x}) \\ &= \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}). \end{aligned}$$

- Essentially, we just need to compute the product of two Gaussian pdfs. Such computation is very useful and it occurs often when one needs to perform inference

Product of normal distributions

As in previous cases, the product turns out to be normal also. However, unlike them, the product is not a pdf and so it does not normalize to 1. So we have to compute both the scaling factor and the exponent explicitly. Let us start with the exponent.

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})$$

Product of normal distributions

As in previous cases, the product turns out to be normal also. However, unlike them, the product is not a pdf and so it does not normalize to 1. So we have to compute both the scaling factor and the exponent explicitly. Let us start with the exponent.

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})$$
$$\propto \exp\left(-\frac{1}{2}[(\mathbf{x}-\mathbf{y}_1)^T \Lambda_{\mathbf{Y}_1}(\mathbf{x}-\mathbf{y}_1) + (\mathbf{x}-\mathbf{y}_2)^T \Lambda_{\mathbf{Y}_2}(\mathbf{x}-\mathbf{y}_2)]\right)$$

Product of normal distributions

As in previous cases, the product turns out to be normal also. However, unlike them, the product is not a pdf and so it does not normalize to 1. So we have to compute both the scaling factor and the exponent explicitly. Let us start with the exponent.

$$
\begin{aligned}
&\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) \\
&\propto \exp\left(-\frac{1}{2}[(\mathbf{x}-\mathbf{y}_1)^T\Lambda_{\mathbf{Y}_1}(\mathbf{x}-\mathbf{y}_1)+(\mathbf{x}-\mathbf{y}_2)^T\Lambda_{\mathbf{Y}_2}(\mathbf{x}-\mathbf{y}_2)]\right) \\
&\propto \exp\left(-\frac{1}{2}[\mathbf{x}^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})\mathbf{x}-(\mathbf{y}_2^T\Lambda_{\mathbf{Y}_2}+\mathbf{y}_1^T\Lambda_{\mathbf{Y}_1})\mathbf{x}-\mathbf{x}^T(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1)]\right)
\end{aligned}
$$

Product of normal distributions

As in previous cases, the product turns out to be normal also. However, unlike them, the product is not a pdf and so it does not normalize to 1. So we have to compute both the scaling factor and the exponent explicitly. Let us start with the exponent.

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})$$

$$\propto \exp\left(-\frac{1}{2}[(\mathbf{x}-\mathbf{y}_1)^T\Lambda_{\mathbf{Y}_1}(\mathbf{x}-\mathbf{y}_1)+(\mathbf{x}-\mathbf{y}_2)^T\Lambda_{\mathbf{Y}_2}(\mathbf{x}-\mathbf{y}_2)]\right)$$

$$\propto \exp\left(-\frac{1}{2}[\mathbf{x}^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})\mathbf{x}-(\mathbf{y}_2^T\Lambda_{\mathbf{Y}_2}+\mathbf{y}_1^T\Lambda_{\mathbf{Y}_1})\mathbf{x}-\mathbf{x}^T(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1)]\right)$$

$$\propto e^{-\frac{1}{2}[(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))]}$$

Product of normal distributions

As in previous cases, the product turns out to be normal also. However, unlike them, the product is not a pdf and so it does not normalize to 1. So we have to compute both the scaling factor and the exponent explicitly. Let us start with the exponent.

$$
\begin{aligned}
&\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})\\
\propto &\exp\left(-\frac{1}{2}[(\mathbf{x}-\mathbf{y}_1)^T\Lambda_{\mathbf{Y}_1}(\mathbf{x}-\mathbf{y}_1)+(\mathbf{x}-\mathbf{y}_2)^T\Lambda_{\mathbf{Y}_2}(\mathbf{x}-\mathbf{y}_2)]\right)\\
\propto &\exp\left(-\frac{1}{2}[\mathbf{x}^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})\mathbf{x}-(\mathbf{y}_2^T\Lambda_{\mathbf{Y}_2}+\mathbf{y}_1^T\Lambda_{\mathbf{Y}_1})\mathbf{x}-\mathbf{x}^T(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1)]\right)\\
\propto &e^{-\frac{1}{2}[(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))]}\\
\propto &\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1), (\Lambda_{\mathbf{Y}_2}+\Lambda_{\mathbf{Y}_1})^{-1})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})\\
= &K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1), (\Lambda_{\mathbf{Y}_2}+\Lambda_{\mathbf{Y}_1})^{-1})
\end{aligned}
$$

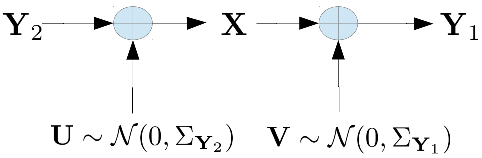
for some scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ independent of $\mathbf{x}$.

Product of normal distributions

- One can compute the scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ directly

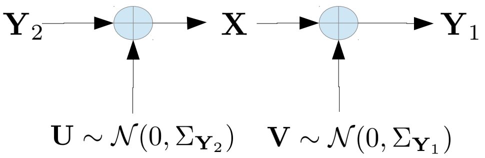
Product of normal distributions

- One can compute the scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ directly
- However, it is much easier to take advantage for the following setup when $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$ as shown below



Product of normal distributions

- One can compute the scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ directly
- However, it is much easier to take advantage for the following setup when $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$ as shown below

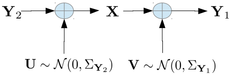


- Since $\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) = \mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2})$ and $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$, we have

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) = \underbrace{\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})}_{p(y_1|x)=p(y_1|x,y_2)}\underbrace{\mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2})}_{p(x|y_2)} = p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)$$

Product of normal distributions

- Then, marginalizing $\mathbf{x}$ out from $p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)$, we have

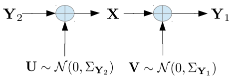


$p(\mathbf{y}_1|\mathbf{y}_2) = \int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x}$. However, from the figure,

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$$

Product of normal distributions

- Then, marginalizing $\mathbf{x}$ out from $p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)$, we have $p(\mathbf{y}_1|\mathbf{y}_2) = \int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x}$. However, from the figure,



$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$$

- On the other hand,

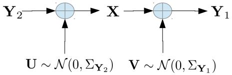
$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = \int \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})d\mathbf{x}$$

$$= \int K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})d\mathbf{x}$$

$$= K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}).$$

Product of normal distributions

- Then, marginalizing $\mathbf{x}$ out from $p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)$, we have $p(\mathbf{y}_1|\mathbf{y}_2) = \int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x}$. However, from the figure,



$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x} = p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$$

- On the other hand,

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x} = \int \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1}) \mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) d\mathbf{x}$$

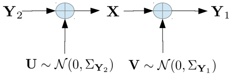
$$= \int K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}) \mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}) d\mathbf{x}$$

$$= K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}).$$

- Thus we have $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$

Product of normal distributions

- Then, marginalizing $\mathbf{x}$ out from $p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)$, we have



$p(\mathbf{y}_1|\mathbf{y}_2) = \int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x}$. However, from the figure,

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x} = p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$$

- On the other hand,

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) d\mathbf{x} = \int \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1}) \mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) d\mathbf{x}$$

$$= \int K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}) \mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}) d\mathbf{x}$$

$$= K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}).$$

- Thus we have $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$ and so

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1}) \mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})$$

$$= \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1}) \mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})$$

Product of normal distributions

Let us try to interpret the product as the overall likelihood after making two observations. Consider the simpler case when $\mathbf{X}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are all scaler

Product of normal distributions

Let us try to interpret the product as the overall likelihood after making two observations. Consider the simpler case when $\mathbf{X}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are all scaler

- The mean considering both observations, $(\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y)$, is essential a weighted average of observations $\mathbf{y}_2$ and $\mathbf{y}_1$
  - The weight is higher when the precision $\Lambda_{\mathbf{Y}_2}$ or $\Lambda_{\mathbf{Y}_1}$ is larger

Product of normal distributions

Let us try to interpret the product as the overall likelihood after making two observations. Consider the simpler case when $\mathbf{X}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are all scaler

- The mean considering both observations, $(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}y)$, is essential a weighted average of observations $\mathbf{y}_2$ and $\mathbf{y}_1$
  - The weight is higher when the precision $\Lambda_{\mathbf{Y}_2}$ or $\Lambda_{\mathbf{Y}_1}$ is larger
- The overall variance $(\Lambda_{\mathbf{Y}_2}+\Lambda_{\mathbf{Y}_1})^{-1}$ is always smaller than the individual variance $\Sigma_{\mathbf{Y}_2}$ and $\Sigma_{\mathbf{Y}_1}$
  - We are more certain with $\mathbf{x}$ after considering both $\mathbf{y}_1$ and $\mathbf{y}_2$

Product of normal distributions

Let us try to interpret the product as the overall likelihood after making two observations. Consider the simpler case when $\mathbf{X}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are all scaler

- The mean considering both observations, $(\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y)$, is essential a weighted average of observations $\mathbf{y}_2$ and $\mathbf{y}_1$
  - The weight is higher when the precision $\Lambda_{\mathbf{Y}_2}$ or $\Lambda_{\mathbf{Y}_1}$ is larger
- The overall variance $(\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}$ is always smaller than the individual variance $\Sigma_{\mathbf{Y}_2}$ and $\Sigma_{\mathbf{Y}_1}$
  - We are more certain with $\mathbf{x}$ after considering both $\mathbf{y}_1$ and $\mathbf{y}_2$
- The scaling factor, $\mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})$, can be interpreted as how much one can believe on the overall likelihood.
  - The value is reasonable since when the two observations are far away with respect to the overall variance $\Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1}$, the likelihood will become less reliable
  - The scaling factor is especially useful when we deal with mixture of Gaussian to be discussed next

Division of normal distributions

- To compute $\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}$, note that from the product formula earlier

$$
\begin{aligned}
&\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})\\
=&\mathcal{N}(\boldsymbol{\mu}_1;(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)
\end{aligned}
$$

Division of normal distributions

- To compute $\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}$, note that from the product formula earlier

$$\begin{aligned}&\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})\\ =&\mathcal{N}(\boldsymbol{\mu}_1;(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)\end{aligned}$$

- Therefore,

$$\begin{aligned}\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}&=\frac{\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1,(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2);\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})}\\&=\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1;\boldsymbol{\mu},\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})},\end{aligned}$$

where $\boldsymbol{\mu}=(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2)$

Division of normal distributions

- To compute $\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}$, note that from the product formula earlier

$$\begin{aligned}&\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})\\ =&\mathcal{N}(\boldsymbol{\mu}_1;(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)\end{aligned}$$

- Therefore,

$$\begin{aligned}\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}&=\frac{\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1,(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2);\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})}\\ &=\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1;\boldsymbol{\mu},\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})},\end{aligned}$$

  where $\boldsymbol{\mu}=(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2)$

- Note that the final pdf will be Gaussian-like if $\Lambda_1 \succeq \Lambda_2$. Otherwise, one can still write out the pdf using the precision matrix. But the covariance matrix will not be defined (Try plot some pdfs out yourselves)

Mixture of Gaussians

Consider an electrical system that outputs signal of different statistics when it is on and off

- When the system is on, the output signal S behaves like $\mathcal{N}(5, 1)$. When the system is off is off, S behaves like $\mathcal{N}(0, 1)$

Mixture of Gaussians

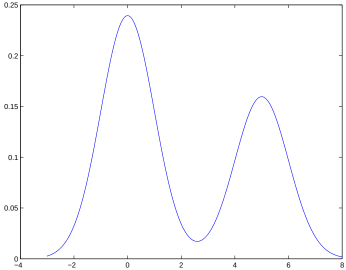
Consider an electrical system that outputs signal of different statistics when it is on and off

- When the system is on, the output signal S behaves like $\mathcal{N}(5, 1)$. When the system is off is off, S behaves like $\mathcal{N}(0, 1)$
- If someone measuring the signal does not know the status of the system but only knows that the system is on 40% of the time, then to the observer, the signal S behaves like a mixture of Gaussians

Mixture of Gaussians

Consider an electrical system that outputs signal of different statistics when it is on and off

- When the system is on, the output signal S behaves like $\mathcal{N}(5,1)$. When the system is off is off, S behaves like $\mathcal{N}(0,1)$
- If someone measuring the signal does not know the status of the system but only knows that the system is on 40% of the time, then to the observer, the signal S behaves like a mixture of Gaussians
- The pdf of S will be $0.4\mathcal{N}(s;5,1)+0.6\mathcal{N}(s;0,1)$ as shown below



Mixture of Gaussians

- A main limitation of normal distribution is that it is unimodal

Mixture of Gaussians

- A main limitation of normal distribution is that it is unimodal
- Mixture of Gaussian distribution allows multimodal and can virtually model any pdfs. But there is a computational cost for this gain

Mixture of Gaussians

- A main limitation of normal distribution is that it is unimodal
- Mixture of Gaussian distribution allows multimodal and can virtually model any pdfs. But there is a computational cost for this gain
- Let us illustrate this with the following example:
    - Consider two mixtures of Gaussian likelihood of x given two observations y_1 and y_2 as follows:

$$p(y_1|x) = 0.6\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 5, 1);$$
$$p(y_2|x) = 0.5\mathcal{N}(x; -2, 1) + 0.5\mathcal{N}(x; 4, 1).$$

    What is the overall likelihood, $p(y_1, y_2|x)$?

Mixture of Gaussians

- A main limitation of normal distribution is that it is unimodal
- Mixture of Gaussian distribution allows multimodal and can virtually model any pdfs. But there is a computational cost for this gain
- Let us illustrate this with the following example:
  - Consider two mixtures of Gaussian likelihood of x given two observations y_1 and y_2 as follows:

$$p(y_1|x) = 0.6\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 5, 1);$$
$$p(y_2|x) = 0.5\mathcal{N}(x; -2, 1) + 0.5\mathcal{N}(x; 4, 1).$$

  What is the overall likelihood, $p(y_1, y_2|x)$?
- As usual, it is reasonable to assume the observations to be conditionally independent given x. Then,

$$\begin{aligned} p(y_1, y_2|x) &= p(y_1|x)p(y_2|x) \\ &= (0.6\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 5, 1))(0.5\mathcal{N}(x; -2, 1) + 0.5\mathcal{N}(x; 4, 1)) \\ &= 0.3\mathcal{N}(x; 0, 1)\mathcal{N}(x; -2, 1) + 0.2\mathcal{N}(x; 5, 1)\mathcal{N}(x; -2, 1) \\ &\quad + 0.3\mathcal{N}(x; 0, 1)\mathcal{N}(x; 4, 1) + 0.2\mathcal{N}(x; 5, 1)\mathcal{N}(x; 4, 1) \end{aligned}$$

Explosion of Gaussians

- The last step involves computing products of Gaussians but we have learned it in previous sections. Using the previous result,

$$
\begin{aligned}
p(y_1, y_2|x) = &\ 0.3\mathcal{N}(-2; 0, 2)\mathcal{N}(x; -1, 0.5) + 0.2\mathcal{N}(-2; 5, 2)\mathcal{N}(x; 1.5, 0.5) \\
&+ 0.3\mathcal{N}(4; 0, 2)\mathcal{N}(x; 2, 0.5) + 0.2\mathcal{N}(4; 5, 2)\mathcal{N}(x; 4.5, 0.5).
\end{aligned}
$$

Explosion of Gaussians

- The last step involves computing products of Gaussians but we have learned it in previous sections. Using the previous result,

$$
\begin{aligned}
p(y_1, y_2|x) = {} & 0.3\mathcal{N}(-2; 0, 2)\mathcal{N}(x; -1, 0.5) + 0.2\mathcal{N}(-2; 5, 2)\mathcal{N}(x; 1.5, 0.5) \\
& + 0.3\mathcal{N}(4; 0, 2)\mathcal{N}(x; 2, 0.5) + 0.2\mathcal{N}(4; 5, 2)\mathcal{N}(x; 4.5, 0.5).
\end{aligned}
$$

So we have the overall likelihood is a mixture of four Gaussians

Explosion of Gaussians

- The last step involves computing products of Gaussians but we have learned it in previous sections. Using the previous result,

$$\begin{aligned} p(y_1, y_2|x) = &\ 0.3\mathcal{N}(-2; 0, 2)\mathcal{N}(x; -1, 0.5) + 0.2\mathcal{N}(-2; 5, 2)\mathcal{N}(x; 1.5, 0.5) \\ &+ 0.3\mathcal{N}(4; 0, 2)\mathcal{N}(x; 2, 0.5) + 0.2\mathcal{N}(4; 5, 2)\mathcal{N}(x; 4.5, 0.5). \end{aligned}$$

  So we have the overall likelihood is a mixture of four Gaussians
- Let's repeat our discussion but with n observations instead. The overall likelihood will be a mixture of 2^n Gaussians!
  - Therefore, the computation will quickly become intractable as the number of observations increases

Explosion of Gaussians

- The last step involves computing products of Gaussians but we have learned it in previous sections. Using the previous result,

$$\begin{aligned} p(y_1, y_2|x) = &\ 0.3\mathcal{N}(-2; 0, 2)\mathcal{N}(x; -1, 0.5) + 0.2\mathcal{N}(-2; 5, 2)\mathcal{N}(x; 1.5, 0.5) \\ &+ 0.3\mathcal{N}(4; 0, 2)\mathcal{N}(x; 2, 0.5) + 0.2\mathcal{N}(4; 5, 2)\mathcal{N}(x; 4.5, 0.5). \end{aligned}$$

  So we have the overall likelihood is a mixture of four Gaussians
- Let's repeat our discussion but with n observations instead. The overall likelihood will be a mixture of 2^n Gaussians!
  - Therefore, the computation will quickly become intractable as the number of observations increases
  - Fortunately, in reality, some of the Gaussians in the mixture tend to have a very small weight

Reduce number of components in Gaussian mixtures

- For instance, in our previous numerical example, if we continue our numerical computation for the two observation example, we have

$$\begin{aligned} p(y_1, y_2|x) = {} & 0.4163\mathcal{N}(x; -1, 0.5) + 3.5234 \times 10^{-6}\mathcal{N}(x; 1.5, 0.5) \\ & + 0.0202\mathcal{N}(x; 2, 0.5) + 0.5734\mathcal{N}(x; 4.5, 0.5). \end{aligned}$$

Reduce number of components in Gaussian mixtures

- For instance, in our previous numerical example, if we continue our numerical computation for the two observation example, we have

$$p(y_1, y_2|x) = 0.4163\mathcal{N}(x; -1, 0.5) + 3.5234 \times 10^{-6}\mathcal{N}(x; 1.5, 0.5) + 0.0202\mathcal{N}(x; 2, 0.5) + 0.5734\mathcal{N}(x; 4.5, 0.5).$$

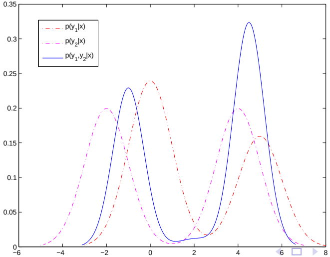
- We can see that the weight for the component at mean 1.5 is very small. And the component at mean 2 has a rather small weight also.

Reduce number of components in Gaussian mixtures

- For instance, in our previous numerical example, if we continue our numerical computation for the two observation example, we have

$$p(y_1, y_2|x) = 0.4163\mathcal{N}(x; -1, 0.5) + 3.5234 \times 10^{-6}\mathcal{N}(x; 1.5, 0.5) + 0.0202\mathcal{N}(x; 2, 0.5) + 0.5734\mathcal{N}(x; 4.5, 0.5).$$

- We can see that the weight for the component at mean 1.5 is very small. And the component at mean 2 has a rather small weight also.
- Even with the four Gaussian components, the overall likelihood is essentially just a bimodal distribution as shown in the figure below



Reduce number of components in Gaussian mixtures

- Therefore, we may approximate $p(y_1, y_2|x)$ with only two of its original component as $0.4163/(0.4163 + 0.5734)\mathcal{N}(x; -1, 0.5) + 0.5734/(0.4163 + 0.5734)\mathcal{N}(x; 4.5, 0.5) = 0.4206\mathcal{N}(x; -1, 0.5) + 0.5794\mathcal{N}(x; 4.5, 0.5)$

Reduce number of components in Gaussian mixtures

- Therefore, we may approximate $p(y_1, y_2|x)$ with only two of its original component as $0.4163/(0.4163 + 0.5734)\mathcal{N}(x; -1, 0.5) + 0.5734/(0.4163 + 0.5734)\mathcal{N}(x; 4.5, 0.5) = 0.4206\mathcal{N}(x; -1, 0.5) + 0.5794\mathcal{N}(x; 4.5, 0.5)$
- However, it is not always a good approximation strategy just to dump away the small components in a Gaussian mixture

Another example

Consider

$$p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1).$$

Another example

Consider

$$p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1).$$

- Let say we want to reduce $p(x)$ to only a mixture of two Gaussians. It is tempting to just dumping four smallest one and renormalized the weight. For example, if we choose to remove the first four components, we have

$$\hat{p}(x) = 1/6\mathcal{N}(x; 0.2, 1) + 5/6\mathcal{N}(x; 5, 1)$$

Another example

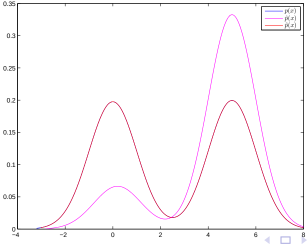
Consider

$$p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1).$$

- Let say we want to reduce $p(x)$ to only a mixture of two Gaussians. It is tempting to just dumping four smallest one and renormalized the weight. For example, if we choose to remove the first four components, we have

$$\hat{p}(x) = 1/6\mathcal{N}(x; 0.2, 1) + 5/6\mathcal{N}(x; 5, 1)$$

- The approximation $\hat{p}(x)$ is significantly different from $p(x)$ as shown below



Merging components

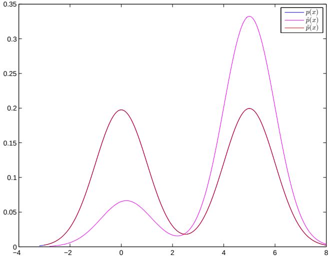
- The problem is that while the first five components are all relatively small compared to the last one, they are all quite similar and their combined contribution is comparable to the latter

Merging components

- The problem is that while the first five components are all relatively small compared to the last one, they are all quite similar and their combined contribution is comparable to the latter
- Actually the first five components are so similar that their combined contribution can be accurately modeled as one Gaussian

Merging components

- The problem is that while the first five components are all relatively small compared to the last one, they are all quite similar and their combined contribution is comparable to the latter
- Actually the first five components are so similar that their combined contribution can be accurately modeled as one Gaussian
- So rather than discarding the components, one can get a much more accurate approximation by merging them. The approximation is illustrated as $\tilde{p}(x)$ in the figure below



Merging components

To successfully obtain such approximation $\tilde{p}(x)$, we have to answer two questions:

- which components to merge?
- how to merge them?

Which Components to Merge?

It is reasonable to pick similar components to merge. The question is how do will gauge the similarity between two components.

Which Components to Merge?

It is reasonable to pick similar components to merge. The question is how do will gauge the similarity between two components.

- Consider two pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$, note that we can define an inner product of $p(\mathbf{x})$ and $q(\mathbf{x})$ by

$$\langle p(\mathbf{x}), q(\mathbf{x}) \rangle = \int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}$$

Which Components to Merge?

It is reasonable to pick similar components to merge. The question is how do will gauge the similarity between two components.

- Consider two pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$, note that we can define an inner product of $p(\mathbf{x})$ and $q(\mathbf{x})$ by

$$\langle p(\mathbf{x}), q(\mathbf{x}) \rangle = \int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}$$

- Note that the inner product is well defined and $\langle p(\mathbf{x}), p(\mathbf{x}) \rangle \geq 0$

Which Components to Merge?

It is reasonable to pick similar components to merge. The question is how do will gauge the similarity between two components.

- Consider two pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$, note that we can define an inner product of $p(\mathbf{x})$ and $q(\mathbf{x})$ by

$$\langle p(\mathbf{x}), q(\mathbf{x})\rangle = \int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}$$

- Note that the inner product is well defined and $\langle p(\mathbf{x}), p(\mathbf{x})\rangle \geq 0$
- By Cauchy-Schwartz inequality,

$$\frac{\langle p(\mathbf{x}), q(\mathbf{x})\rangle}{\sqrt{\langle p(\mathbf{x}), p(\mathbf{x})\rangle\langle q(\mathbf{x}), q(\mathbf{x})\rangle}} = \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}} \leq 1$$

Which Components to Merge?

It is reasonable to pick similar components to merge. The question is how do will gauge the similarity between two components.

- Consider two pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$, note that we can define an inner product of $p(\mathbf{x})$ and $q(\mathbf{x})$ by

$$\langle p(\mathbf{x}), q(\mathbf{x})\rangle = \int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}$$

- Note that the inner product is well defined and $\langle p(\mathbf{x}), p(\mathbf{x})\rangle \geq 0$
- By Cauchy-Schwartz inequality,

$$\frac{\langle p(\mathbf{x}), q(\mathbf{x})\rangle}{\sqrt{\langle p(\mathbf{x}), p(\mathbf{x})\rangle \langle q(\mathbf{x}), q(\mathbf{x})\rangle}} = \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}} \leq 1$$

- The inner product maximizes $(= 1)$ when $p(\mathbf{x}) = q(\mathbf{x})$. This suggests a very reasonable similarity measure between two pdfs

Similarity measure

- Let's define

$$Sim(p(\mathbf{x}), q(\mathbf{x})) \triangleq \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}}$$

Similarity measure

- Let's define

$$Sim(p(\mathbf{x}), q(\mathbf{x})) \triangleq \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}}$$

- In particular, if $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_p, \Sigma_p)$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_q, \Sigma_q)$, we have (please verify)

$$Sim(\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p), \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)) = \frac{\mathcal{N}(\boldsymbol{\mu}_p; \boldsymbol{\mu}_q, \Sigma_p + \Sigma_q)}{\sqrt{\mathcal{N}(0; 0, 2\Sigma_p)\mathcal{N}(0; 0, 2\Sigma_q)}},$$

which can be computed very easily and is equal to one only when means and covariances are the same

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^{n} w_i$

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^{n} w_i$
- Combined mean will simply be $\sum_{i=1}^{n} \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^{n} w_i$
- Combined mean will simply be $\sum_{i=1}^{n} \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$
- For combined covariance, it may be tempting to approximate it as $\sum_{i=1}^{n} \hat{w}_i \Sigma_i$.

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^n w_i$
- Combined mean will simply be $\sum_{i=1}^n \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$
- For combined covariance, it may be tempting to approximate it as $\sum_{i=1}^n \hat{w}_i \Sigma_i$.
    - However, it is an underestimate

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^n w_i$
- Combined mean will simply be $\sum_{i=1}^n \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$
- For combined covariance, it may be tempting to approximate it as $\sum_{i=1}^n \hat{w}_i \Sigma_i$.
  - However, it is an underestimate
  - Because the weighted sum only counted the contribution of variation among each component, it did not take into account the variation due to different means across components.

How to Merge Components?

Say we have n components $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, $\cdots$, $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ with weights $w_1, w_2, \cdots, w_n$. What should the combined component be like?

- Combined component weight should equal to total weight $\sum_{i=1}^n w_i$
- Combined mean will simply be $\sum_{i=1}^n \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$
- For combined covariance, it may be tempting to approximate it as $\sum_{i=1}^n \hat{w}_i \Sigma_i$.
  - However, it is an underestimate
  - Because the weighted sum only counted the contribution of variation among each component, it did not take into account the variation due to different means across components.
  - Instead, let's denote $\mathbf{X}$ as the variable sampled from the mixture. That is, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ with probability $\hat{w}_i$. Then, we have (please verify)

$$\Sigma = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$$
$$= \sum_{i=1}^n \hat{w}_i(\Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) - \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i \hat{w}_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j^T.$$

Now, go back to our previous numerical example

- Recall that $p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1)$

Now, go back to our previous numerical example

- Recall that $p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1)$
- If we merge the five smallest components (one can easily check that they are also more similar to each other than to the last component), we have $\tilde{p}(x) = 0.5\mathcal{N}(x; 0, 1.02) + 0.5\mathcal{N}(x; 5, 1)$ as shown again below. The approximate pdf is virtually indistinguishable from the original