Review

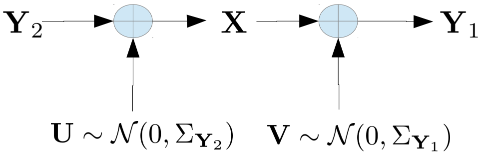
- PCA (assume zero mean)
  - Via eigen-decomposition
    1. $\Sigma \approx \frac{1}{m}\mathcal{X}^T\mathcal{X}$
    2. $P^T \Sigma P = D$
    3. $Y = P^T X$
  - Via SVD
    1. $U^T \mathcal{X} V = D$
    2. $Y = V^T X$
- Marginalization of a normal distribution is still a normal distribution
- Conditioning of normal distribution:
  $\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$
- Product of normal distribution:
  $\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) =$
  $\mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})$

Correction: 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

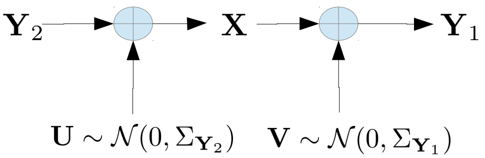
Correction: product of normal distributions

- One can compute the scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ directly
- However, recall that $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$, it is model the variables as shown below



Correction: 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, recall that $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$, it is model the variables as shown below

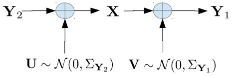


- Since $\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) = \mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2})$ and $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 | \mathbf{X}$, we have

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

Correction: product of normal distributions

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



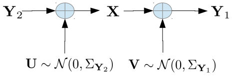
$p(\mathbf{y}_1|\mathbf{y}_2) = \int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x}$. However, from the figure,

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

Correction: 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,

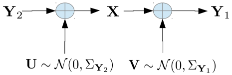
$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = \int \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})d\mathbf{x}$$

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

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

Correction: 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,

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

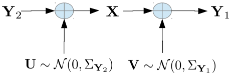
$$= \int K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})d\mathbf{x}$$

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

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

Correction: 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

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

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

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

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

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

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

- Therefore,

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

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

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

- Therefore,

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

  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

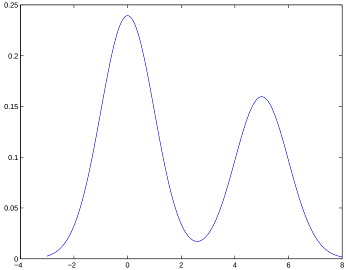
Consider an electrical system that outputs signal of different statistics when it is on and off

- 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

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

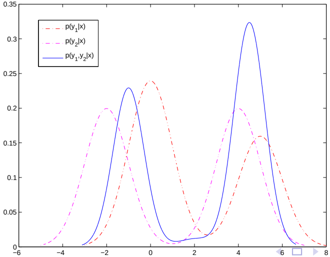
- We can see that the weight for the component at mean 1.5 is very small. And the component at mean 2 has a rather small weight also.

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

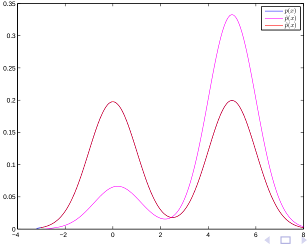
Consider

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

- 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

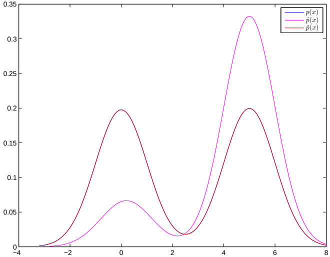
- The problem is that while the first five components are all relatively small compared to the last one, they are all quite similar and their combined contribution is comparable to the latter

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)

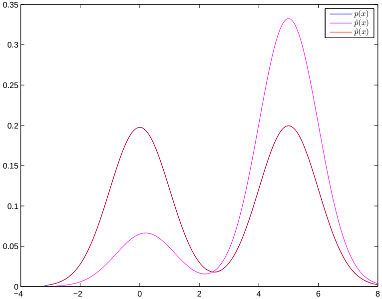
$$\Sigma = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$$
$$= \sum_{i=1}^n \hat{w}_i(\Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) - \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i \hat{w}_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j^T.$$

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



Review multivariate normal

- Marginalization of a normal distribution is still a normal distribution
- Conditioning of normal distribution:
  $\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$
- Product of normal distribution:
  $\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) =$
  $\mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1} y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})$
- Division of normal distribution:

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

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

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

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$.

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

- More concisely, we can write it as

$$Bern(x|p) = p^x(1 - p)^{1-x},$$

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

- More concisely, we can write it as

$$Bern(x|p) = p^x (1-p)^{1-x},$$

- The mean and variance are

$$E[X] = p \cdot 1 + (1-p) \cdot 0 = p$$
$$Var[X] = p \cdot (1-p)^2 + (1-p) \cdot p^2 = p(1-p)$$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p, N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p, N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p, N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p, N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p, N-1)$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p, N-1)$
    $= Np$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p, N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p,N-2) = N(N-1)p^2$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p,N-2) = N(N-1)p^2$
  - Therefore, $Var[X] = E[X^2] - E[X]^2$

Binomial distribution (N trials)

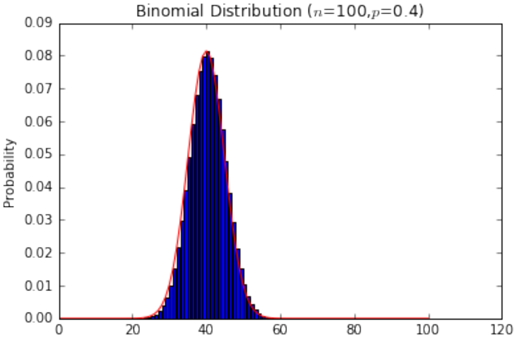
- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p, N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p, N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p, N-2) = N(N-1)p^2$
  - Therefore, $Var[X] = E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2 = N(N-1)p^2 + Np - (Np)^2 = Np(1-p)$

Binomial distribution

As shown below, the binomial distribution can be model well with a normal distribution $\mathcal{N}(Np, Np(1-p))$ for large N



The binomial distribution is shown in blue and an approximation by normal distribution is shown in red

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u, v|p) = \arg\max_p p^u(1-p)^v$$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u, v|p) = \arg\max_p p^u(1-p)^v$$

- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead

$$\hat{p} = \arg\max_p p(u, v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute
$$\hat{p} = \arg\max_p p(u, v|p) = \arg\max_p p^u(1-p)^v$$
- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead
$$\hat{p} = \arg\max_p p(u, v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$
- It is very difficult to determine the prior unanimously. Actually it can be controversial just to determine the form of it

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u, v|p) = \arg\max_p p^u(1-p)^v$$

- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead

$$\hat{p} = \arg\max_p p(u, v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$

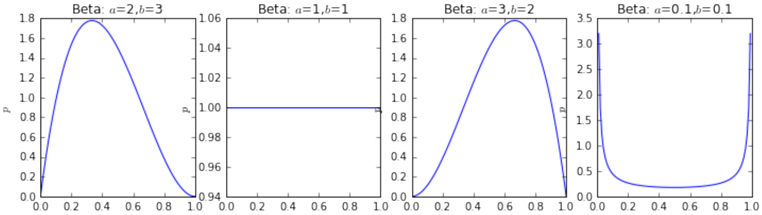
- It is very difficult to determine the prior unanimously. Actually it can be controversial just to determine the form of it
- However, if we select $p(p)$ of a form $p(p) \propto p^a(1-p)^b$, then the resulting posterior distribution with the same form as before. This choice is often chosen for practical purposes, and a prior with same "form" as its likelihood (and thus posterior) is known as the conjugate prior

Beta distribution

- The conjugate prior of both Bernoulli and binomial distributions is the beta distribution. Its pdf is given by

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},$$

where $X \in [0,1]$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

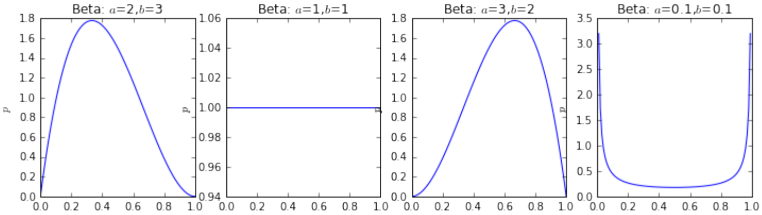


Beta distribution

- The conjugate prior of both Bernoulli and binomial distributions is the beta distribution. Its pdf is given by

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},$$

where $X \in [0,1]$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$



- Note that with $a = b = 1$, $Beta(x|1,1) = 1$. It is the same as no prior

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}\Big|_0^\infty = 1$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x}$$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x}$$

$$= -x^{z-1} e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx$$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x}$$

$$= -x^{z-1} e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx$$

$$= (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx = (z-1)\Gamma(z-1)$$

□

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x}$$
$$= -x^{z-1} e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx$$
$$= (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx = (z-1)\Gamma(z-1)$$
□

- Therefore, for integer $z > 1$, $\Gamma(z) = (z-1)!$

Mode of beta distribution

The mode is the peak of a distribution. Recall that $Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$. Set

$$\frac{\partial Beta(x|a,b)}{\partial x} = \frac{(a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2}}{B(a,b)} = 0,$$

we have $(a-1)(1-x) = (b-1)x \Rightarrow x = \frac{a-1}{a+b-2}$

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
This gives us a handy trick to manipulate beta distribution

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^{1} xBeta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^{1} xBeta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a+1}(1-x)^{b-1}dx$

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^{1} xBeta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a+1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}$.

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^{1} xBeta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a+1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}$. Thus,

$$Var[X] = E[X^2] - E[X]^2 = \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2}$$
$$= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}$$