Appendix D

Appendix D: Multivariate Gaussian Distribution

D.1 Introduction

Gaussian or Normal distribution is the most important and widely used distribution in engineering. In this chapter, we will present the basic tools to manipulate the multivariate Gaussian distribution.

Just a quick note on convention. Vectors are in bold. Random variables are in upper case and realizations of random variables are in lower case. Therefore, a vector random variable is in bold upper case.

D.2 Probability density function

The probability density function (pdf) of a multivariate Gaussian random variable $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ is given by

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{D.1}$$

For convenience, we will also denote the multivariate Gaussian pdf with $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma)$. Note that $\mathbf{x}$ and $\boldsymbol{\mu}$ are symmetric in $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma)$. We have $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma) = \mathcal{N}(\boldsymbol{\mu};\mathbf{x},\Sigma) = \mathcal{N}(\boldsymbol{\mu}-\mathbf{x};0,\Sigma) = \mathcal{N}(0;\boldsymbol{\mu}-\mathbf{x},\Sigma)$. These equations are trivial but are also very handy at times.

Without confusion in notation, a common convention is to use $p(\mathbf{x})$ to denote $p_{\mathbf{X}}(\mathbf{x})$[1]. Of course, for the above definition to be well-defined, we need to make sure that Σ^{-1} exists. Since Σ is symmetric, the eigenvalues are real and the eigenvectors can be made orthogonal. And as far as all eigenvalues are strictly larger than 0, then Σ^{-1} exists. On the contrary, if there is a zero eigenvalue,

[1]Note that some books may also use $f_{\mathbf{X}}(\mathbf{x})$ instead of $p_{\mathbf{X}}(\mathbf{x})$ for pdfs.

it means that there is no variation along the direction of the corresponding eigenvector. That is, if we project the variable along that eigenvector, the projected value is just a constant (instead of stochastic). So if we ignore these degenerated cases, we can safely assume that Σ^{-1} exists and the pdf is well-defined.

D.3 Marginalization

Consider $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{Z}, \Sigma_\mathbf{Z})$ and let say $\mathbf{X}$ is a segment of $\mathbf{Z}$. That is, $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ for some $\mathbf{Y}$. Then how should $\mathbf{X}$ behave?

We can find the pdf of $\mathbf{X}$ by just marginalizing that of $\mathbf{Z}$. That is

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \tag{D.2}$$

$$= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int \exp\left(-\frac{1}{2} \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu}_\mathbf{X} \\ \mathbf{y} - \boldsymbol{\mu}_\mathbf{Y} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu}_\mathbf{X} \\ \mathbf{y} - \boldsymbol{\mu}_\mathbf{Y} \end{pmatrix} \right) d\mathbf{y}. \tag{D.3}$$

Let us denote Σ^{-1} as Λ, which is usually known to be the precision matrix. And partition both Σ and Λ into $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}$ and $\Lambda = \begin{pmatrix} \Lambda_{\mathbf{XX}} & \Lambda_{\mathbf{XY}} \\ \Lambda_{\mathbf{YX}} & \Lambda_{\mathbf{YY}} \end{pmatrix}$. Then we have

$$p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int \exp\left(-\frac{1}{2} \left[(\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T \Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}) \right.\right. \tag{D.4}$$

$$+ (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YX}} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}) + (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T \Lambda_{\mathbf{XY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}) \tag{D.5}$$

$$\left.\left. + (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}) \right] \right) d\mathbf{y} \tag{D.6}$$

$$= \frac{e^{-\frac{(\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T \Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int \exp\left(-\frac{1}{2} \left[(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YX}} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}) \right.\right. \tag{D.7}$$

$$\left.\left. + (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T \Lambda_{\mathbf{XY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}) + (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}) \right] \right) d\mathbf{y} \tag{D.8}$$

To proceed, we use the "completing square" trick that probably one learns in high school. Basically, for a quadratic expression $ax^2 + 2bx + c$, we may rewrite it as $a(x^2 + 2\frac{b}{a}x) + c = a(x + \frac{b}{a})^2 + c - \frac{b^2}{a}$. By doing that, we immediately can see that the minimum (assuming a is positive) of $ax^2 + 2bx + c$ is $c - \frac{b^2}{a}$ and occurs when $x = -\frac{b}{a}$.

Now let's apply the completing square trick on $(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YX}} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}) + (\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T \Lambda_{\mathbf{XY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}) + (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})^T \Lambda_{\mathbf{YY}} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y})$. For the ease of exposition, let us denote $\tilde{\mathbf{x}}$ as $\mathbf{x} - \boldsymbol{\mu}_\mathbf{X}$ and $\tilde{\mathbf{y}}$ as $\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}$. We have

$$\tilde{\mathbf{y}}^T \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T \Lambda_{\mathbf{XY}} \tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T \Lambda_{\mathbf{YY}} \tilde{\mathbf{y}} \tag{D.9}$$

$$= (\tilde{\mathbf{y}} + \Lambda_{\mathbf{YY}}^{-1} \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}})^T \Lambda_{\mathbf{YY}} (\tilde{\mathbf{y}} + \Lambda_{\mathbf{YY}}^{-1} \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}}) - \tilde{\mathbf{x}}^T \Lambda_{\mathbf{XY}} \Lambda_{\mathbf{YY}}^{-1} \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}}, \tag{D.10}$$

where we use the fact that $\Lambda = \Sigma^{-1}$ is symmetric and so $\Lambda_{\mathbf{XY}} = \Lambda_{\mathbf{YX}}$. Therefore, we have

$$p(\mathbf{x}) = \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}} d\mathbf{y} \quad \text{(D.11)}$$

$$= \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}\right) \quad \text{(D.12)}$$

$$\overset{(a)}{=} \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right) \quad \text{(D.13)}$$

$$\overset{(b)}{=} \frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right) \quad \text{(D.14)}$$

$$= \frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T\Sigma_{\mathbf{XX}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})}{2}\right), \quad \text{(D.15)}$$

where (a) is due to Lemma D.1 and (b) is due to Corollary D.2. In conclusion, $X \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{XX}})$, which is probably what one may expect from the beginning.

Note that for illustrative purpose, we kept track of the normalization factor $\frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}}$ in the above deriation but it was really not necessary. Because we know $p(\mathbf{x})$ should still be a density function and thus will be normalized to one. In the future sections, we will mostly just keep track of the exponent.

D.4 Conditioning

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 section, we have $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{YY}})$. Therefore,

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left[\begin{pmatrix}\tilde{\mathbf{x}}\\ \tilde{\mathbf{y}}\end{pmatrix}^T \Sigma^{-1} \begin{pmatrix}\tilde{\mathbf{x}}\\ \tilde{\mathbf{y}}\end{pmatrix} - \tilde{\mathbf{y}}^T\Sigma_{\mathbf{YY}}^{-1}\tilde{\mathbf{y}}\right]\right) \quad \text{(D.16)}$$

$$\propto \exp\left(-\frac{1}{2}[\tilde{\mathbf{x}}^T\Lambda_{\mathbf{XX}}\tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}}]\right), \quad \text{(D.17)}$$

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

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T\Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right) \quad \text{(D.18)}$$

$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}))^T\Lambda_{\mathbf{XX}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}))\right) \quad \text{(D.19)}$$

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 $\Lambda_{\mathbf{XX}}\Sigma_{\mathbf{XY}} + \Lambda_{\mathbf{XY}}\Sigma_{\mathbf{YY}} = 0$, $\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}} = -\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}$ and from Lemma D.1, we have

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}). \tag{D.20}$$

One can make some intuitive interpretation for the conditioning result above. Let say both $\mathbf{X}$ and $\mathbf{Y}$ are scalar[1]. When the observation of $\mathbf{Y}$ 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 $\mathbf{Y}$ for the 1-D case. This is reasonable as 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 $\mathbf{Y}$ to the variation of $\mathbf{X}$. In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are negatively correlated, the direction of the adjustment will be shifted. 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 ($\mathbf{X}$ and $\mathbf{Y}$ are more correlated).

Corollary D.1. *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.*

Proof. Without loss of generality, we can assume the variables are all zero-mean with unit variance. 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 (D.20), we have

$$\begin{aligned} \Sigma_{\binom{X}{Y}|Z} &= \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{XZ} & \rho_{YZ} \end{pmatrix} \sigma_{YY}^{-1} \begin{pmatrix} \rho_{XZ} \\ \rho_{YZ} \end{pmatrix} \\ &= \begin{pmatrix} 1-\rho_{XZ}^2 & \rho_{XY} - \rho_{XZ}\rho_{YZ} \\ \rho_{XY} - \rho_{XZ}\rho_{YZ} & 1-\rho_{YZ}^2 \end{pmatrix} \end{aligned}$$

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. □

D.5 Product of Gaussian pdfs

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

[1]For the consistency of notations, We will stick with the vector notation for the rest of this section though.

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

$$p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x}) = p(\mathbf{y}_1|\mathbf{x})p(\mathbf{y}_2|\mathbf{x}) \tag{D.21}$$

$$= \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}). \tag{D.22}$$

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.

As in previous sections, the product turns out to be "Gaussian" also. However, unlike previous case, 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.

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) \tag{D.23}$$

$$\propto \exp\left(-\frac{1}{2}[(\mathbf{x}-\mathbf{y}_1)^T\Lambda_{\mathbf{Y}_1}(\mathbf{x}-\mathbf{y}_1) + (\mathbf{x}-\mathbf{y}_2)^T\Lambda_{\mathbf{Y}_2}(\mathbf{x}-\mathbf{y}_2)]\right) \tag{D.24}$$

$$\propto \exp\left(-\frac{1}{2}[\mathbf{x}^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})\mathbf{x} - (\mathbf{y}_2^T\Lambda_{\mathbf{Y}_2} + \mathbf{y}_1^T\Lambda_{\mathbf{Y}_1})\mathbf{x} - \mathbf{x}^T(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}\mathbf{y}_1)]\right) \tag{D.25}$$

$$\propto e^{-\frac{1}{2}[(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))^T(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})(\mathbf{x}-(\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2+\Lambda_{\mathbf{Y}_1}\mathbf{y}_1))]} \tag{D.26}$$

$$\propto \mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}\mathbf{y}_1), (\Lambda_{\mathbf{Y}_2}+\Lambda_{\mathbf{Y}_1})^{-1}). \tag{D.27}$$

Therefore,

$$\begin{aligned}&\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) \\ =&K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1}+\Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}\mathbf{y}_1), (\Lambda_{\mathbf{Y}_2}+\Lambda_{\mathbf{Y}_1})^{-1})\end{aligned} \tag{D.28}$$

for some scaling factor $K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})$ independent of $\mathbf{x}$. And note that we used $\Lambda_{\mathbf{Y}_1} = \Sigma_{\mathbf{Y}_1}^{-1}$ and $\Lambda_{\mathbf{Y}_2} = \Sigma_{\mathbf{Y}_2}^{-1}$ to denote the precision matrices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ above.

Of course, 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 realize that

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) = \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2}) = p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) \tag{D.29}$$

for $\mathbf{X}$ and $\mathbf{Y}_1$ to be conditionally independent given $\mathbf{Y}_2$ with the setup as shown in Figure D.1.

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 Figure D.1,

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1}) \tag{D.30}$$

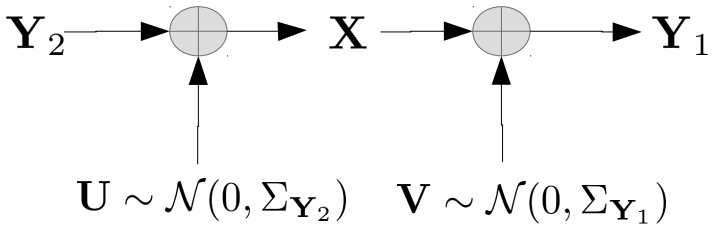

Figure D.1: The conditional pdf $p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2) = p(\mathbf{y}_1|\mathbf{x})p(\mathbf{x}|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2})$ if $\mathbf{X} = \mathbf{U} + \mathbf{Y}_2$ and $\mathbf{Y}_1 = \mathbf{V} + \mathbf{X}$, where $\mathbf{U} \sim (0, \Sigma_{\mathbf{Y}_2})$ is independent of $\mathbf{Y}_2$ and $\mathbf{V} \sim (0, \Sigma_{\mathbf{Y}_1})$ is independent of $\mathbf{X}$.

but from (D.28),

$$\int p(\mathbf{y}_1, \mathbf{x}|\mathbf{y}_2)d\mathbf{x} = \int \mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})d\mathbf{x} \tag{D.31}$$

$$= \int K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})d\mathbf{x} \tag{D.32}$$

$$= K(\mathbf{y}_1, \mathbf{y}_2, \Sigma_{\mathbf{Y}_1}, \Sigma_{\mathbf{Y}_2}). \tag{D.33}$$

In summary,

$$\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2})$$
$$= \mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}). \tag{D.34}$$

Let us try to interpret the product as the overall likelihood after making two observations. For simplicity, let us also assume that $\mathbf{X}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are all scaler[1]. 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 weigthed average of observations $\mathbf{y}_2$ and $\mathbf{y}_1$. And the weight is higher when the precision $\Lambda_{\mathbf{Y}_2}$ or $\Lambda_{\mathbf{Y}_1}$ is larger. And the overall variance $(\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}$ is always smaller than the individual variance $\Sigma_{\mathbf{Y}_2}$ and $\Sigma_{\mathbf{Y}_1}$. This can be understood since we are more certain with $\mathbf{x}$ after considering both $\mathbf{y}_1$ and $\mathbf{y}_2$. Finally, 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 important when we deal with mixture of Gaussian in Section D.7.

[1]Again, for the consistency of notation, we will keep using the vector convention for the rest of this section.

D.6 Division of Gaussian pdfs

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 (D.34)

$$\begin{aligned}&\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})\\ =&\mathcal{N}(\boldsymbol{\mu}_1;(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1).\end{aligned} \tag{D.35}$$

Therefore,

$$\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}=\frac{\mathcal{N}(\mathbf{x};(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2),(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1,(\Lambda_1-\Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1-\Lambda_2\boldsymbol{\mu}_2);\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})} \tag{D.36}$$

$$=\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},(\Lambda_1-\Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1;\boldsymbol{\mu},\Lambda_2^{-1}+(\Lambda_1-\Lambda_2)^{-1})}, \tag{D.37}$$

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.

D.7 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. And the pdf of S will be $0.4\mathcal{N}(s;5,1)+0.6\mathcal{N}(s;0,1)$ as shown in Figure D.2.

The main limitation of Gaussian distribution is that it is unimodal. By mixing Gaussian pdfs of different means, mixture of Gaussian pdfs are multimodal and can virtually model any pdfs. But there is a computational cost for this extra power. 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:

$$p(y_1|x)=0.6\mathcal{N}(x;0,1)+0.4\mathcal{N}(x;5,1); \tag{D.38}$$

$$p(y_2|x)=0.5\mathcal{N}(x;-2,1)+0.5\mathcal{N}(x;4,1). \tag{D.39}$$

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, the overall likelihood $p(y_1,y_2|x)$ just equal to the

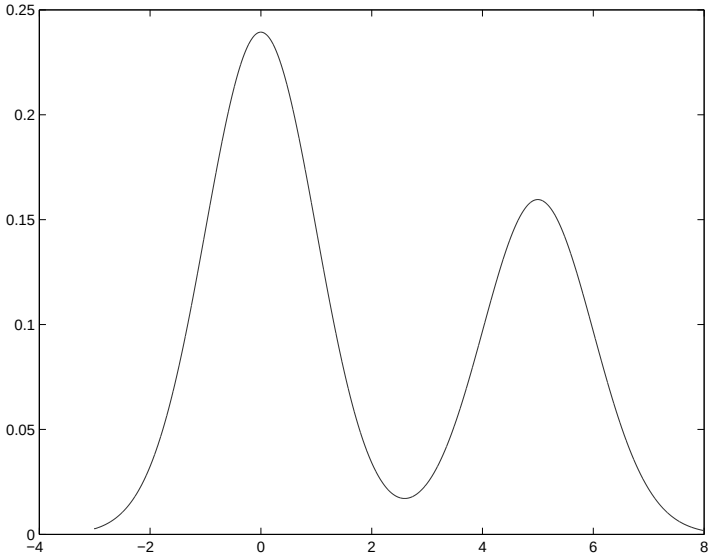

Figure D.2: The pdf of a mixture of Gaussians $(0.4\mathcal{N}(5,1) + 0.6\mathcal{N}(0,1))$

product of likelihoods $p(y_1|x)p(y_2|x)$. That is,

$$p(y_1, y_2|x) = (0.6\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 5, 1))(0.5\mathcal{N}(x; -2, 1) + 0.5\mathcal{N}(x; 4, 1)) \tag{D.40}$$

$$\begin{aligned} &= 0.3\mathcal{N}(x; 0, 1)\mathcal{N}(x; -2, 1) + 0.2\mathcal{N}(x; 5, 1)\mathcal{N}(x; -2, 1) \\ &\quad + 0.3\mathcal{N}(x; 0, 1)\mathcal{N}(x; 4, 1) + 0.2\mathcal{N}(x; 5, 1)\mathcal{N}(x; 4, 1). \end{aligned} \tag{D.41}$$

This involves computing products of Gaussians but we have learned it in previous sections. Using (D.34),

$$\begin{aligned} p(y_1, y_2|x) = &\ 0.3\mathcal{N}(-2; 0, 2)\mathcal{N}(x; -1, 0.5) + 0.2\mathcal{N}(-2; 5, 2)\mathcal{N}(x; 1.5, 0.5) \\ &+ 0.3\mathcal{N}(4; 0, 2)\mathcal{N}(x; 2, 0.5) + 0.2\mathcal{N}(4; 5, 2)\mathcal{N}(x; 4.5, 0.5). \end{aligned} \tag{D.42}$$

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

D.7.1 Reduce Number of Components in Gaussian Mixtures

Let say we have a likelihood of x given an observation is a mixture of two Gaussians as just discussed. And we have n such similar observations. The overall likelihood will be a mixture of 2^n Gassians! 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. For instance, in our previous numerical example, if we continue

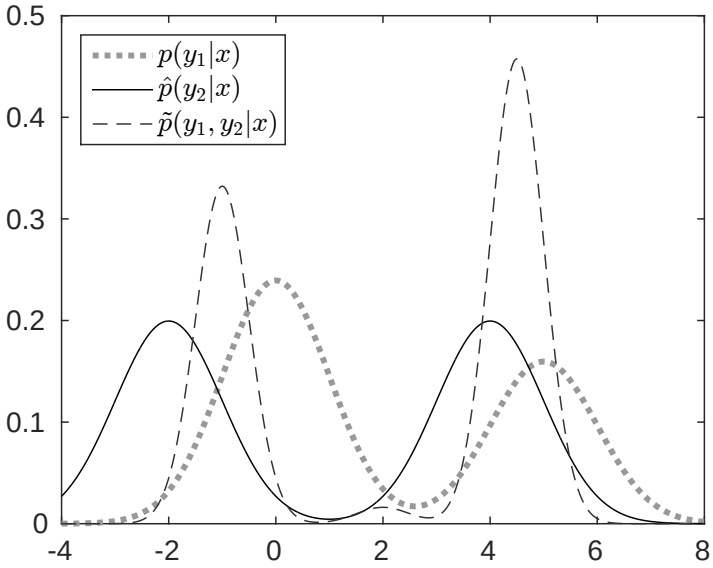

Figure D.3: Likelihood functions: $p(y_1|x) = 0.6\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 5, 1), p(y_2|x) = 0.5\mathcal{N}(x; -2, 1) + 0.5\mathcal{N}(x; 4, 1), p(y_1, y_2|x) = p(y_1|x)p(y_2|x) = 0.4163\mathcal{N}(x; -1, 0.5) + 3.5234 \times 10^{-6}\mathcal{N}(x; 1.5, 0.5) + 0.0202\mathcal{N}(x; 2, 0.5) + 0.5734\mathcal{N}(x; 4.5, 0.5)$.

our numerical computation in (D.42), we have

$$p(y_1, y_2|x) = 0.4163\mathcal{N}(x; -1, 0.5) + 3.5234 \times 10^{-6}\mathcal{N}(x; 1.5, 0.5) + 0.0202\mathcal{N}(x; 2, 0.5) + 0.5734\mathcal{N}(x; 4.5, 0.5). \quad \text{(D.43)}$$

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 Figure D.3. 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. For example, consider

$$p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1). \quad \text{(D.44)}$$

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

$$\hat{p}(x) = 1/6\mathcal{N}(x; 0.2, 1) + 5/6\mathcal{N}(x; 5, 1), \quad \text{(D.45)}$$

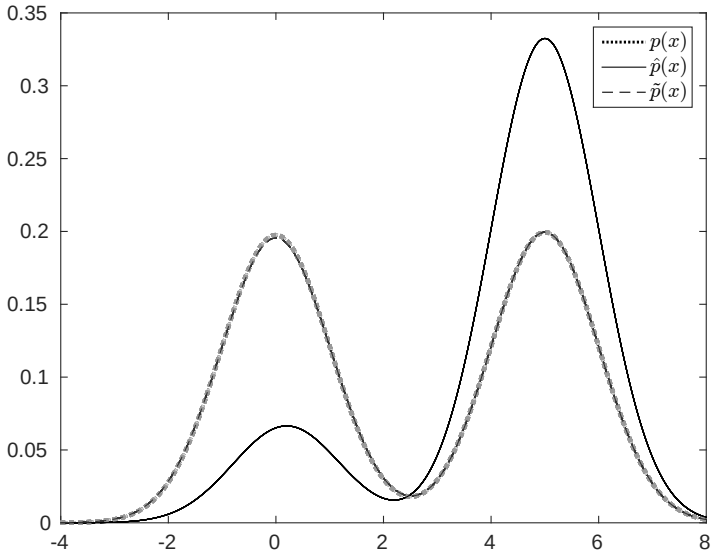

Figure D.4: Approximate $p(x) = 0.1\mathcal{N}(x; -0.2, 1) + 0.1\mathcal{N}(x; -0.1, 1) + 0.1\mathcal{N}(x; 0, 1) + 0.1\mathcal{N}(x; 0.1, 1) + 0.1\mathcal{N}(x; 0.2, 1) + 0.5\mathcal{N}(x; 5, 1)$ by discarding smallest weight components ($\hat{p}(x) = 1/6\mathcal{N}(x; 0.2, 1) + 5/6\mathcal{N}(x; 5, 1)$) and by merging similar components ($\tilde{p}(x) = 0.5\mathcal{N}(x; 0, 1.02) + 0.5\mathcal{N}(x; 5, 1)$). The latter approximation does so well that $p(x)$ and $\tilde{p}(x)$ essentially overlap each other.

which is significantly different from $p(x)$ as shown in Figure D.4.

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, as one can see from Figure D.4, the first five components are so similar that their combined contribution can be accurately modeled as one Gaussian. So one can get a much more accurate approximation by merging these components rather than discarding them. Such approximation $\tilde{p}(x)$ is also illustrated in Figure D.4. However, to successfully obtain such approximation $\tilde{p}(x)$, we have to answer two questions: which components to merge? And how to merge them? We will address these in the following [106].

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

$$\langle p(\mathbf{x}), q(\mathbf{x}) \rangle = \int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}. \tag{D.46}$$

Such inner product is well defined, in particular $\langle p(\mathbf{x}), p(\mathbf{x})\rangle \geq 0$. Therefore, by Cauchy-Schwartz inequality,

$$\frac{\langle p(\mathbf{x}), q(\mathbf{x})\rangle}{\sqrt{\langle p(\mathbf{x}), p(\mathbf{x})\rangle\langle q(\mathbf{x}), q(\mathbf{x})\rangle}} = \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}} \leq 1. \tag{D.47}$$

Moreover, the latter equality holds only when $p(\mathbf{x}) = q(\mathbf{x})$. This suggests a very reasonable similarity measure between two pdfs. Let's define

$$Sim(p(\mathbf{x}), q(\mathbf{x})) \triangleq \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}}. \tag{D.48}$$

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

$$Sim(\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p), \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)) = \frac{\mathcal{N}(\boldsymbol{\mu}_p; \boldsymbol{\mu}_q, \Sigma_p + \Sigma_q)}{\sqrt{\mathcal{N}(0; 0, 2\Sigma_p)\mathcal{N}(0; 0, 2\Sigma_q)}}, \tag{D.49}$$

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? First of all, the combined component obviously will have weight equal to the combined weight $\sum_{i=1}^n w_i$. And its mean will simply be $\sum_{i=1}^n \hat{w}_i \boldsymbol{\mu}_i$, where $\hat{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$.

It is tempting to write the combined covariance as $\sum_{i=1}^n \hat{w}_i \Sigma_i$. However, the covariance is more than that. Because the sum only count 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 the combined covariance Σ given by

$$\Sigma = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T \tag{D.50}$$

$$= \sum_{i=1}^n \hat{w}_i(\Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) - \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i\hat{w}_j\boldsymbol{\mu}_i\boldsymbol{\mu}_j^T. \tag{D.51}$$

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 in Figure D.4, where the approximate pdf is virtually indistinguishable from the original.

[43, 44, 42, 129, 41]

D.8 Summary

Below we assume $\mathbf{Y}_2 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\mathbf{Y}_2 = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_\mathbf{X} \\ \boldsymbol{\mu}_\mathbf{Y} \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}$, $\Sigma^{-1} = \Lambda = \begin{pmatrix} \Lambda_{\mathbf{XX}} & \Lambda_{\mathbf{XY}} \\ \Lambda_{\mathbf{YX}} & \Lambda_{\mathbf{YY}} \end{pmatrix}$.

Marginal pdf of X:

$$\boxed{\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \Sigma_{\mathbf{XX}})} \tag{D.52}$$

Conditional pdf of X given observation y:

$$\boxed{\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})} \tag{D.53}$$

Product of Gaussian pdfs:

$$\boxed{\begin{aligned} &\mathcal{N}(\mathbf{x}; \mathbf{y}_1, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2}) \\ =&\mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}\mathbf{y}_1), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1}) \end{aligned}} \tag{D.54}$$

Division of Gaussian pdfs:

$$\boxed{\frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1)}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)} = \frac{\mathcal{N}(\mathbf{x}; (\Lambda_1 - \Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1 - \Lambda_2\boldsymbol{\mu}_2), (\Lambda_1 - \Lambda_2)^{-1})}{\mathcal{N}(\boldsymbol{\mu}_1, (\Lambda_1 - \Lambda_2)^{-1}(\Lambda_1\boldsymbol{\mu}_1 - \Lambda_2\boldsymbol{\mu}_2); \Lambda_2^{-1} + (\Lambda_1 - \Lambda_2)^{-1})}} \tag{D.55}$$

Measure Similarity between Two Gaussian pdfs:

$$\boxed{Sim(\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p), \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)) = \frac{\mathcal{N}(\boldsymbol{\mu}_p; \boldsymbol{\mu}_q, \Sigma_p + \Sigma_q)}{\sqrt{\mathcal{N}(0; 0, 2\Sigma_p)\mathcal{N}(0; 0, 2\Sigma_q)}}} \tag{D.56}$$

Merging n Gaussian Components in a Mixture:

Merging 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 $w_1, w_2, \cdots, w_n$. And let the combined weight and combined component be w and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Then, w, $\boldsymbol{\mu}$, and Σ are given by the following:

$$\boxed{\begin{aligned} w &= \sum_{i=1}^{n} w_i && \text{(D.57)} \\ \boldsymbol{\mu} &= \sum_{i=1}^{n} \frac{w_i}{w}\boldsymbol{\mu}_i && \text{(D.58)} \\ \Sigma &= \frac{1}{w}\sum_{i=1}^{n} w_i(\Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) - \frac{1}{w^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \boldsymbol{\mu}_i\boldsymbol{\mu}_j^T && \text{(D.59)} \end{aligned}}$$

Appendix: Matrix Equations

Throughout this section, we assume $\Sigma^{-1} = \Lambda$, $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}$, and $\Lambda = \begin{pmatrix} \Lambda_{\mathbf{XX}} & \Lambda_{\mathbf{XY}} \\ \Lambda_{\mathbf{YX}} & \Lambda_{\mathbf{YY}} \end{pmatrix}$. And note that many of the equations have "symmetry" and other similar forms hold as well. For example, apparently we also have $\Sigma_{\mathbf{YY}} = \Lambda_{\mathbf{YY}} - \Lambda_{\mathbf{YX}}\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}$ from Lemma D.1. Without loss of generality, we simply pick one arbritary form for each lemma in this section.

Lemma D.1. $\Sigma_{\mathbf{XX}}^{-1} = \Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}$

Proof. Since $\Lambda = \Sigma^{-1}$, we have $\Sigma_{\mathbf{XX}}\Lambda_{\mathbf{XY}} + \Sigma_{\mathbf{XY}}\Lambda_{\mathbf{YY}} = 0$ and $\Sigma_{\mathbf{XX}}\Lambda_{\mathbf{XX}} + \Sigma_{\mathbf{XY}}\Lambda_{\mathbf{YX}} = I$. Insert an identity into the latter equation, we have $\Sigma_{\mathbf{XX}}\Lambda_{\mathbf{XX}} + \Sigma_{\mathbf{XY}}(\Lambda_{\mathbf{YY}}\Lambda_{\mathbf{YY}}^{-1})\Lambda_{\mathbf{YX}} = \Sigma_{\mathbf{XX}}\Lambda_{\mathbf{XX}} - (\Sigma_{\mathbf{XX}}\Lambda_{\mathbf{XY}})\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}} = \Sigma_{\mathbf{XX}}(\Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}) = I$. □

Lemma D.2. $\det(\Sigma) = \det(\Sigma_{\mathbf{YY}})\det(\Lambda_{\mathbf{XX}}^{-1})$

Proof.

$$\det(\Sigma) = \det\begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix} \tag{D.60}$$

$$= \det\left(\begin{pmatrix} I & 0 \\ 0 & \Sigma_{\mathbf{YY}} \end{pmatrix}\begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} & I \end{pmatrix}\right) \tag{D.61}$$

$$= \det\left(\begin{pmatrix} I & 0 \\ 0 & \Sigma_{\mathbf{YY}} \end{pmatrix}\begin{pmatrix} I & \Sigma_{\mathbf{XY}} \\ 0 & I \end{pmatrix}\begin{pmatrix} \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} & 0 \\ \Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} & I \end{pmatrix}\right) \tag{D.62}$$

$$= \det\begin{pmatrix} I & 0 \\ 0 & \Sigma_{\mathbf{YY}} \end{pmatrix}\det\begin{pmatrix} I & \Sigma_{\mathbf{XY}} \\ 0 & I \end{pmatrix}\det\begin{pmatrix} \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} & 0 \\ \Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} & I \end{pmatrix} \tag{D.63}$$

$$= \det\Sigma_{\mathbf{YY}}\det(\Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}) \tag{D.64}$$

$$= \det\Sigma_{\mathbf{YY}}\det\Lambda_{\mathbf{XX}}^{-1}, \tag{D.65}$$

where the last equality is from Lemma D.1. □

Note that since the width (height) of Σ is equal to the sum of the widths of $\Sigma_{\mathbf{XX}}$ and $\Sigma_{\mathbf{YY}}$. The equation below follows immediately.

Corollary D.2. $\det(a\Sigma) = \det(a\Sigma_{\mathbf{YY}})\det(a\Lambda_{\mathbf{XX}}^{-1})$ *for any constant* a.