Lecture Notes on the Gaussian Distribution

Hairong Qi

The Gaussian distribution is also referred to as the *normal distribution* or the *bell curve distribution* for its bell-shaped density curve. There's a saying that within the image processing and computer vision area, you can answer all questions asked using a Gaussian. The Gaussian distribution is also the most popularly used distribution model in the field of pattern recognition. So let's take a closer look at it.

1 The Definition

The formula for a d-dimensional Gaussian probability distribution is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}{2}) \tag{1}$$

where $\mathbf{x}$ is a d-element column vector of variables along each dimension, μ is the mean vector, calculated by

$$\mu = E[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x}$$

and Σ is the $d \times d$ covariance matrix, calculated by

$$\Sigma = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu^{\mathrm{T}}] = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}p(\mathbf{x})d\mathbf{x}$$

with the following form.

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} \tag{2}$$

The covariance matrix is always symmetric and positive semidefinite, where *positive semidefinite* means that for all non-zero $\mathbf{x} \in R^d$, $\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x} \geq 0$. We normally only deal with covariance matrices that are *positive definite* where for all non-zero $\mathbf{x} \in R^d$, $\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x} > 0$, such that the determinant $|\Sigma|$ will be strictly positive. The diagonal elements σ_{ii} are the variances of the respective x_i, i.e., σ_i^2, and the off-diagonal elements, σ_{ij}, are the covariances of x_i and x_j. If the variables along each dimension are statistically independent, then $\sigma_{ij} = 0$, and we would have a diagonal covariance matrix,

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix} \tag{3}$$

If the covariances along each dimension is the same, then we'll have an identify matrix multiplied by a scalar,

$$\sigma^2 I \tag{4}$$

With Eq. 4, the determinant of Σ becomes

$$|\Sigma| = \sigma^{2d} \tag{5}$$

and the inverse of Σ becomes

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix} \tag{6}$$

For 2-d Gaussian where $d = 2, \mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathrm{T}}, |\Sigma| = \sigma^4$, the formulation becomes

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x_1^2 + x_2^2}{2\sigma^2}) \tag{7}$$

We often denote a Gaussian distribution of Eq. 1 as $p(\mathbf{x}) \sim N(\mu, \Sigma)$.

The following figure illustrates how samples, $\mathbf{x}$'s, drawn from a 2-dimensional Gaussian, distribute. The center of the data cluster is determined by the mean vector and the shape of the cluster is outlined by the covariance matrix. The loci of points of constant density are hyperellipsoids for which the quadratic form $r^2 = (\mathbf{x} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \mu)$ is constant. The principal axes of these hyperellipsoids are given by the eigenvectors of Σ and the eigenvalues determine the lengths of the axes. r^2 is also referred to as the *Mahalanobis distance* from $\mathbf{x}$ to μ.

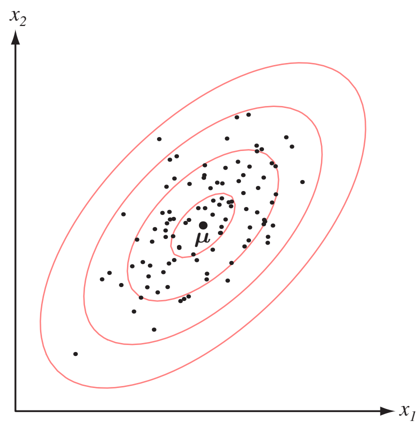

FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

2 The Whitening Transform

The linear transformation of an arbitrary Gaussian distribution will result in another Gaussian distribution. In particular, if A is a $d \times k$ matrix, and $\mathbf{y} = A^{\mathrm{T}}\mathbf{x}$, then $p(\mathbf{y}) \sim N(A^{\mathrm{T}}\mu, A^{\mathrm{T}}\Sigma A)$. In the special case where $k = 1$, A becomes a column vector $\mathbf{a}$, then the transformation actually projects $\mathbf{x}$ onto a line in the direction of $\mathbf{a}$.

If $A = \Phi\Lambda^{-1/2}$ where Φ is the matrix with columns the orthonormal eigenvectors of Σ, and Λ the diagonal matrix of the corresponding eigenvalues, then the transformed distribution has covariance matrix equal to the identify matrix. In signal processing, we refer to this process as a *whitening transform* and the corresponding transformation matrix the whitening matrix, A_w. See the following figure for an illustration.

3 The 68-95-99.7 Rule for Gaussian Distributions

The integral of any probability distribution functions (PDF) from $-\infty$ to $+\infty$ is always 1. The Gaussian distribution follows the same rule, that is,

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \tag{8}$$

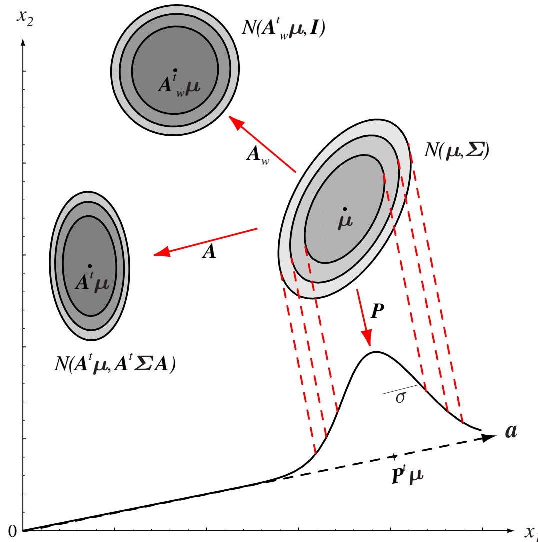

FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection $\mathbf{P}$ onto a line defined by vector $\mathbf{a}$—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

where $p(x)$ is a 1-d Gaussian. Another interpretation is that the area covered underneath the pdf curve is 1.

The 68-95-99.7 rule states that the area covered underneath the pdf curve that is bounded by $x \in [\mu - \sigma, \mu + \sigma]$ is 68% of the entire area (or 1); for $x \in [\mu - 2\sigma, \mu + 2\sigma]$, the area portion is 95%; and for $x \in [\mu - 3\sigma, \mu + 3\sigma]$, the area portion is 99.7%. That is, for the case of zero mean,

$$\begin{array}{l} \int_{-\sigma}^{\sigma} p(x) = 0.68 \\ \int_{-2\sigma}^{2\sigma} p(x) = 0.95 \\ \int_{-3\sigma}^{3\sigma} p(x) = 0.997 \end{array} \tag{9}$$

See the following figure for an illustration.

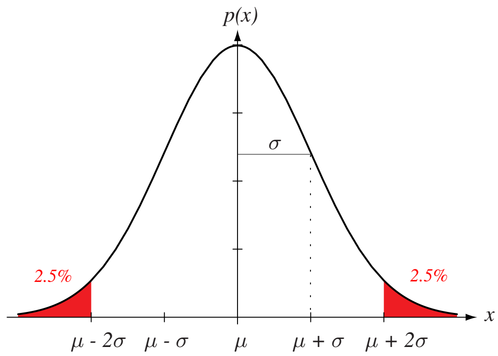

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

4 The Gaussian Blur Kernel

Because of the low-pass nature of the Gaussian, it becomes a natual choice for the construction of a weighted average filter in either the spatial domain or the frequency domain, as the Fourier transform of Gaussian is still a Gaussian. We can create a Gaussian average mask based on Eq. 7 with (x, y) taken from the corresponding coordinates of the mask. Assume the center of the mask has a coordinate of $(0, 0)$, a 3×3 mask can then be constructed by

$$\frac{1}{2\pi\sigma^2}\begin{bmatrix} \exp(-\frac{2}{2\sigma^2}) & \exp(-\frac{1}{2\sigma^2}) & \exp(-\frac{2}{2\sigma^2}) \\ \exp(-\frac{1}{2\sigma^2}) & 1 & \exp(-\frac{1}{2\sigma^2}) \\ \exp(-\frac{2}{2\sigma^2}) & \exp(-\frac{1}{2\sigma^2}) & \exp(-\frac{2}{2\sigma^2}) \end{bmatrix} \tag{10}$$

based on the following coordinate pattern

$$\begin{bmatrix} (-1,-1) & (-1,0) & (-1,1) \\ (0,-1) & (0,0) & (0,1) \\ (1,-1) & (1,0) & (1,1) \end{bmatrix}$$

According to Eq. 10, a typical 3×3 Gaussian mask

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

is generated with $\sigma = 0.85$, which is roughly 70% of the entire area covered underneath the Gaussian pdf.

Now, let's say you want to generate a 5×5 Gaussian mask that would keep, say, 95% of the content, what would the σ be? Based on the 68-95-99.7 rule, to keep 95% of the content below the Gaussian, x should be within the range of $[-2\sigma, 2\sigma]$, and for a 5×5 kernel, x is between -2 and 2, therefore, $-2\sigma = -2$, which yields $\sigma = 1$. With this σ value, you should be able to generate a 5×5 Gaussian mask.

5 How to Generate Gaussian Noise

TBC