6 Probability Density Functions (PDFs)

In many cases, we wish to handle data that are represented in terms of real-valued random variables, or a real-valued vector $\mathbf{x} = [x_1, x_2, ..., x_n]^T$. Most of the intuitions from discrete variables transfer directly to the continuous case, although there are some subtleties.

We describe the probabilities of a real-valued scalar variable x with a probability density function (PDF), written $p(x)$. Any real-valued function $p(x)$ that satisfies:

$$p(x) \geq 0 \qquad \text{for all } x \tag{1}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{2}$$

is a valid PDF. we will use the convention of upper-case P for discrete probabilities, and lower-case p for PDFs.

With the PDF we can specify the probability that the random variable x falls within a given range:

$$P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} p(x)dx \tag{3}$$

This can be visualized by plotting the curve $p(x)$. Then, to determine the probability that x falls within a given range, we compute the area under the curve for that range.

The PDF can be thought of as the infinite limit of a discrete distribution, i.e., a discrete distribution with an infinite number of possible outcomes. Specifically, suppose we create a discrete distribution with N possible outcomes, each corresponding to a range on the real number line. Then, suppose we increase N towards infinity, so that each outcome shrinks to a single real number. A PDF is defined as the limiting case of this discrete distribution.

There is an important subtlety here. A probability density is *not* a probability per se. For one thing, there is no requirement that $p(x) \leq 1$. Moreover, the probability that x attains any one specific value out of the infinite set of possible values is always zero, e.g. $P(x = 5) = \int_5^5 p(x)dx = 0$ for any PDF $p(x)$. People are sometimes sloppy in referring to $p(x)$ as a probability, but it is not a probability — rather, it is a function that can be used in computing probabilities.

Joint distributions are defined in a natural way. For two variables x and y, the joint PDF $p(x, y)$ defines the probability that (x, y) lies in a given domain $\mathcal{D}$:

$$P((x, y) \in \mathcal{D}) = \int_{(x,y)\in\mathcal{D}} p(x, y)dxdy \tag{4}$$

For example, the probability that a 2D coordinate (x, y) lies in the domain $(0 \leq x \leq 1, 0 \leq y \leq 1)$ is $\int_{0\leq x\leq 1} \int_{0\leq y\leq 1} p(x, y)dxdy$. The PDF over a vector may also be written as a joint PDF of its variables. For example, for a 2D-vector $\mathbf{a} = [x, y]^T$, the PDF $p(\mathbf{a})$ is equivalent to the PDF $p(x, y)$.

Conditional distributions are defined as well; $p(x|\mathbf{A})$ is the PDF over x, if the statement $\mathbf{A}$ is true. This statement may be an expression on a continuous value, e.g. "$y = 5$." As a short-hand, we often just write $p(x|y)$, which provides a PDF for x as a function of the value of y. (It must be the case that $\int p(x|y)dx = 1$, since $p(x|y)$ is a PDF over values of x.)

In general, for all of the rules for manipulating discrete distributions there are analogous rules for continuous distributions:

Probability rules for PDFs:
- $p(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} p(x)dx = 1$
- $P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} p(x)dx$
- **Sum rule:** $\int_{-\infty}^{\infty} p(x)dx = 1$
- **Product rule:** $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$.
- **Marginalization:** $p(y) = \int_{-\infty}^{\infty} p(x, y)dx$
- We can also add conditional information, e.g. $p(y|z) = \int_{-\infty}^{\infty} p(x, y|z)dx$
- **Independence:** Variables x and y are independent if $p(x, y) = p(x)p(y)$.

6.1 Mathematical Expectation, Mean, and Variance

Some very brief definitions of ways to describe a PDF:

Given a function $f(\mathbf{x})$ of a random variable $\mathbf{x}$, the **expected value** of the function with repect to a PDF $p(\mathbf{x})$ is defined as:

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] \equiv \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \,. \tag{5}$$

Intuitively, this is the value that we roughly "expect" $\mathbf{x}$ to have.

The mean $\boldsymbol{\mu}$ of a distribution $p(\mathbf{x})$ is the expected value of $\mathbf{x}$:

$$\boldsymbol{\mu} = \mathbb{E}_{p(\mathbf{x})}[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} \,. \tag{6}$$

The variance of a scalar variable x is the expected squared deviation from the mean:

$$\mathbb{E}_{p(x)}[(x-\mu)^2] = \int (x-\mu)^2 p(x)dx \,. \tag{7}$$

The variance of a distribution tells us how uncertain, or "spread-out" the distribution is. For a very narrow distribution $\mathbb{E}_{p(x)}[(x-\mu)^2]$ will be small.

The **covariance** of a vector $\mathbf{x}$ is a matrix:

$$\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}) = \mathbb{E}_{p(\mathbf{x})}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T p(x)d\mathbf{x} \,. \tag{8}$$

By inspection, we can see that the diagonal entries of the covariance matrix are the variances of the individual entries of the vector:

$$\boldsymbol{\Sigma}_{ii} = \mathrm{var}(x_{ii}) = \mathbb{E}_{p(\mathbf{x})}[(x_i-\mu_i)^2] \,. \tag{9}$$

The off-diagonal terms are covariances:

$$\boldsymbol{\Sigma}_{ij} = \mathrm{cov}(x_i, x_j) = \mathbb{E}_{p(x)}[(x_i-\mu_i)(x_j-\mu_j)] \,, \tag{10}$$

between variables x_i and x_j. If the covariance is a large positive number, then we expect x_i to be larger than μ_i when x_j is larger than μ_j. If the covariance is zero and we know no other information, then knowing $x_i > \mu_i$ does not tell us whether or not it is likely that $x_j > \mu_j$.

One goal of statistics is to infer properties of distributions. In the simplest case, the **sample mean** of a collection of N data points $\mathbf{x}_{1:N}$ is just their average: $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i$. The **sample covariance** of a set of data points is $\frac{1}{N}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$. The covariance of the data points tells us how "spread-out" the data are.

6.2 Uniform Distributions

Perhaps the simplest PDF is the **uniform distribution**. Intuitively, this distribution states that all values within a given range $[x_0, x_1]$ are equally likely. Formally, the uniform distribution on the interval $[x_0, x_1]$ is:

$$p(x) \;=\; \begin{cases} \frac{1}{x_1 - x_0} & \text{if } x_0 \le x \le x_1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

It is easy to see that this is a valid PDF (because $p(x) > 0$ and $\int p(x)dx = 1$).

We can also write this distribution with this alternative notation; i.e.,

$$x|x_0, x_1 \;\sim\; \mathcal{U}(x_0, x_1) \tag{12}$$

Equations (11) and (12) are equivalent. The latter simply says that x is distributed uniformly in the range x_0 and x_1, and it is impossible that x lies outside that range. The mean of a uniform distribution $\mathcal{U}(x_0, x_1)$ is $(x_1 + x_0)/2$. The variance is $(x_1 - x_0)^2/12$.

6.3 Gaussian Distributions

Arguably the single most important PDF is the **Normal** (a.k.a., **Gaussian**) distribution. Among the reasons for its popularity are that it is theoretically elegant, and arises naturally in a number of situations. It is the distribution that maximizes entropy,[1] and it is also tied to the Central Limit Theorem: Roughly speaking, the distribution of the sum of a number of random variables approaches the Gaussian distribution as that number tends to infinity (Fig. 1).

Perhaps most importantly, the Gaussian has very attractive analytical properties. Gaussians are easy to manipulate, and their form so well understood, that we often assume quantities are Gaussian distributed, even though they are not, in order to turn an intractable model, or problem, into something that is easier to work with.

The simplest case is a Gaussian PDF over a scalar variable x:

$$p(x|\mu, \sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) . \tag{13}$$

(The notation $\exp(a)$ is the same as e^a). The 1D Gaussian has two parameters, namely, the mean μ, and the variance σ^2. The mean specifies the center of the distribution, and the variance tells us how "spread-out" the PDF is.

[1] Over distributions defined on the real line with fixed variance.

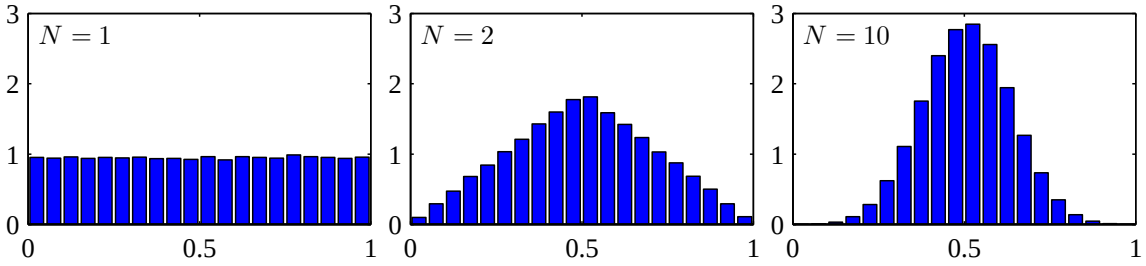

Figure 1: Histogram plots of the mean of N uniformly distributed numbers for various values of N. The effect of the Central Limit Theorem is seen: as N increases, the distribution becomes more Gaussian. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

The PDF for a d-dimensional random vector $\mathbf{x}$, the elements of which are jointly distributed according to a Gaussian denity function, is specified in terms of a mean vector, $\boldsymbol{\mu}\mathbb{R}^d$, and a $d \times d$ covariance matrix, $\boldsymbol{\Sigma}$:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \;, \tag{14}$$

where $|A|$ denotes the determinant of matrix A.

An important special case is when the Gaussian is istropic, or rotationally invariant. In this case the PDF is spheroidal in the sense that all points squidistant from the mean have the same density. In this case the covariance matrix can be written as $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ where $\mathbf{I}$ is the identity matrix, so the PDF reduces to.

$$p(\mathbf{x}|\boldsymbol{\mu}, \sigma^2) \;=\; \frac{1}{\sqrt{(2\pi)^d \sigma^{2d}}} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{x}-\boldsymbol{\mu}||^2\right) \;. \tag{15}$$

The Gaussian distribution is used frequently enough that it is useful to denote its PDF in a simple way. We will define a function G to be the Gaussian density function, i.e.,

$$G(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;\equiv\; \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \;, \tag{16}$$

When formulating problems and manipulating PDFs this functional notation will be useful. When we want to specify that a random vector $\mathbf{x}$ has a Gaussian density, with mean μ and covariance $\boldsymbol{\Sigma}$, it is common to use the notation:

$$\mathbf{x} \;\sim\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;. \tag{17}$$

Equations (14) and (17) essentially say the same thing. Equation (17) says that $\mathbf{x}$ is Gaussian, and Equation (14) specifies (evaluates) the density as a function of $\mathbf{x}$. The covariance matrix $\boldsymbol{\Sigma}$ of a Gaussian must be symmetric and positive definite.

6.3.1 Diagonalization

The multi-dimensional Gaussian density is essentially an ellipsoidal density function. That is, its level surfaces of constant density, i.e., points $\mathbf{x}$ satisfying $G(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \alpha$ for some positive

constant α, are ellpsoidal. This results from the quadratic nature of the exponent in Eqn. (14), and is the key to understanding the properties of the Gaussian PDF.

In more detail, suppose for some $\alpha > 0$ we find the surface comprising all points $\mathbf{x}$ such that

$$G(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; \alpha \,. \tag{18}$$

From Eqn. (14) you can take the negative logarithm of both sides of the equation, collect all constant terms (which don't depend on $\mathbf{x}$) on the right hand side, and you obtain a quadratic, i.e.,

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \;=\; c \,, \tag{19}$$

for constant c. This is a quadratic equation with a symmetric positive definite matrix $\boldsymbol{\Sigma}$. In other words, it is the equation of an ellipsoid in d dimensions. As depicted in 2D in Fig. 2, the center of the ellipsoid is $\boldsymbol{\mu}$. The shape of the ellipsoid is naturally expressed in terms of its principal directions ((its major and minor axes in 2D) and its elongation (or radii) along those directions. These properties of the ellipsoid are obtained through the eigenvector diagonalization of the covariance matrix $\boldsymbol{\Sigma}$.

As a reminder, the eigendecomposition of a real-valued symmetric matrix $\boldsymbol{\Sigma}$ yields a set of orthonormal vectors $\mathbf{v}_i$ and scalars λ_i such that

$$\boldsymbol{\Sigma} \mathbf{u}_i \;=\; \lambda_i \mathbf{u}_i \,. \tag{20}$$

Equivalently, if we combine the eigenvalues and eigenvectors into matrices $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_N]$ and $\mathbf{S} = \mathrm{diag}(\lambda_1, ...\lambda_N)$, then we have

$$\boldsymbol{\Sigma} \mathbf{U} \;=\; \mathbf{U} \mathbf{S} \,. \tag{21}$$

Since $\mathbf{U}$ is orthonormal (i.e., $\mathbf{U}^{-1} = \mathbf{U}^T$), it follows that:

$$\boldsymbol{\Sigma} \;=\; \mathbf{U} S \mathbf{U}^T \quad \text{ and } \quad \boldsymbol{\Sigma}^{-1} \;=\; \left(\mathbf{U}\mathbf{S}\mathbf{U}^T\right)^{-1} \;=\; \mathbf{U}\mathbf{S}^{-1}\mathbf{U}^T \,. \tag{22}$$

(If any of these steps are not familiar to you, you should refresh your memory of them.)

And these eigenvectors of $\boldsymbol{\Sigma}$ are the principal axes of the ellipsoid (19). They can be used, along with the mean, to center and align the distribution with the coordinate axes. To this end, let's return to the negative log of the Gaussian (ignoring constant terms), but subsitute the factorization of $\boldsymbol{\Sigma}$ in Eqn. (22) to obtain:

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{U}\mathbf{S}^{-1}\mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}) \;=\; c \,. \tag{23}$$

This suggests that with a change of variables,

$$\mathbf{y} \;=\; \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}) \,, \tag{24}$$

as depicted in 2D in Fig. 2, the quadratic reduces to

$$\frac{1}{2} \mathbf{y}^T \mathbf{S}^{-1} \mathbf{y} \;=\; c \,. \tag{25}$$

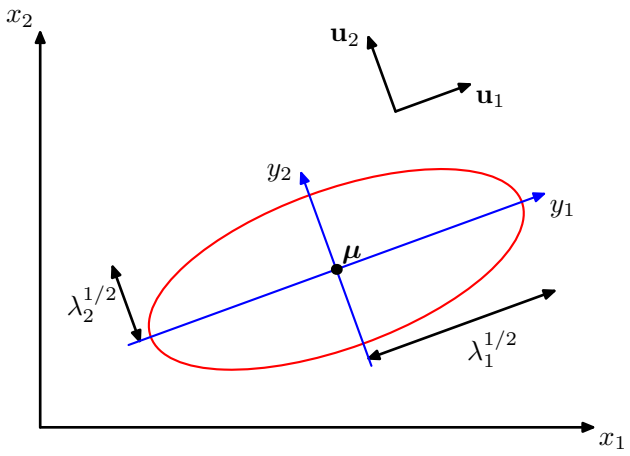

Figure 2: The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

Because the change of variables is orthogonal, $\mathbf{y}$ is also Gaussianly distributed. But $\mathbf{y}$ has zero mean and a diagonal covariance matrix:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}) \; . \tag{26}$$

The distribution remains ellipsoidal, but now the principal axes are aligned with the new coordinate axes. With a diagonal covariance matrix, this also means that the eigenvalues are the variances along corresponding principal axes, and the elements of $\mathbf{y}$ (the random variables y_i), are uncorrelated. Or equivelanetly, the y_i are independent Gaussian random variables, so the joint distribution can be expressed as a product of simpler 1D densities:

$$G(\mathbf{y}; \mathbf{0}, \mathbf{S}) \;=\; \prod_{i=1}^{d} G(y_i; 0, \lambda_i) \; . \tag{27}$$

Another interesting change of variances, often called a whitening transform, is given by

$$\mathbf{z} \;=\; \mathrm{diag}(\lambda_1^{-\frac{1}{2}}, ..., \lambda_N^{-\frac{1}{2}})\, \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{28}$$

This yields a quadratic of the form $\mathbf{z}^T\mathbf{z}/2 = \sum_i z_i^2/2$. Given variables $\mathbf{x}$, we can convert them to the $\mathbf{z}$ representation by applying Eq. (28), and, if all eigenvalues are nonzero, we can convert back by inverting Eq. (28). Hence, we can write our Gaussian in this new coordinate system as[2]:

$$\frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}||\mathbf{z}||^2\right) \;=\; \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right) \tag{29}$$

[2]The normalizing $|\boldsymbol{\Sigma}|$ disappears due to the nature of change-of-variables in PDFs, which we won't discuss here.

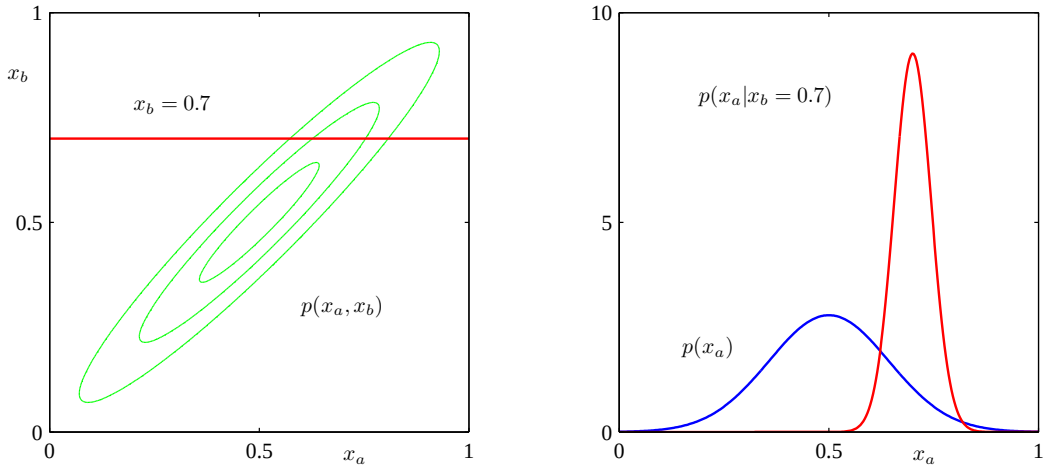

Figure 3: Left: The contours of a Gaussian distribution $p(x_a, x_b)$ over two variables. Right: The marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve). (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

Here, level sets of constant density for the PDF are hyperspheres. Equivalently, $\mathbf{z}$ is Gaussian with an isotropic covariance, and the different elements of $\mathbf{z}$ are uncorrelated. We have transformed the original random vector, whose elements interact (i.e., are mutually dependendent), into a into a simpler Gaussian comprising d independent variables z_i, each mean zero with unit variance.

6.3.2 Marginal and Conditional Distributions

Two important properties of the multi-dimensional Gaussian density are that its marginal and conditional distributions are both also Gaussian. More precisely, if we parition the elements of $\mathbf{x}$ into two subsets, denoted $\mathbf{x}_a$ and $\mathbf{x}_b$, then if the joint distributon is Gaussian (i.e., $\mathbf{x}$ is Gaussian), then the conditional distribution of one set, conditioned on the other, is Gaussian. And the marginal distribution of either set is also Gaussian.

In more detail, suppose that $\mathbf{x}$, along with its mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and its *precision matrix*, defined by $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$, are given in block matrix form as follows:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_a} \\ \boldsymbol{\mu_b} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}, \tag{30}$$

It is straightforward to show that the two marginal PDFs for $\mathbf{x}_a$ and $\mathbf{x}_b$ are Gaussian, i.e.,

$$\mathbf{x}_a \sim \mathcal{N}(\boldsymbol{\mu_a}, \Sigma_{aa}) , \qquad \mathbf{x}_b \sim \mathcal{N}(\boldsymbol{\mu_b}, \Sigma_{bb}) . \tag{31}$$

With a little more work one can also show that the conditional distributions are Gaussian. For example, the conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b$ satisfies

$$\mathbf{x}_a \,|\, \mathbf{x}_b \;\sim\; \mathcal{N}(\boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1}) , \tag{32}$$

where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu_a} - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$. Note that Λ_{aa}^{-1} is not simply Σ_{aa}. Figure 3 shows the marginal and conditional distributions applied to a two-dimensional Gaussian.

Finally, another important property of Gaussian functions is that the product of two Gaussian functions is another Gaussian function (although no longer normalized to integrate to one and hence not a proper density function per se):

$$G(\mathbf{x};\, \boldsymbol{\mu}_1, \Sigma_2)\, G(\mathbf{x};\, \boldsymbol{\mu}_2, \Sigma_2) \;\propto\; G(\mathbf{x};\, \boldsymbol{\mu}, \Sigma)\;, \tag{33}$$

where

$$\boldsymbol{\mu} = \Sigma \left(\Sigma_1^{-1} \boldsymbol{\mu_1} + \Sigma_2^{-1} \boldsymbol{\mu_2} \right), \tag{34}$$

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}\;. \tag{35}$$

Note: be careful to remember that the product of two Gaussian random variables is not Gaussian in general. I.e., this is not the same as the product of two Gaussian functions.

Finally, note that the linear transformation of a Gaussian random variable is also Gaussian. For example, if we apply a transformation such that $\mathbf{y} = A\mathbf{x}$ where $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, we have $\mathbf{y} \sim \mathcal{N}(\mathbf{y}|A\boldsymbol{\mu}, A\Sigma A^T)$.

