Probability

Data as vectors

The usual way you would recieve data is a matrix with rows being observations and columns being features. Every observation can be thought of as a vector, and these vectors are the empirical values for the “random variable” that is a general vector which is a list of such features. For example, for this matrix $\mathbf{X}$:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
$$\begin{bmatrix} 4 & 5 \\ 5 & 6 \end{bmatrix}$$

The random variable $\mathbf{x} \in \mathbb{R}^2$, which is a list of features (scalar random variables), and has the observed values of $[1\ 2]^T$, $[3\ 4]^T$, $[4\ 5]^T$ and $[5\ 6]^T$.

Note : Any random vector $\mathbf{x}$ is a column vector, unless stated otherwise.

The probability of of $\mathbf{x}$ taking a certain value $\mathbf{x_1}$ (in its full domain, not only in the observed values) is denoted as $p_{\mathbf{x}}(\mathbf{x_1})$, or for our convenience, simply as $p(\mathbf{x})$.

If the case that the domain of $\mathbf{x}$ contains infinite elements and the probability for one value is infinitesimal, $p(\mathbf{x})$ refers to the probability distribution function, but I’ll still refer to it as the “probability” as if we are always taking about a discrete dirtribution.

For an event described by multiple random variables, the probability is written as $p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$. In such a setting, if there are less random variables inside the bracket than required to describe the event, then we are talking about the probability of any of the events with the specified random

event, then we are talking about the probability of any of the events with the specified random variable taking the values in the brackets, regardless of what values the other random variable take. For example, if an event is described by three random variables $\mathbf{x}, \mathbf{y}, \mathbf{z}$, then $p(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

Since we can always concatenate multiple random variables describing an event in a single big vector, so we only need to talk about 2 random variables at max.

We use ' | ' to mean "given" it is used to denote conditional probability, defined (yes, I'm using this equation to define the probability, rather than the usual set theory based definition) as

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})}$$

This is the probability of and event having a certain $\mathbf{x}$, given that it has $\mathbf{y}$ as specified.

From here it's easy to derive Baye's rule.

Bayes' rule

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

One good example where it's used is the Gaussian Naive Bayes' Classifier.

A simpler example would be the β-distribution, which is probability distribution of the probability θ of a person winning a match, if we know that he has won α matches and lost β assuming that this probability of winning stays the same in every match. Here we are trying to find $p(\theta \mid \alpha, \beta)$. we

already know that $p(\alpha, \beta \mid \theta) = {}^{\alpha+\beta}C_{\alpha}\, \theta^{\alpha}\, (1-\theta)^{\beta}$ and we can find $p(\alpha, \beta)$ by integrating $p(\alpha, \beta \mid \theta)$ over θ , which gives us ${}^{\alpha+\beta}C_{\alpha}({}^{\alpha+\beta}C_{\beta}(\alpha+\beta+1))^{-1} = (\alpha+\beta-1)^{-1}$ and we assume that $p(\theta) = 1 \ \ \forall\ 0 \leq \theta \leq 1$ and 0 otherwise. This gives us that :

$$p(\theta \mid \alpha, \beta) = \frac{p(\alpha, \beta \mid \theta)\ p(\theta)}{p(\alpha, \beta)} = (\alpha+\beta+1)^{\alpha+\beta}C_{\alpha}\ \theta^{\alpha}\, (1-\theta)^{\beta}$$

This is called the beta distribution.

Using Bayes' rule if you have some prior pdf for $\mathbf{x}$, given more data in the form of vectors $\mathbf{y}_i$ used as row vectors in the matrix $\mathbf{Y}$, you can update $\mathbf{x}$. We want to find the posterior pdf for $\mathbf{x}$, that is $p(\mathbf{x} \mid \mathbf{Y})$. This is calculated as $\frac{p(\mathbf{Y} \mid \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{Y})}$. Since the observations are all independent to each other, thus we can break this as $\prod_i \frac{p(\mathbf{y_i} \mid \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{y_i})}$. This means we need to know $p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{y})}$ first. If $\mathbf{y}$ was in the same domain as $\mathbf{x}$, then we could just calculate $p(\mathbf{y} \mid \mathbf{x})$ to be the delta function in $\mathbf{x}$ centred around $\mathbf{y}$, and then we would get the same for $p(\mathbf{x} \mid \mathbf{y})$. So it makes no sense to take them to be in the same set of features. Although, we can certainly take them to be in a set of dependent features.

Statistically Independent variables :

Two random variables $\mathbf{x}, \mathbf{y}$ are statistically independent iff $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\, p(\mathbf{y})$.

Expected values

For a function $f(\mathbf{x})$ of the random variable $\mathbf{x}$, the expected value is given by :

$$\mathrm{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} f(\mathbf{x}) p(\mathbf{x})$$

It's easy to see that $\mathrm{E}(\mathbf{x})$ is the mean of $\mathbf{x}$.

Note : The definition is only valid for discrete distributions

Covariance and variance

Suppose you you want to know how much are 2 feature x, y similar (linearly dependent), given some data in the form of a matrix $\mathbf{X}$ with these two features present in the observations, then you can just take the dot product of the column vectors of these two features after subtracting the mean from each entry (This is important to do as if these quantities don't vary a lot but have high values all the time, then the covariance of such quantities must intuitively be small). Then, divide by the length of the column vectors since you don't want this value to depend on the size of the data-set. This quantity is called the covariance. Basically,

$$\mathrm{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x[i] - \mathrm{E}(x))(y[i] - \mathrm{E}(y)) = \mathrm{E}((x - \mathrm{E}(x))(y - \mathrm{E}(y)))$$

Here $[i]$ means "in the i^{th} observation" where there are n observations.

And the covariance of a feature with itself is its variance.

$$\mathrm{Var}(x) = \mathrm{Cov}(x, x) = \frac{1}{n} \sum_{i=1}^{n} (x[i] - \mathrm{E}(x))^2 = \mathrm{E}((x - \mathrm{E}(x))^2) = \mathrm{E}(x^2) -$$

$$(\mathrm{E}(x))^2$$

The last equality is obtained by expanding the square inside the summation and using the fact that $\mathrm{E}(x)$ is a constant.

The covariance can be used as a definition of an inner product between two random variables. Then the induced norm would be the square root of the variance. And the angle θ between two random variables x, y would be given by :

$$\cos(\theta) = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)\ \ \mathrm{Var}(y)} = \mathrm{corr}(x, y)$$

This value is called the “Correlation” .

When the random variable $\mathbf{x}$ is a vector, we can express the covariances of all pairs of features x_i, x_j in a covariance matrix, which is basically the “Variance” of $\mathbf{x}$.

$$\mathrm{Var}(\mathbf{x}) = \frac{1}{n}\mathbf{X^T X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}[i](\mathbf{x}[i])^T = \frac{1}{n}\sum_{\mathbf{x}}\mathbf{x}\mathbf{x}^T = \begin{bmatrix} \mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_2) & \dots & \mathrm{Cov}(x_1, x_n) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Cov}(x_2, x_2) & \dots & \mathrm{Cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \mathrm{Cov}(x_n, x_2) & \dots & \mathrm{Cov}(x_n, x_n) \end{bmatrix}$$

Again, $\mathbf{X}$ is the matrix representing the data-set, with rows being $\mathbf{x}[i]^T$

Now, for taking the covariance of two vector valued random variables, $\mathbf{x}, \mathbf{v}$, we do :

Now, for taking the covariance of two vector valued random variables, $\mathbf{x}, \mathbf{y}$, we do :

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{\mathbf{x},\mathbf{y}}(\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{y} - \mathrm{E}(\mathbf{y}))^T =$$
$$\mathrm{E}((\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{y} - \mathrm{E}(\mathbf{y}))^T) =$$
$$\mathrm{E}((\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{y}^T - \mathrm{E}(\mathbf{y})^T)) =$$
$$\mathrm{E}(\mathbf{x}\mathbf{y}^\mathbf{T}) - \mathrm{E}(\mathbf{x}\mathrm{E}(\mathbf{y})^T) - \mathrm{E}(\mathrm{E}(\mathbf{x})\mathbf{y}^T) + \mathrm{E}(\mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{y})^T) =$$
$$\mathrm{E}(\mathbf{x}\mathbf{y}^\mathbf{T}) - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{y})^T$$

Suppose two such random variables are linearly related as $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, then the covariance is :

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathrm{E}(\mathbf{x}(\mathbf{x}^\mathbf{T}\mathbf{A}^\mathbf{T} + \mathbf{b}^\mathbf{T})) - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{A}\mathbf{x} + \mathbf{b})^T = \mathrm{E}(\mathbf{x}\mathbf{x}^\mathbf{T})\mathbf{A}^\mathbf{T} + \mathrm{E}(\mathbf{x})\mathbf{b}^\mathbf{T} -$$
$$\mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x})^T\mathbf{A}^\mathbf{T} - \mathrm{E}(\mathbf{x})\mathbf{b}^\mathbf{T}$$
$$= (\mathrm{E}(\mathbf{x}\mathbf{x}^\mathbf{T}) - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x})^T)\mathbf{A}^\mathbf{T} = \mathrm{Var}(\mathbf{x})\mathbf{A}^\mathbf{T} = \mathbf{\Sigma}\mathbf{A}^\mathbf{T}$$

Here $\mathrm{Var}(\mathbf{x}) = \mathbf{\Sigma}$ is the covariance matrix of the features that $\mathbf{x}$ is composed of.

And thus ,

$$\mathrm{Var}(\mathbf{y}) = \mathrm{Cov}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{y}) = \mathbf{A}\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\mathbf{T}$$

The first equality is easy to prove because $\mathrm{Cov}(\mathbf{x}, \mathbf{y})$ is linear in $\mathbf{x}$ for a fixed $\mathbf{y}$.

Sum of variables

For any two random variables, $\mathbf{x}, \mathbf{y}$, the variable $\mathbf{z} = \mathbf{x} + \mathbf{y}$ is a random variable with $\mathrm{E}(\mathbf{z}) = \mathrm{E}(\mathbf{x}) + \mathrm{E}(\mathbf{y})$ and

$\mathrm{E}(\mathbf{x}) + \mathrm{E}(\mathbf{y})$ and

$$\mathrm{Var}(\mathbf{z}) = \mathrm{Var}(\mathbf{x}) + \mathrm{Var}(\mathbf{y}) + \mathrm{Cov}(\mathbf{x}, \mathbf{y}) + \mathrm{Cov}(\mathbf{y}, \mathbf{x}) .$$

Statistically independent variables are linearly independent

For statistically independent variables $\mathbf{x_i}, \mathbf{x_j}$

$\mathrm{Cov}(\mathbf{x_i}, \mathbf{x_j}) = \mathrm{E}((\mathbf{x_i} - \mathrm{E}(\mathbf{x_i}))((\mathbf{x_j} - \mathrm{E}(\mathbf{x_j}))^T) = \sum_{\mathbf{x_i}, \mathbf{x_j}} (\mathbf{x_i} - \mathrm{E}(\mathbf{x_i}))(\mathbf{x_j} - \mathrm{E}(\mathbf{x_j}))^T p(\mathbf{x_i}, \mathbf{x_j}) = \sum_{\mathbf{x_i}, \mathbf{x_j}} (\mathbf{x_i} - \mathrm{E}(\mathbf{x_i}))(\mathbf{x_j} - \mathrm{E}(\mathbf{x_j}))^T p(\mathbf{x_i}) p(\mathbf{x_j}) = \sum_{\mathbf{x_i}} (\mathbf{x_i} - \mathrm{E}(\mathbf{x_i})) p(\mathbf{x_i}) \sum_{\mathbf{x_j}} (\mathbf{x_j} - \mathrm{E}(\mathbf{x_j}))^T p(\mathbf{x_j}) = \mathrm{E}(\mathbf{x_i} - \mathrm{E}(\mathbf{x_i})) \mathrm{E}(\mathbf{x_j} - \mathrm{E}(\mathbf{x_j}))^T = \mathbf{0}\mathbf{0}^\mathbf{T} = \mathbf{0}$

I.I.D. random variables

Random variables $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ are independent and identically distributed if :

- They are statistically independent , i.e. $p(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}) = p(\mathbf{x_1})\, p(\mathbf{x_2}) \ldots p(\mathbf{x_k}) \forall k \leq n$
- All of their distributions are the same function. This can be summarised as : $\mathbf{x_i} = \mathbf{x_j} \implies p(\mathbf{x_i}) = p(\mathbf{x_j})$.

It's easy to see that in a case like this, the variables are also linearly independent .

This is useful when you want you take n observations with each observation coming from the same distribution and predict beforehand the properties of these observations. For example, the expected value of the computed mean of observations $\mathbf{x_i}$ is the same as the expected value for a single observation, say $\boldsymbol{\mu}$, and the variance is thus $\mathrm{E}((\frac{1}{n}\sum_\mathbf{i} \mathbf{x_i} - \boldsymbol{\mu})(\frac{1}{n}\sum_i \mathbf{x_i} - \boldsymbol{\mu})^T) = \frac{1}{n^2}\mathrm{E}((\sum_i (\mathbf{x_i} - \boldsymbol{\mu_i}))(\sum_i (\mathbf{x_i} - \boldsymbol{\mu_i}))^T) = \frac{1}{n^2}\sum_i \sum_j \mathrm{E}((\mathbf{x_i} - \boldsymbol{\mu_i})(\mathbf{x_j} - \boldsymbol{\mu_j})^\mathbf{T}) = \frac{1}{n^2}\sum_{i,j} \mathrm{Cov}(\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{n^2}\sum_i \mathrm{Cov}(\mathbf{x_i}, \mathbf{x_i}) = \frac{1}{n^2}\sum_i \mathrm{Var}(\mathbf{x_i}) = \frac{1}{n^2}(n\boldsymbol{\Sigma}) = \frac{1}{n}\boldsymbol{\Sigma}$

The 4th equality is because of the linear independence of the variables

The 4th equality is because of the linear independence of the variables.

Gaussian distribution

This distribution is important to the next theorem. Although it lacks physical meaning, it has immense mathematical meaning.

For a random variable $x \in \mathbb{R}$, if it follows the Gaussian distribution

$$p(\mathbf{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

then it has mean μ and variance σ^2 , where $\sigma = \sqrt{\mathrm{Var}(x)}$ is the standard deviation .

To make life simpler, every time we want to say that a variable x follows a Gaussian distribution with mean and variance μ, σ^2, instead of saying it in words, we write $x \backsim N(\mu, \sigma^2)$.

For a vector valued random variable $\mathbf{x} \in \mathbb{R}^n$, made of statistically independent scalar Gaussian random variables $x_1, x_2, \ldots, x_n$ with $x_i \backsim N(\mu_i, \sigma_i^2)$, we have :

$$p(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) = \frac{1}{(2\pi)^n \prod_i \sigma_i^2} \exp\left(-\sum_i \frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$

then if $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$, we have $\prod_i \sigma_i^2 = |\boldsymbol{\Sigma}|$ and $\sum_i \frac{(x_i-\mu_i)^2}{\sigma_i^2} = (\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\, \boldsymbol{\Sigma}^{-\mathbf{1}}\, (\mathbf{x}-\boldsymbol{\mu})$, giving us :

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

$$(2\pi)^n|\boldsymbol{\Sigma}| \qquad z$$

It turns out that this equation is valid even when x_i, x_j are NOT { independent OR Gaussian } , but $\mathbf{x}$ overall is Gaussian .

Marginals and conditionals of Gaussians

Suppose $\mathbf{x}, \mathbf{y}$ are two random variables, which together (concatenated) form a random variable $\mathbf{z} = [\mathbf{x}^\mathbf{T}\ \mathbf{y}^\mathbf{T}]^\mathbf{T}$ which **we know** to be Gaussian, and suppose we want to find out the distribution of $\mathbf{x}$ given the value of $\mathbf{y}$ (Perhaps you want to predict something about a point with a specific $\mathbf{y}$). We do that by taking the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$. This method isn't specific to Gaussian distributions, but here we are interested in the result obtained when we consider Gaussian distribution specifically. Basically, to find $p(\mathbf{x} \mid \mathbf{y})$, you need to find $p(\mathbf{x})$, which we'll prove to be a Gaussian distribution $N(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}})$.

(Say that $\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{b}} = \mathrm{Cov}(\mathbf{a}, \mathbf{b})$ and $\boldsymbol{\mu}_\mathbf{a} = \mathrm{E}(\mathbf{a})$ for any two random variables $\mathbf{a}, \mathbf{b}$.)

Rather than brute forcing our way through complex linear algebra, we'll prove this inductively, by proving this for a scalar valued y, say the last coordinate of $\mathbf{z}$ to get a reduced set of coordinates being in a guassian distribution, and then using the fact that we can do this over an over and over untill we eventually delete an arbitary number of arbitary coordinates, making this true for any general $\mathbf{x}, \mathbf{y}$.

First let's write the covariance and mean of $\mathbf{z}$ in terms of $\mathbf{x}$ and y

$$\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{z}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{y}} \\ \boldsymbol{\Sigma}_{\mathbf{y},\mathbf{x}} & \sigma_y^2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu}_\mathbf{z} = \begin{bmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \mu_y \end{bmatrix}$$

You are free to verify that the inverse of any positive definite symmetric matrix $S = \begin{bmatrix} A & B \\ \end{bmatrix}$

You are free to verify that the inverse of any positive definite symmetric matrix $\begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$ where A, B, D are sub-matrices is also positive definite symmetric, say $\begin{bmatrix} A' & B' \\ B'^T & D' \end{bmatrix}$.

In our case $A = \mathbf{\Sigma_{x,x}},\ B = \mathbf{\Sigma_{x,y}},\ B^T = \mathbf{\Sigma_{y,x}},\ D = \sigma_y^2$. In favour of my sanity, I'll use A, B, D for the linear algebra rather than bold greek symbols with subscripts all the time. Also, let's call $p = \mathbf{x} - \boldsymbol{\mu}_\mathbf{x}$ and $q = y - \mu_y$.

So now we have :

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^n |S|} \exp(-\frac{1}{2} \begin{bmatrix} p^T & q \end{bmatrix} \begin{bmatrix} A' & B' \\ B'^T & D' \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix})$$

It's easy to see that the stuff inside of the exp function (except the -1/2) evaluates to

$$p^T A' p + 2(p^T B')q + D'q^2 = p^T A' p - \frac{(p^T B')^2}{D'} + (q + \frac{p^T B'}{D'})^2 = p^T A' p - p^T \frac{B'^T B'}{D'} p + (q + \frac{p^T B'}{D'})^2$$

Now since $\int_{-\infty}^{\infty} \exp(-\frac{1}{2}(q + \frac{p^T B'}{D'})^2) dq = \sqrt{2\pi}$ and other stuff is constant wrt q, thus we only have to worry about $p^T (A' - \frac{B'^T B'}{D'}) p$ inside the exp function (except the -1/2) after the integration is done. Namely, $p(\mathbf{x})\ \alpha\ \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T \mathbf{M} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}))$ where $\mathbf{M} = A' - \frac{B'^T B'}{D'}$ is some constant square matrix. Now after normalising the expression to get a pdf, we'll just end up getting a normal distribution $\mathbf{x} \backsim N(\boldsymbol{\mu}_\mathbf{x}, \mathbf{M})$. What this means is that $\mathbf{M} = \mathbf{\Sigma_{x,x}^{-1}}$, which isn't surprising because the way it was defined, it's basically A^{-1} . And thus, using the inductive method I proposed, we can say that for any random variable formed by deleting some coordinate**s** (more

than one) of $\mathbf{z}$, say $\mathbf{x}$, has a normal distribution .

Now, moving towards the conditional:

$$p(\mathbf{x}) \ \alpha \ \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T{\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x}))$$

$$p(\mathbf{x},\mathbf{y}) \ \alpha \ \exp(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{z})^T{\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{z}}}^{-1}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{z}))$$

$$\implies p(\mathbf{x} \mid \mathbf{y}) \ \alpha \ \exp(-\frac{1}{2}((\mathbf{z}-\boldsymbol{\mu}_\mathbf{z})^T{\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{z}}}^{-1}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{z}) \ - \\ (\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T{\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})))$$

it's easy to see that the thing inside the exp function (except the -1/2) is a symmetric quadratic form in $\mathbf{x}$, and can thus be expressed as $(\mathbf{x}-\mathbf{a})^\mathbf{T}\mathbf{M}(\mathbf{x}-\mathbf{a})$ for some symmetric matrix $\mathbf{M}$. Thus $p(\mathbf{x}|\mathbf{y})$ is also a normal distribution. To find the parameters of this distribution, you have to get your hands dirty and do a lot of linear algebra involving block matrices. I am going to skip all of that and just give the results to you :

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{y}}\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{y}}^{-\mathbf{1}}(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})$$

$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x},\mathbf{y}}\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{y}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{x}}$$

Change of variables

For probability density functions $p_\mathbf{y}(\mathbf{y})$ and $p_\mathbf{x}(\mathbf{x})$ for variables $\mathbf{x}, \mathbf{y}$ related as $\mathbf{y} = f(\mathbf{x})$, we have $p_\mathbf{y}(\mathbf{y})\, dy_1\, dy_2 \ldots\, dy_n = p_\mathbf{x}(\mathbf{x})\, dx_1\, dx_2\, dx_3 \ldots\, dx_n$

We know that the determinant of the Jacobian of $\mathbf{y}$ wrt $\mathbf{x}$ gives us the factor by which the volume $dx_1\, dx_2\, dx_3 \ldots\, dx_n$ scales to become the volume $dy_1\, dy_2 \ldots\, dy_n$ ..Thus, we can write

$p_{\mathbf{y}}(\mathbf{y})|\mathbf{J}_f(\mathbf{x})| = p_{\mathbf{x}}(\mathbf{x})$.

Affine Transform of a Gaussian variable

Consider $\mathbf{x} \backsim N(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})$. Then, for a variable $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, we have

- $\mathbf{J}_{\mathbf{y}}(\mathbf{x}) = \mathbf{A}$
- $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} + \mathbf{b} \implies \boldsymbol{\mu}_{\mathbf{x}} = \mathbf{A}^{-1}(\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{b})$
- $\mathrm{Var}(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^{\mathbf{T}} \implies \boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{A}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}(\mathbf{A}^{-1})^{\mathbf{T}}$

and thus

$$p_{\mathbf{y}}(\mathbf{y}) = |\mathbf{A}|^{-1} p_{\mathbf{x}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})) =$$

$$\frac{|\mathbf{A}^{-1}|}{(2\pi)^n |\mathbf{A}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}(\mathbf{A}^{-1})^{\mathbf{T}}|} \exp(-\frac{1}{2}(\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}))^{\mathbf{T}}(\mathbf{A}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}(\mathbf{A}^{-1})^{\mathbf{T}})^{-1}(\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})))$$

$$\frac{1}{(2\pi)^n |\boldsymbol{\Sigma}_{\mathbf{y}}|} \exp(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^{\mathbf{T}}(\mathbf{A}^{-1})^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{A}\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})) =$$

$$\frac{1}{(2\pi)^n |\boldsymbol{\Sigma}_{\mathbf{y}}|} \exp(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})) =$$

Thus $\mathbf{y}$ also follows the Gaussian distribution,

namely $\mathbf{y} \backsim N(\mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^{\mathbf{T}})$.

Standard Gaussian

Consider $\mathbf{y} \backsim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This is the result of an affine transformation on $\mathbf{x} \backsim N(\mathbf{0}, \mathbf{I})$ given by $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\mu}$ for some $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^{\mathbf{T}} = \boldsymbol{\Sigma}$. We know how to find a possible $\mathbf{A}$ using the

Cholesky Decomposition .

Thus all Gaussian variables are just Affine Transformations of the Standard Gaussian $\mathbf{x} \backsim N(\mathbf{0}, \mathbf{I})$.