Derivations of the Univariate and Multivariate Normal Density

Alex Francis & Noah Golmant

Berkeley, California, United States

Contents

1. The Univariate Normal Distribution

It is first useful to visit the single variable case; that is, the well-known continuous probability distribution that depends only on a single random variable X. The normal distribution formula is a function of the mean μ and variance σ^2 of the random variable, and is shown below.

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

This model is ubiquitous in applications ranging from Biology, Chemistry, Physics, Computer Science, and the Social Sciences. It's discovery is dated as early as 1738 (perhaps ironically, the name of the distribution is not attributed to it's founder; that is credited, instead, to Abraham de Moivre). But why is this model still so popular, and why has it seemed to gain relevance over time? Why do we use it as a paradigm for cat images and email data? The next section addresses three applications of the normal distribution, and in the process, derives it's formula using elementary techniques.

1.1. The Continuous Approximation of the Binomial Distribution

Any introductory course in probability introduces counting arguments as a way to discuss probability; fundamentally, probability deals in subsets of larger supersets, so being able to count the cardinality of the subset and superset allows us to build a framework for thinking about probability. One of the first-introduced discrete distributions based on counting arguments is the *binomial distribution*, which counts the number of successes (or failures) in n independent trials that each have a probability p of success. The binomial distribution is defined as,

$$\mathbb{P}\{k \text{ successes in } n \text{ trials}\} = \binom{n}{k} p^k (1-p)^{n-k}$$

The argument is simple: the probability of a specific "k success" outcome is $p^k(1-p)^{n-k}$, by the independence of the outcomes. But there are more than just this specific way to achieve k successes in n trials. Therefore, to account for the undercounting problem, we note that we could fill n bins with k successes by using the combinatorial function

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Which completes the argument. It can be shown that a binomial random variable X has expectation $\mathbb{E}(X) = np$, variance $\mathrm{Var}(X) = np(1-p)$ and *mode* (the number of successes with maximal probability) $\lfloor np + p \rfloor$.

In practice, we often want more than just the probability of a specific number of successes. A classical problem in this area of study is factory production. Let's say I own a factory that produces iPads that are either functional or defective. If I produce defective iPads with 2% probability, what's the probability that more than 50 in 10000 units are defective (assuming independence)? This problem is somewhat difficult to solve using the binomial formula, since I have a sum of combinatorial terms, i.e.,

$$\begin{aligned}\mathbb{P}\{\text{More than 50 defective units}\} &= 1 - \mathbb{P}\{\text{0 defective } \cup \ldots \cup \text{ 50 defective}\} \\ &= 1 - \sum_{i=0}^{50} \left(\binom{10000}{i} ((0.2)^i (1-0.2)^{10000-i}) \right)\end{aligned}$$

For larger n, this can be difficult to compute even using software. Therefore, we are motivated to obtain a continuous distribution that approximates the binomial distribution in question, with well-known *quantiles* (the probability of an observation being less than a certain quantity). This leads to the following theorem.

THEOREM 1.1.1 **(The Normal Approximation to the Binomial Distribution)** The continuous approximation to the binomial distribution has the form of the normal density, with $\mu = np$ and $\sigma^2 = np(1-p)$.

Proof. The proof follows the basic ideas of Jim Pitman in *Probability.*[1]

Define the *height* function H as the ratio between the probability of success in bucket k and the probability of success at the mode of the binomial distribution, m, i.e.

$$H(k) = \frac{\mathbb{P}\{k\}}{\mathbb{P}\{m\}}$$

Define the *consecutive heights ratio* function R as the ratio of the heights of a bucket, or number of successes, k and it's predecessor bucket, $k-1$, i.e.

$$R(k) = \frac{H(k)}{H(k-1)} = \frac{P(k)}{P(k-1)} = \frac{n-k+1}{k} \frac{p}{(1-p)}$$

[1] This is the textbook for Statistics 134; Jim Pitman remains a distinguished member of the faculty in the Department of Statistics at UC Berkeley.

For $k > m$, $H(k)$ is the product of $(m - k)$ consecutive heights ratios.

$$H(k) = \frac{\mathbb{P}\{m+1\}}{\mathbb{P}\{m\}} \frac{\mathbb{P}\{m+2\}}{\mathbb{P}\{m+1\}} \cdots \frac{\mathbb{P}\{k\}}{\mathbb{P}\{k-1\}} = \prod_{i=1}^{k-m} R(m+i)$$

And, for $k < m$,

$$H(k) = \frac{\mathbb{P}\{m-1\}}{\mathbb{P}\{m\}} \frac{\mathbb{P}\{m-2\}}{\mathbb{P}\{m-1\}} \cdots \frac{\mathbb{P}\{k\}}{\mathbb{P}\{k+1\}} = \prod_{i=1}^{m-k} \frac{1}{R(m-i)}$$

Denoting the natural logarithm function as log, we use this operator on the products above to turn them into sums.

$$\log H(k) = \begin{cases} \sum_{i=1}^{k-m} \log R(m+i) & k > m \\ -\sum_{i=1}^{m-k} \log R(m-i) & k < m \end{cases}$$

In order to evaluate the sum more explicitly, we express the logarithm of the consecutive heights ratio in a useful form, using some approximations from calculus. Namely, recall that $\log(1+\delta) \approx \delta$ for small δ. Letting $k = m + x = np + x$, for $k > m$,

$$\begin{aligned}
\log R(k) &= \log\left(\frac{n-k+1}{k}\frac{p}{(1-p)}\right) && (1)\\
&= \log\left(\frac{(n-np-p-x+1)p}{(np+p+x)(1-p)}\right) && (2)\\
&\approx \log\left(\frac{(n-np-x)p}{(np+x)(1-p)}\right) && (3)\\
&= \log\left(np(1-p)-px\right) - \log\left(np(1-p)+(1-p)x\right) && (4)\\
&= \log\left(1+\frac{px}{np(1-p)}\right) - \log\left(1+\frac{(1-p)x}{np(1-p)}\right) && (5)\\
&\approx -\frac{px}{np(1-p)} - \frac{(1-p)x}{np(1-p)} && (6)\\
&= \frac{-x}{np(1-p)} && (7)
\end{aligned}$$

There is an analogous formula for $k < m$. I will leave this case as an exercise for the reader for the remainder of the proof. In the above steps, we also used a few “large value” assumptions to obtain the results: from (2) to (3), we assume $np + p \approx np$, which is a reasonable assumption for large n. We also assume, in the same step, that $k + 1 \approx k$, which is reasonable for large k. Now, plugging in to obtain $H(k)$, we obtain the non-normalized height of the normal density,

$$\begin{aligned}
\log H(k) &= \sum_{i=1}^{k-m} \log R(m+i)\\
&\approx \sum_{i=1}^{k-m} \frac{-i}{np(1-p)}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{np(1-p)}\sum_{i=1}^{k-m} i \\
&= -\frac{(k-m)(k-m+1)}{2np(1-p)} \\
&\approx -\frac{(k-m)^2}{2np(1-p)} \\
&\approx -\frac{(k-np)^2}{2np(1-p)} \\
&= -\frac{(k-\mu)^2}{2\sigma^2}
\end{aligned}
$$

Therefore,

$$H(k) = \exp\left(-\frac{(k-\mu)^2}{2\sigma^2}\right) \tag{8}$$

However, the quantity we seek is not the height of the histogram, but the probability density for some arbitrary k, that is,

$$\mathbb{P}\{k\} = H(k)\mathbb{P}(m)$$

But, we know that

$$\sum_{i=0}^{n} H(k) = \frac{1}{\mathbb{P}(m)}\sum_{i=0}^{n}\mathbb{P}\{i\} = \frac{1}{\mathbb{P}(m)}$$

So,

$$\mathbb{P}\{k\} = \frac{H(k)}{\sum_{i=0}^{n} H(k)}$$

Using a continuous argument on the approximation in (8), we take the integral of the function H over all reals and obtain the normalizing constant $\sigma\sqrt{2\pi}$, which completes the proof.

$$
\begin{aligned}
\sum_{i=0}^{n} H(k) &\approx \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}
\end{aligned}
$$

□

This is a powerful result, albeit one with limitations. Specifically, if k or n are small, the approximation is, obviously, much less accurate than we may otherwise desire. Other results in statistical theory over the two centuries since this approximation was discovered deal with the issue of smaller samples, but will not be covered here.

1.2. The Central Limit Theorem

The Central Limit Theorem is a seminal result that places us one step closer to establishing the practical footing that allows us to understand the underlying processes by which data are generated. Formally, the central limit theorem is stated below.

THEOREM 1.2.1 **(The Central Limit Theorem)** Suppose that $X_1, \dots, X_n$ is a finite sequence of independent, identically distributed random variables with common mean $\mathbb{E}(X_i) = \mu$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$. Then, if we let $S_n = \sum_{i=1}^n X_i$, we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq c\right) = \Phi(c) = \int_{-\infty}^{c} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

Proof. The proof relies on the *characteristic function* from probability. The characteristic function of a random variable X is defined to be,

$$\phi(t) = \mathbb{E}(e^{itX})$$

Where $i = \sqrt{-1}$. For a continuous distribution, using the formula for expectation, we have,

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itX} f_X(x) dx$$

This is the Fourier transform of the probability density function. It completely defines the probability density function, and is useful for deriving analytical results about probability distributions. The characteristic function for the univariate normal distribution is computed from the formula,

$$\begin{aligned}
\phi_X(t) &= \int_{-\infty}^{\infty} e^{itX} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma} - \left(\frac{\mu}{\sigma} + it\sigma\right)\right)^2\right) \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma} - \left(\frac{\mu}{\sigma} + it\sigma\right)\right)^2\right) dx
\end{aligned}$$

Recall from single-variable calculus that the integral $\int_{-\infty}^{\infty} e^{-z^2/a^2} = a\sqrt{\pi}$. Then, using the technique of u-substitution, we have,

$$\phi_X(t) = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right) \tag{9}$$

It is a fact of characteristic functions that the characteristic function of a sum of independent two random variable $Z = X + Y$ is the product of the characteristic functions. That is,

$$\phi_Z(t) = \phi_X(t)\phi_Y(t)$$

The proof of this little lemma is rather simple, so it is included below.

$$\begin{aligned}
\phi_Z(t) &= \mathbb{E}(e^{it(X+Y)}) \\
&= \mathbb{E}(e^{itX}e^{itY}) \\
&= \mathbb{E}(e^{itX})\mathbb{E}(e^{itY})
\end{aligned}$$

$$= \phi_X(t)\phi_Y(t)$$

Then, for the sum of n independent random variables $Z = X_1 + \ldots + X_n$, the characteristic function is,

$$\phi_Z(t) = \prod_{j=1}^{n} \phi_{X_j}(t)$$

For the central limit theorem, we specifically examine the random variable $S_n = X_1 + \ldots + X_n$, when the X_j are independent, and identically distributed. Denote the random variable,

$$Z_j = X_j - \mu$$

It turns out that the characteristic function of the random variable,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(X_1 - \mu) + \ldots + (X_n - \mu)}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} Z_j$$

Is easier to study than the sum S_n, so we consider the characteristic function of this random variable below.

$$\begin{aligned} \phi_{(S_n-\mu)/(\sigma\sqrt{n})}(t) &= \prod_{j=1}^{n} \left(\int_{-\infty}^{\infty} \exp\left(\frac{it(x_j - \mu)}{\sigma\sqrt{n}} \right) p(x_j) dx_j \right) \\ &= \left(\phi_{Z_i}\left(\frac{t}{\sigma\sqrt{n}} \right) \right)^n \end{aligned}$$

Using the MacLaurin series for e^x, and ignoring the normalizing constant for a moment, we have that,

$$\begin{aligned} \phi_{Z_1}(t) &= \int_{-\infty}^{\infty} \exp\left(it(x_1 - \mu)\right) p(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \sum_{j=0}^{\infty} \frac{(it(x_1 - \mu))^j}{j!} p(x_1) dx_1 \\ &= \sum_{j=0}^{\infty} \int_{-\infty}^{\infty} \frac{(it(x_1 - \mu))^j}{j!} p(x_1) dx_1 \\ &= \sum_{j=0}^{\infty} \frac{(it)^j}{j!} \mathbb{E}((X_1 - \mu)^j) \end{aligned}$$

Where the definition of expectation was used in the ultimate step. Incorporating the normalizing constant,

$$\phi_{Z_i}\left(\frac{t}{\sigma\sqrt{n}} \right) = \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{it}{\sigma\sqrt{n}} \right)^j \mathbb{E}((X_1 - \mu)^j)$$

Finally, we take $n \to \infty$ in order to get the limit theorem in question. Note that $\mathbb{E}(X_1 - \mu) = 0$ and that $\mathbb{E}((X_1 - \mu)^2) = \text{Var}(X_1 - \mu) = \sigma^2$. Critically, all terms with order higher than two in the MacLaurin polynomial, which we denote ϵ_n go to zero as $n \to \infty$. Then,

$$\lim_{n\to\infty} \phi_{(S_n-\mu)/(\sigma\sqrt{n})}(t) = \lim_{n\to\infty} \left(\phi_{Z_i}\left(\frac{t}{\sigma\sqrt{n}} \right) \right)^n$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \left(\sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{it}{\sigma\sqrt{n}} \right)^j \mathbb{E}((X_1 - \mu)^j) \right)^n \\
&= \lim_{n\to\infty} \left(1 - \frac{t^2}{2n} + \epsilon_n \right)^n \\
&= \lim_{n\to\infty} \left(1 - \frac{t^2}{2n} \right)^n \\
&= \exp\left(-\frac{t^2}{2} \right) \qquad (10)
\end{aligned}
$$

Note that (10) is, by way of the derivation of the characteristic function for the normal density in (9),

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1)$$

Which completes the proof. We could also use the *inverse Fourier transform* to compute the closed form probability density function, which would be the probability density function of the $\mathcal{N}(0,1)$ random variable, but this is left as an exercise for the reader. □

The idea of the Central Limit Theorem is unintuitive. It states that, even for the most poorly-behaved of distributions (non-symmetric, bimodal, etc.), the distribution of the sum is normal. This result has in some sense defined classical statistical inference. It allows us to develop the idea of a *hypothesis test*, which allows us to compare data to a theoretical distribution from which we hypothesize the data has been drawn, leading to a transition out of probability theory and into the realm of theoretical statistics. It is useful, therefore, for comparing proposed means, variances, and proportions to ones actually drawn from the population, which leads to a variety of applications that span from A/B testing at technology companies to polling results that allow scientists to obtain reasonable predictions about the results of an election (except, it seems, in 2016).[2]

The Central Limit Theorem also has fascinating applications in signal processing. Consider the following (potentially contrived) example. When using a cell phone, I make contact with a single cell tower for which my cell phone is in broadcast range. At each time step t, I send a deterministic (that is, not stochastic/random) signal X to the tower. However, the process is fairly noisy, so the signal that is received by the tower Y is,

$$Y = X + \epsilon$$

Where ϵ is an arbitrary zero-mean noise distribution. Then, assuming independence of signal and noise, we have that,

$$
\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(X) + \mathbb{E}(\epsilon) = X \\
\mathrm{Var}(Y) &= \mathrm{Var}(X + \epsilon)
\end{aligned}
$$

[2] In practice, for the A/B testing example, a variant of classical hypothesis testing called Bayesian Inference is utilized, for technical reasons.

$$
\begin{aligned}
&= \mathrm{Var}(X) + \mathrm{Var}(\epsilon) \\
&= \mathrm{Var}(\epsilon)
\end{aligned}
$$

If $\mathrm{Var}(\epsilon)$ is high, the signal that is received at the tower may be quite noisy. However, say I disperse two hundred tools for measuring cell signals throughout the area, all of which feed data to the single cell tower. Let the measurements be, $Y_1, \ldots, Y_{200}$, with,

$$Y_i = X + \epsilon_i$$

Where ϵ_i is the same error distribution as before. Then, the cell tower can compute,

$$\bar{Y} = \frac{1}{200}\sum_{i=1}^{200} Y_i = X + \frac{1}{200}\sum_{i=1}^{200} \epsilon_i$$

The Central Limit Theorem tells us that this quantity is normally distributed, with expectation and variance,

$$
\begin{aligned}
\mathbb{E}(\bar{Y}) &= \mathbb{E}(X + \frac{1}{200}\sum_{i=1}^{200} \epsilon_i) \\
&= X \\
\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{200}\sum_{i=1}^{200} \epsilon_i\right) \\
&= \frac{1}{200^2} 200\sigma^2 \\
&= \frac{\sigma^2}{200}
\end{aligned}
$$

Therefore, the magic of the central limit theorem allows us to more precisely estimate the signal sent from one electronic device by using arithmetic averages.

1.3. The Noisy Process

The final example in the preceding question leads us to the ultimate and most useful interpretation of the Gaussian distribution, as an error curve. This leads us into a story whose characters are the scientific celebrities of the late 18th and early 19th century. The search for a model for error and imprecision was pioneered by astronomers. Galileo observed in 1632 that inexactness was ubiquitous in measurement, and he conjectured that an error distribution would capture five inherent truths:

1. There is only one number which gives the distance of a star from the center of the earth, the *true distance*.
2. All observations are encumbered with errors, due to the observer, the instruments, and the other observational conditions.
3. All observations are distributed symmetrically about the true value; that is, the errors are distributed symmetrically about zero.
4. Small errors occur more frequently than large errors.

5. The calculated distance is a function of the direct angular observations such that small adjustments of the observations may result in a large adjustment of the distance.

Following the precedent of the philosophy of Galileo, and in light of celestial events nearly 170 years later, a young mathematician by the name of Carl Friedrich Gauss gained fame in Europe for accurately predicting the orbit of the "heavenly body" Ceres.[3] He delineated to the scientific community that he used a "least squares" estimate in order to locate the orbit that best fit the observations. His theory was grounded in the following three assumptions, which resemble the points of Galileo:

1. Small errors are more likely than large errors.
2. For any real number ϵ, the likelihood of errors of magnitudes ϵ and $-\epsilon$ are equal.
3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average (from the "least squares" formulation - it can be shown that the average minimizes the sum of squares).

Based on these observations, Gauss settled on the error curve that we now know as the normal density.

THEOREM 1.3.1 **(The Normal Distribution as an Error Curve)** The probability density for the error curve is,

$$\phi(x) = \frac{h}{\sqrt{\pi}} e^{h^2 x^2}$$

Where h is a non-negative constant that represents the "precision of the measurement process".

Proof. Begin the proof by assuming that we have measurements from a process, with true value m. Denote the measurements of a random process $m_1, \ldots, m_n$, and denote $\phi(x)$ as the probability density function of the random errors. Since the distribution is symmetric, the function is even, so $\phi(x) = \phi(-x)$. Assume that the function $\phi(x)$ is differentiable, and that the derivative is denoted $\phi'(x)$. Assuming independent measurements, the likelihood of a particular set of observations, given the true value, is,

$$\mathcal{L}(m_1, \ldots, m_n \mid m) = \prod_{i=1}^{n} \phi(m_i - m)$$

Gauss assumed (from bullet 3 above) that the most likely value for the quantity m, given the n observations, was the maximum likelihood estimate of m, or the mean of the measurements. Therefore, we can rewrite the likelihood of a particular set of observations as the product of differences from the *mean*,

$$\mathcal{L}(m_1, \ldots, m_n \mid m) = \prod_{i=1}^{n} \phi(m_i - \bar{m})$$

[3] Laplace and Simpson also took a stab at determining an error curve in the decades that preceded and followed Gauss' success. If you're interested in learning about these stories, I would highly recommend visiting the following URL: `http://www.ww.ingeniousmathstat.org/sites/default/files/pdf/upload_library/22/Allendoerfer/stahl96.pdf`

Differentiating $\mathcal{L}$ with respect to m gives an important characteristic of the error curve.

$$\begin{aligned} \left.\frac{\partial \mathcal{L}}{\partial m}\right|_{m=\bar{m}} &= 0 \\ \left.\frac{\partial}{\partial m}\prod_{i=1}^{n}\phi(m_i - m)\right|_{m=\bar{m}} &= 0 \\ -\left(\sum_{i=1}^{n}\frac{\phi'(m_i-\bar{m})}{\phi(m_i-\bar{m})}\right)\mathcal{L}(m_1,\ldots,m_n \mid \bar{m}) &= 0 \end{aligned}$$

It follows that,

$$\sum_{i=1}^{n}\frac{\phi'(m_i-\bar{m})}{\phi(m_i-\bar{m})} = 0$$

It can be shown that this condition implies that the ratio of the derivative of the function and the function itself is linear, i.e.,

$$\frac{\phi'(x)}{\phi(x)} = kx$$

For some arbitrary real constant k. Solving this differential equation, we have,

$$\begin{aligned} \int \frac{\phi'(x)}{\phi(x)} &= \int kx \\ \ln\phi(x) &= \frac{k}{2}x^2 + C \\ \phi(x) &= A\exp\left(\frac{k}{2}x^2\right) \end{aligned}$$

For some positive constant A. In order for the distribution to be symmetric about zero (maximal for $x = 0$, $k/2$ must be some negative constant. So we let $k/2 = -h^2$ for some constant h. Then, the final form of the distribution is obtained from integrating over the density to obtain the normalizing constant A.

$$\int_{-\infty}^{\infty}\exp\left(-h^2x^2\right)dx = \frac{\sqrt{\pi}}{h}$$

Therefore,

$$\phi(x) = \frac{h}{\sqrt{\pi}}e^{h^2x^2}$$

□

Of course, this is exactly the form of the normal density, when the mean is zero (which Gauss assumed), and the constant h is $\frac{1}{\sigma\sqrt{2}}$.

The results of this section are highly relevant to our exploration of digits, images, and emails in Machine Learning throughout the semester. The data generation and measurement process can be represented as the sum of a "truth" and a linear combination of symmetric error curves. The Central Limit theorem and the definition of the noisy process in this section provide us with some assurance that the normal distribution adequately models these errors, and therefore motivates the usage of them in statistical models. Next, we turn to the derivation of the Multivariate Gaussian, which is an adaptation of the univariate density for high-dimensional data vectors.

2. The Multivariate Normal Distribution

Our goal in this section is to derive the well-known density function for the multivariate normal distribution, which deals with random vectors instead of just individual random variables. To do this, we will start with a collection (a vector) of independent, normally distributed random variables and work ourselves up to the general case where they are no longer independent.

2.1. The Basic Case: Independent Univariate Normals

To derive the general case of the multivariate normal, we will start with a vector consisting of N independent, normally distributed random variables and mean 0: $\mathbf{Z} = (Z_1, Z_2, \dots, Z_N)$, where $Z_i \sim \mathcal{N}(0, \sigma_i^2)$. We denote the density of a single Z_i as f_{Z_i}. Then, since the variables are independent, the joint probability density function, $f_{\mathbf{Z}}$ of all N variables will just be the product of their densities. That is,:

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{x}) &= \Pi_{i=1}^{N} f_{Z_i}(x_i) \\ &= \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\frac{1}{2}\frac{x_i^2}{\sigma_i^2}\} \\ &= \frac{1}{\sqrt{(2\pi)^N \Pi_{i=1}^{N}\sigma_i^2}} \exp\{-\frac{1}{2}\mathbf{x}^T(\mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)^{-1}\mathbf{x}\} \\ &= \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\{-\frac{1}{2}\mathbf{x}^\intercal \Sigma^{-1}\mathbf{x}\} \end{aligned}$$

Where $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)$. In this case we say that $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Unfortunately, this derivation is restricted to the case where these entries are independent and 0-centered. However, we will see in the next few sections that we can derive the general case using this result.

2.2. Affine Transformations of a Random Vector

Consider an affine transformation $L : \mathbb{R}^N \to \mathbb{R}^N$, $L(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for an invertible matrix $A \in \mathbb{R}^{N\times N}$ and a constant vector $\mathbf{b} \in \mathbb{R}^N$. It is easy to verify that when we apply this transformation to a random variable $\mathbf{Z}$, with mean $\mathbb{E}\mathbf{Z} = \mu$ and covariance $\mathrm{cov}(\mathbf{Z}) = \Sigma_Z$, we get a new random variable $\mathbf{X} = L(\mathbf{Z})$ such that:

$$\begin{aligned} &\mathbb{E}\mathbf{X} = L(\mathbb{E}\mathbf{Z}) \\ &\Sigma_X = \mathrm{cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\intercal] = A\Sigma_Z A^\intercal \end{aligned}$$

In our case, the affine transformation will consist of a rotation followed by a scaling along the principal axes of our rotation, with one translation. That is, for a symmetric, positive definite matrix Σ and constant vector μ, we will be looking at the transformation $\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \mu$. It is interesting to note that, given an orthogonal decomposition $\Sigma = U\Lambda U^T$, where U is orthogonal and Λ is a diagonal matrix consisting of the eigenvalues of Σ, entry x_i of the new random vector is a weighted sum of originally independent random variables in $\mathbf{Z}$. Let A_i denote the ith row of a matrix A. Then,

$$\begin{aligned} x_i &= (\Sigma^{1/2}\mathbf{Z} + \mu)_i \\ &= \sqrt{\lambda_i} U_i \cdot \mathbf{Z} + \mu_i \\ &= \sum_{j=1}^{N} \sqrt{\lambda_i} U_{ij} z_j + \mu_i \end{aligned}$$

This is, in a sense, the source of the covariance between entries of $\mathbf{X}$.

We now just need one more fact about a change of variables to derive the general multivariate normal PDF for this new random vector.

2.3. PDF of a Transformed Random Vector

Suppose that $\mathbf{Z}$ is a random vector taking on values in a subset $S \subseteq \mathbb{R}^N$, with a continuous probability density function f. Suppose that $\mathbf{X} = r(\mathbf{Z})$ where r is a differentiable function from S onto some other subset $T \subseteq \mathbb{R}^N$. Then the probability density function g of $\mathbf{X}$ is given by:

$$\begin{aligned} g(\mathbf{x}) &= f(\mathbf{z}) |\det(\frac{d\mathbf{z}}{d\mathbf{x}})| \\ &= f(r^{-1}(\mathbf{x})) |\det(\frac{d\mathbf{z}}{d\mathbf{x}})| \end{aligned}$$

Where if you recall from multivariable calculus, $\frac{d\mathbf{z}}{d\mathbf{x}}$ is the Jacobian of the inverse of r, and $\det(\cdot)$ denotes the determinant of a matrix [4].

Returning to our previous discussion, where $\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \mu$, we can see that the inverse transformation is given by $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \mu)$. It is easily verifiable that the determinant of the Jacobian of this inverse is given by $\frac{1}{\sqrt{\det(\Sigma)}}$. Now we have everything we need for the general case.

2.4. The Multivariate Normal PDF

Consider the random vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$ where I is the identity matrix. As before we let $\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \mu$ for positive definite Σ and a constant vector μ. We can now find the density function g of $\mathbf{X}$ from the known density function f for $\mathbf{Z}$.

$$g(\mathbf{x}) = f(r^{-1}(\mathbf{x})) |\det(\frac{d\mathbf{z}}{d\mathbf{x}})|$$

[4] For a proof of this fact, see `http://www.math.uah.edu/stat/dist/Transformations.html`

$$
\begin{aligned}
&= f(\Sigma^{-1/2}(\mathbf{x}-\mu))\frac{1}{\sqrt{|\Sigma|}} \\
&= \frac{1}{\sqrt{(2\pi)^N}}\frac{1}{\sqrt{|\Sigma|}}\exp\{-\frac{1}{2}(\Sigma^{-1/2}(\mathbf{x}-\mu))^\intercal(\Sigma^{-1/2}(\mathbf{x}-\mu))\} \\
&= \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp\{-\frac{1}{2}(\mathbf{x}-\mu)^\intercal\Sigma^{-1}(\mathbf{x}-\mu)\}
\end{aligned}
$$

This is probability density function for a multivariate normal distribution with mean vector μ and a covariance matrix Σ. We say that $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

One important insight is that by performing an affine transformation on a vector consisting of independent, normally distributed variables, we have "induced" a measure of dependence, the covariance, between the entries of our resulting random vector. By using some properties related to a change of variables, we then derived the density function for the resulting distribution.

References

[1] Caldwell, Allen. "Gaussian Distribution." SpringerReference (n.d.): n. pag. Max-Plank-Institut Fr Physik. 4 May 2009. Web. 17 Mar. 2017.

[2] Pitman, Jim. Probability. Beijing: World Corporation, 2009. Print.

[3] Rice, John A. Mathematical Statistics and Data Analysis. New Delhi: Cengage Learning/Brooks/Cole, 2007. Print.

[4] Siegrist, Kyle. "Transformations of Variables." Transformations of Variables. N.p., n.d. Web. 17 Mar. 2017.

[5] Stahl, Saul. "The Normal Distribution." Mathematics Magazine 72.2 (2006): 99-106. Mathematical Association of America. Mathematical Association of America, 2006. Web. 17 Mar. 2017.