5 Probability

Probability theory provides powerful tools for modeling and dealing with uncertainty.

5.1 Basics

Suppose we have some sort of randomized experiment (e.g. a coin toss, die roll) that has a fixed set of possible outcomes. This set is called the **sample space** and denoted Ω.

We would like to define probabilities for some **events**, which are subsets of Ω. The set of events is denoted $\mathcal{F}$.[9] The **complement** of the event A is another event, $A^c = \Omega \setminus A$.

Then we can define a **probability measure** $\mathbb{P} : \mathcal{F} \to [0, 1]$ which must satisfy

(i) $\mathbb{P}(\Omega) = 1$

(ii) **Countable additivity**: for any countable collection of disjoint sets $\{A_i\} \subseteq \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.[10]

If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely** (often abbreviated a.s.).[11], and conversely A occurs **almost never** if $\mathbb{P}(A) = 0$.

From these axioms, a number of useful rules can be derived.

Proposition 26. *Let A be an event. Then*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(ii) *If B is an event and $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.*

(iii) $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$

Proof. (i) Using the countable additivity of $\mathbb{P}$, we have

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \,\dot\cup\, A^c) = \mathbb{P}(\Omega) = 1$$

To show (ii), suppose $B \in \mathcal{F}$ and $B \subseteq A$. Then

$$\mathbb{P}(A) = \mathbb{P}(B \,\dot\cup\, (A \setminus B)) = \mathbb{P}(B) + \mathbb{P}(A \setminus B) \geq \mathbb{P}(B)$$

as claimed.

For (iii): the middle inequality follows from (ii) since $\varnothing \subseteq A \subseteq \Omega$. We also have

$$\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing \,\dot\cup\, \varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing)$$

by countable additivity, which shows $\mathbb{P}(\varnothing) = 0$. □

[9] $\mathcal{F}$ is required to be a σ-algebra for technical reasons; see [2].
[10] Note that a probability space is simply a measure space in which the measure of the whole space equals 1.
[11] This is a probabilist's version of the measure-theoretic term *almost everywhere*.

Proposition 27. *If A and B are events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Proof. The key is to break the events up into their various overlapping and non-overlapping parts.

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B) \dot\cup (A \setminus B) \dot\cup (B \setminus A)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\end{aligned}$$

□

Proposition 28. *If $\{A_i\} \subseteq \mathcal{F}$ is a countable set of events, disjoint or not, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i)$$

This inequality is sometimes referred to as **Boole's inequality** or the **union bound**.

Proof. Define $B_1 = A_1$ and $B_i = A_i \setminus (\bigcup_{j<i} A_j)$ for $i > 1$, noting that $\bigcup_{j \leq i} B_j = \bigcup_{j \leq i} A_j$ for all i and the B_i are disjoint. Then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\bigcup_i B_i\right) = \sum_i \mathbb{P}(B_i) \leq \sum_i \mathbb{P}(A_i)$$

where the last inequality follows by monotonicity since $B_i \subseteq A_i$ for all i. □

5.1.1 Conditional probability

The **conditional probability** of event A given that event B has occurred is written $\mathbb{P}(A|B)$ and defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

assuming $\mathbb{P}(B) > 0$.[12]

5.1.2 Chain rule

Another very useful tool, the **chain rule**, follows immediately from this definition:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

5.1.3 Bayes' rule

Taking the equality from above one step further, we arrive at the simple but crucial **Bayes' rule**:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

[12] In some cases it is possible to define conditional probability on events of probability zero, but this is significantly more technical so we omit it.

It is sometimes beneficial to omit the normalizing constant and write

$$\mathbb{P}(A|B) \propto \mathbb{P}(A)\mathbb{P}(B|A)$$

Under this formulation, $\mathbb{P}(A)$ is often referred to as the **prior**, $\mathbb{P}(A|B)$ as the **posterior**, and $\mathbb{P}(B|A)$ as the **likelihood**.

In the context of machine learning, we can use Bayes' rule to update our "beliefs" (e.g. values of our model parameters) given some data that we've observed.

5.2 Random variables

A **random variable** is some uncertain quantity with an associated probability distribution over the values it can assume.

Formally, a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function[13] $X : \Omega \to \mathbb{R}$.[14]

We denote the range of X by $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. To give a concrete example (taken from [3]), suppose X is the number of heads in two tosses of a fair coin. The sample space is

$$\Omega = \{hh, tt, ht, th\}$$

and X is determined completely by the outcome ω, i.e. $X = X(\omega)$. For example, the event $X = 1$ is the set of outcomes $\{ht, th\}$.

It is common to talk about the values of a random variable without directly referencing its sample space. The two are related by the following definition: the event that the value of X lies in some set $S \subseteq \mathbb{R}$ is

$$X \in S = \{\omega \in \Omega : X(\omega) \in S\}$$

Note that special cases of this definition include X being equal to, less than, or greater than some specified value. For example

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

A word on notation: we write $p(X)$ to denote the entire probability distribution of X and $p(x)$ for the evaluation of the function p at a particular value $x \in X(\Omega)$. Hopefully this (reasonably standard) abuse of notation is not too distracting. If p is parameterized by some parameters θ, we write $p(X; \theta)$ or $p(x; \theta)$, unless we are in a Bayesian setting where the parameters are considered a random variable, in which case we condition on the parameters.

5.2.1 The cumulative distribution function

The **cumulative distribution function** (c.d.f.) gives the probability that a random variable is at most a certain value:

$$F(x) = \mathbb{P}(X \leq x)$$

The c.d.f. can be used to give the probability that a variable lies within a certain range:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a)$$

[13] The function must be measurable.

[14] More generally, the codomain can be any measurable space, but $\mathbb{R}$ is the most common case by far and sufficient for our purposes.

5.2.2 Discrete random variables

A **discrete random variable** is a random variable that has a countable range and assumes each value in this range with positive probability. Discrete random variables are completely specified by their **probability mass function** (p.m.f.) $p : X(\Omega) \to [0, 1]$ which satisfies

$$\sum_{x \in X(\Omega)} p(x) = 1$$

For a discrete X, the probability of a particular value is given exactly by its p.m.f.:

$$\mathbb{P}(X = x) = p(x)$$

5.2.3 Continuous random variables

A **continuous random variable** is a random variable that has an uncountable range and assumes each value in this range with probability zero. Most of the continuous random variables that one would encounter in practice are **absolutely continuous random variables**[15], which means that there exists a function $p : \mathbb{R} \to [0, \infty)$ that satisfies

$$F(x) \equiv \int_{-\infty}^{x} p(z) \, \mathrm{d}z$$

The function p is called a **probability density function** (abbreviated p.d.f.) and must satisfy

$$\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1$$

The values of this function are not themselves probabilities, since they could exceed 1. However, they do have a couple of reasonable interpretations. One is as relative probabilities; even though the probability of each particular value being picked is technically zero, some points are still in a sense more likely than others.

One can also think of the density as determining the probability that the variable will lie in a small range about a given value. This is because, for small $\epsilon > 0$,

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} p(z) \, \mathrm{d}z \approx 2\epsilon p(x)$$

using a midpoint approximation to the integral.

Here are some useful identities that follow from the definitions above:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} p(x) \, \mathrm{d}x$$
$$p(x) = F'(x)$$

5.2.4 Other kinds of random variables

There are random variables that are neither discrete nor continuous. For example, consider a random variable determined as follows: flip a fair coin, then the value is zero if it comes up heads, otherwise draw a number uniformly at random from $[1, 2]$. Such a random variable can take on uncountably many values, but only finitely many of these with positive probability. We will not discuss such random variables because they are rather pathological and require measure theory to analyze.

[15] Random variables that are continuous but not absolutely continuous are called **singular random variables**. We will not discuss them, assuming rather that all continuous random variables admit a density function.

5.3 Joint distributions

Often we have several random variables and we would like to get a distribution over some combination of them. A **joint distribution** is exactly this. For some random variables $X_1, \ldots, X_n$, the joint distribution is written $p(X_1, \ldots, X_n)$ and gives probabilities over entire assignments to all the X_i simultaneously.

5.3.1 Independence of random variables

We say that two variables X and Y are **independent** if their joint distribution factors into their respective distributions, i.e.

$$p(X, Y) = p(X)p(Y)$$

We can also define independence for more than two random variables, although it is more complicated. Let $\{X_i\}_{i \in I}$ be a collection of random variables indexed by I, which may be infinite. Then $\{X_i\}$ are independent if for every finite subset of indices $i_1, \ldots, i_k \in I$ we have

$$p(X_{i_1}, \ldots, X_{i_k}) = \prod_{j=1}^{k} p(X_{i_j})$$

For example, in the case of three random variables, X, Y, Z, we require that $p(X, Y, Z) = p(X)p(Y)p(Z)$ as well as $p(X, Y) = p(X)p(Y)$, $p(X, Z) = p(X)p(Z)$, and $p(Y, Z) = p(Y)p(Z)$.

It is often convenient (though perhaps questionable) to assume that a bunch of random variables are **independent and identically distributed** (i.i.d.) so that their joint distribution can be factored entirely:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i)$$

where $X_1, \ldots, X_n$ all share the same p.m.f./p.d.f.

5.3.2 Marginal distributions

If we have a joint distribution over some set of random variables, it is possible to obtain a distribution for a subset of them by "summing out" (or "integrating out" in the continuous case) the variables we don't care about:

$$p(X) = \sum_{y} p(X, y)$$

5.4 Great Expectations

If we have some random variable X, we might be interested in knowing what is the "average" value of X. This concept is captured by the **expected value** (or **mean**) $\mathbb{E}[X]$, which is defined as

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} xp(x)$$

for discrete X and as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$$

for continuous X.

In words, we are taking a weighted sum of the values that X can take on, where the weights are the probabilities of those respective values. The expected value has a physical interpretation as the "center of mass" of the distribution.

5.4.1 Properties of expected value

A very useful property of expectation is that of linearity:

$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha_i X_i + \beta\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] + \beta$$

Note that this holds even if the X_i are not independent!

But if they are independent, the product rule also holds:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

5.5 Variance

Expectation provides a measure of the "center" of a distribution, but frequently we are also interested in what the "spread" is about that center. We define the variance $\mathrm{Var}(X)$ of a random variable X by

$$\mathrm{Var}(X) = \mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^2\right]$$

In words, this is the average squared deviation of the values of X from the mean of X. Using a little algebra and the linearity of expectation, it is straightforward to show that

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

5.5.1 Properties of variance

Variance is not linear (because of the squaring in the definition), but one can show the following:

$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \,\mathrm{Var}(X)$$

Basically, multiplicative constants become squared when they are pulled out, and additive constants disappear (since the variance contributed by a constant is zero).

Furthermore, if $X_1, \ldots, X_n$ are uncorrelated[16], then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$$

5.5.2 Standard deviation

Variance is a useful notion, but it suffers from that fact the units of variance are not the same as the units of the random variable (again because of the squaring). To overcome this problem we can use **standard deviation**, which is defined as $\sqrt{\mathrm{Var}(X)}$. The standard deviation of X has the same units as X.

[16] We haven't defined this yet; see the Correlation section below

5.6 Covariance

Covariance is a measure of the linear relationship between two random variables. We denote the covariance between X and Y as $\text{Cov}(X, Y)$, and it is defined to be

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Note that the outer expectation must be taken over the joint distribution of X and Y.

Again, the linearity of expectation allows us to rewrite this as

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Comparing these formulas to the ones for variance, it is not hard to see that $\text{Var}(X) = \text{Cov}(X, X)$.

A useful property of covariance is that of **bilinearity**:

$$\text{Cov}(\alpha X + \beta Y, Z) = \alpha \,\text{Cov}(X, Z) + \beta \,\text{Cov}(Y, Z)$$
$$\text{Cov}(X, \alpha Y + \beta Z) = \alpha \,\text{Cov}(X, Y) + \beta \,\text{Cov}(X, Z)$$

5.6.1 Correlation

Normalizing the covariance gives the **correlation**:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Correlation also measures the linear relationship between two variables, but unlike covariance always lies between -1 and 1.

Two variables are said to be **uncorrelated** if $\text{Cov}(X, Y) = 0$ because $\text{Cov}(X, Y) = 0$ implies that $\rho(X, Y) = 0$. If two variables are independent, then they are uncorrelated, but the converse does not hold in general.

5.7 Random vectors

So far we have been talking about **univariate distributions**, that is, distributions of single variables. But we can also talk about **multivariate distributions** which give distributions of **random vectors**:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

The summarizing quantities we have discussed for single variables have natural generalizations to the multivariate case.

Expectation of a random vector is simply the expectation applied to each component:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

The variance is generalized by the **covariance matrix**:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}$$

That is, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Since covariance is symmetric in its arguments, the covariance matrix is also symmetric. It's also positive semi-definite: for any $\mathbf{x}$,

$$\mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x} = \mathbf{x}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]\mathbf{x} = \mathbb{E}[\mathbf{x}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x}] = \mathbb{E}[((\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x})^2] \geq 0$$

The inverse of the covariance matrix, $\boldsymbol{\Sigma}^{-1}$, is sometimes called the **precision matrix**.

5.8 Estimation of Parameters

Now we get into some basic topics from statistics. We make some assumptions about our problem by prescribing a **parametric** model (e.g. a distribution that describes how the data were generated), then we fit the parameters of the model to the data. How do we choose the values of the parameters?

5.8.1 Maximum likelihood estimation

A common way to fit parameters is **maximum likelihood estimation** (MLE). The basic principle of MLE is to choose values that "explain" the data best by maximizing the probability/density of the data we've seen as a function of the parameters. Suppose we have random variables $X_1, \dots, X_n$ and corresponding observations $x_1, \dots, x_n$. Then

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \mathcal{L}(\theta)$$

where $\mathcal{L}$ is the **likelihood function**

$$\mathcal{L}(\theta) = p(x_1, \dots, x_n; \theta)$$

Often, we assume that $X_1, \dots, X_n$ are i.i.d. Then we can write

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$$

At this point, it is usually convenient to take logs, giving rise to the **log-likelihood**

$$\log \mathcal{L}(\theta) = \sum_{i=1}^n \log p(x_i; \theta)$$

This is a valid operation because the probabilities/densities are assumed to be positive, and since log is a monotonically increasing function, it preserves ordering. In other words, any maximizer of $\log \mathcal{L}$ will also maximize $\mathcal{L}$.

For some distributions, it is possible to analytically solve for the maximum likelihood estimator. If $\log \mathcal{L}$ is differentiable, setting the derivatives to zero and trying to solve for θ is a good place to start.

5.8.2 Maximum a posteriori estimation

A more Bayesian way to fit parameters is through **maximum a posteriori estimation** (MAP). In this technique we assume that the parameters are a random variable, and we specify a prior distribution $p(\theta)$. Then we can employ Bayes' rule to compute the posterior distribution of the parameters given the observed data:

$$p(\theta|x_1, \dots, x_n) \propto p(\theta)p(x_1, \dots, x_n|\theta)$$

Computing the normalizing constant is often intractable, because it involves integrating over the parameter space, which may be very high-dimensional. Fortunately, if we just want the MAP estimate, we don't care about the normalizing constant! It does not affect which values of θ maximize the posterior. So we have

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta)p(x_1, \dots, x_n|\theta)$$

Again, if we assume the observations are i.i.d., then we can express this in the equivalent, and possibly friendlier, form

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left(\log p(\theta) + \sum_{i=1}^{n} \log p(x_i|\theta) \right)$$

A particularly nice case is when the prior is chosen carefully such that the posterior comes from the same family as the prior. In this case the prior is called a **conjugate prior**. For example, if the likelihood is binomial and the prior is beta, the posterior is also beta. There are many conjugate priors; the reader may find this table of conjugate priors useful.

5.9 The Gaussian distribution

There are many distributions, but one of particular importance is the **Gaussian distribution**, also known as the **normal distribution**. It is a continuous distribution, parameterized by its mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, with density

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

Note that in the special case $d = 1$, the density is written in the more recognizable form

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

We write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\mathbf{X}$ is normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

5.9.1 The geometry of multivariate Gaussians

The geometry of the multivariate Gaussian density is intimately related to the geometry of positive definite quadratic forms, so make sure the material in that section is well-understood before tackling this section.

First observe that the p.d.f. of the multivariate Gaussian can be rewritten as

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}})$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}$ and $g(z) = [(2\pi)^d \det(\boldsymbol{\Sigma})]^{-\frac{1}{2}} \exp\left(-\frac{z}{2}\right)$. Writing the density in this way, we see that after shifting by the mean $\boldsymbol{\mu}$, the density is really just a simple function of its precision matrix's quadratic form.

Here is a key observation: this function g is **strictly monotonically decreasing** in its argument. That is, $g(a) > g(b)$ whenever $a < b$. Therefore, small values of $\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$ (which generally correspond to points where $\tilde{\mathbf{x}}$ is closer to $\mathbf{0}$, i.e. $\mathbf{x} \approx \boldsymbol{\mu}$) have relatively high probability densities, and vice-versa. Furthermore, because g is *strictly* monotonic, it is injective, so the c-isocontours of $p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the $g^{-1}(c)$-isocontours of the function $\mathbf{x} \mapsto \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$. That is, for any c,

$$\{\mathbf{x} \in \mathbb{R}^d : p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\} = \{\mathbf{x} \in \mathbb{R}^d : \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}} = g^{-1}(c)\}$$

In words, these functions have the same isocontours but different isovalues.

Recall the executive summary of the geometry of positive definite quadratic forms: the isocontours of $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ are ellipsoids such that the axes point in the directions of the eigenvectors of $\mathbf{A}$, and the lengths of these axes are proportional to the inverse square roots of the corresponding eigenvalues. Therefore in this case, the isocontours of the density are ellipsoids (centered at $\boldsymbol{\mu}$) with axis lengths proportional to the inverse square roots of the eigenvalues of $\boldsymbol{\Sigma}^{-1}$, or equivalently, the square roots of the eigenvalues of $\boldsymbol{\Sigma}$.