Tutorial: Statistical distance and Fisher information

Pieter Kok
Department of Materials, Oxford University, Parks Road, Oxford OX1 3PH, UK

Statistical distance

We wish to construct a *space* of probability distributions with a *distance* defined on it. A function $s(a,b)$ is a distance between two points a and b if s obeys four requirements:

1. $s(a,b) \geq 0$
2. $s(a,b) = 0 \quad \Leftrightarrow \quad a = b$
3. $s(a,b) = s(b,a)$
4. $s(a,c) \leq s(a,b) + s(b,c)$ (the triangle inequality)

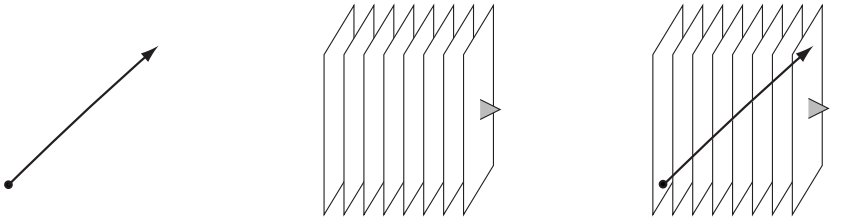
Furthermore, the distance in a metric space is a *metric* g_{kl}. In infinitesimal form:

$$ds^2 = \sum_{kl} g_{kl}\, da^k da^l \ , \tag{1}$$

where the da^k's are the components of the *tangent vector* to a. For a Euclidean space, the metric is the identity tensor, and the distance between points a and $a + da$ becomes Pythagoras' theorem:

$$ds^2 = \sum_{k} da_k da^k \ . \tag{2}$$

Note that there is a difference between the upper and the lower indices. A vector with lower indices is a *covariant* vector, while a vector with upper indices is *contravariant*. We can represent this pictorially as [2]:

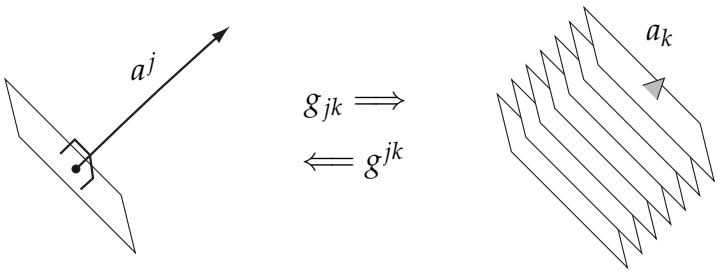


contravariant covariant scalar product

The scalar product is a product (a *contraction*) between a contravariant and a covariant vector (upper and lower indices). It corresponds to the number of sheets in the stack that the arrow pierces. It is now easy to see that transformations on the space leave the scalar product invariant: Distorting the arrow and the stack of sheets in equal measure will not change the number of sheets that the arrow pierces. The contravariant vector is the *dual* of the covariant vector, and they are related via the metric [1]:

$$a_j = \sum_k g_{jk}\, a^k \quad \text{and} \quad a^j = \sum_k g^{jk}\, a_k \tag{3}$$

In other words, the contravariant and covariant forms of the metric are "raising" and "lowering" operators on the indices, and pictorially the vectors transform as

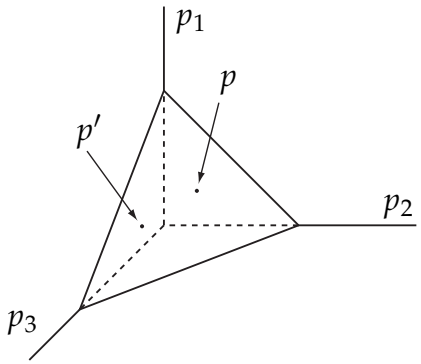


From the raising and lowering of the indices, we can derive a key property of the metric that we will use in this lecture:

$$a_j = \sum_k g_{jk} \left(\sum_l g^{kl} a_l \right) = \sum_{kl} g_{jk}\, g^{kl}\, a_l \quad \Rightarrow \quad \sum_l g_{jk}\, g^{kl} = \delta_j{}^l \equiv \delta_{jl}, \tag{4}$$

where the second identity follows from the linear independence of the a_j. The symbol δ_{jl} is the Kronecker delta. So the contravariant form of the metric g^{jk} is the inverse of the covariant metric g_{jk}.

Next, the *probability simplex* is the convex space of probability distributions:



This is a metric space, but the (Euclidean) distance between the two points p and p' is not necessarily the most useful distance function for probability distributions. We can write down a general distance function between two probability distributions p and $p' = p + dp$ in terms of the metric:

$$ds^2 = \sum_{jk} g_{jk}\, dp^j dp^k \ . \tag{5}$$

In general, the sum may be an integral.

Our next task is to find the *most natural* metric g_{jk} of the simplex [3]. To this end we construct the natural dual to the tangent vectors dp of a probability distribution p, namely a random variable. In particular, *observables* (A) can be considered random variables, with expectation value

$$\langle A \rangle \equiv \sum_j A_j p^j \, . \tag{6}$$

The points of constant expectation value form surfaces in the dual space to dp. Consequently, surfaces with incrementally increasing expectation values form a stack, which makes for a natural covariant vector.

The metric is determined by a quadratic form of tangent (contravariant) vectors, or alternatively, a quadratic form of covariant vectors. We choose to derive the metric from the latter. For a pair of observables, the natural quadratic form is the so-called *correlation*

$$\sum_{jk} A_j B_k g^{jk} = \langle AB \rangle = \sum A_j B_j p^j \ . \tag{7}$$

From this we see immediately that $g^{jk} = \delta_j{}^k p^j$. Using $\sum_k g^{jk} g_{kl} = \delta^j{}_l$ we have

$$g_{jk} = \frac{\delta^j{}_k}{p^j} \ . \tag{8}$$

And therefore

$$ds^2 = \sum_{jk} g_{jk}\, dp^j dp^k = \sum_j \frac{(dp^j)^2}{p^j} \tag{9}$$

is the statistical distance between two infinitesimally close probability distributions p and $p + dp$.

In practice, we are typically interested in two probability distributions that are separated by a *finite* distance. How can we say anything about that? We can substitute $p^j = (r^j)^2$ and using $d(r^j)^2 = 2 r^j dr^j$:

$$ds^2 = 4 \sum_j (dr^j)^2 \ , \tag{10}$$

which can now be integrated directly:

$$s^2 = 4\sum_j \left| r^j - r'^j \right|^2 = 4\sum_j \left[(r^j)^2 + (r'^j)^2 - 2r^j r'^j \right] = 8(1 - r \cdot r') \ . \tag{11}$$

So we expressed the statistical distance as the natural Euclidean distance for *probability amplitudes*! Notice that we have not talked about quantum mechanics at all at this point: Probability amplitudes arise naturally in the classical theory.

Suppose we are given an ensemble of systems characterised by one of two probability distributions p and p'. The statistical distance between two probability distributions is a measure of how hard it is to distinguish between them given a certain number of samples. In general, two probability distributions are distinguishable after N samples when[1]

$$Nds^2 \geq 1 \ . \tag{12}$$

This is a form of the *Cramér-Rao bound*, and we would like to derive this bound more formally.

Fisher information and the Cramér-Rao bound

In this section, I follow the derivation of Braunstein and Caves [4] to get the Cramér-Rao bound. One of the steps in the derivation will lead to the concept of the *Fisher information*.

We consider the situation where we wish to distinguish two probability distributions $p(\theta)$ and $p(\theta')$ on the basis of measurement outcomes x. We can then define the *conditional* probability $p(x|\theta)$ of finding the outcome x given a system prepared in a state $\rho(\theta)$. In quantum mechanics, this conditional probability is obtained from the state $\rho(\theta)$ and the POVM for the measurement outcome $\hat{E}_x$ according to

$$p(x|\theta) = \mathrm{Tr}\left[\hat{E}_x \rho(\theta)\right] \ , \tag{13}$$

with $\int dx\, \hat{E}_x = \mathbb{1}$. This can be related to the conditional probability of having a system in state $\rho(\theta)$ given the measurement outcomes x via Bayes' rule:

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} \equiv L(\theta|x) \ , \tag{14}$$

where L is the *likelihood* of θ. Telling the difference between the distributions $p(\theta)$ and $p(\theta + \delta\theta)$ therefore amounts to the estimation of the parameter θ if $p(\theta)$ and $p(x)$ are known (or can be inferred).

[1] We state this again in infinitesimal form, since the distinguishability criterion is tight only for infinitesimal distances. However, this formula does provide bounds for distinguishing probability distributions separated by a finite distance.

After these preliminary notes, we are now ready to proceed with the derivation of the Cramér-Rao bound. Let $T(x)$ be an estimator for θ based on the measurement outcome x, and let $\Delta T \equiv T(x) - \langle T \rangle_\theta$. The estimator is called *unbiased* when $\langle T \rangle_\theta = \theta$, where

$$\langle A \rangle_\theta \equiv \int dx\, p(x|\theta) A \ . \tag{15}$$

For *any* estimator T (biased or unbiased), and for N independent samples yielding measurement outcomes $x_1, \ldots, x_N$ we have

$$\int dx_1 \cdots dx_N\, p(x_1|\theta) \cdots p(x_N|\theta)\, \Delta T = 0 \ . \tag{16}$$

Taking the derivative to θ and using the chain rule yields

$$\sum_{i=1}^{N} \int dx_1 \cdots dx_N\, p(x_1|\theta) \cdots p(x_N|\theta)\, \frac{1}{p(x_i|\theta)} \frac{\partial p(x_i|\theta)}{\partial \theta}\, \Delta T - \left\langle \frac{d\langle T \rangle_\theta}{d\theta} \right\rangle = 0 \ , \tag{17}$$

where we used that T does not depend on θ. This can be rewritten as

$$\int dx_1 \cdots dx_N\, p(x_1|\theta) \cdots p(x_N|\theta) \left(\sum_{i=1}^{N} \frac{\partial \ln p(x_i|\theta)}{\partial \theta} \right) \Delta T = \left\langle \frac{d\langle T \rangle_\theta}{d\theta} \right\rangle \ . \tag{18}$$

To this equation we can now apply the Schwarz inequality [5]:

$$|\langle f, g \rangle|^2 \leq \langle f, f \rangle \langle g, g \rangle \ . \tag{19}$$

This yields

$$\int dx_1 \cdots dx_N\, p(x_1|\theta) \cdots p(x_N|\theta) \left(\sum_{i=1}^{N} \frac{\partial \ln p(x_i|\theta)}{\partial \theta} \right)^2 \times \int dx_1 \cdots dx_N\, p(x_1|\theta) \cdots p(x_N|\theta)\, (\Delta T)^2 \ \geq \ \left| \left\langle \frac{d\langle T \rangle_\theta}{d\theta} \right\rangle \right|^2 \tag{20}$$

or

$$N\, F(\theta)\, \langle (\Delta T)^2 \rangle_\theta \geq \left| \frac{d\langle T \rangle_\theta}{d\theta} \right|^2 , \tag{21}$$

where we introduced the *Fisher information* $F(\theta)$:

$$F(\theta) \equiv \int dx\, p(x|\theta) \left(\frac{\partial \ln p(x|\theta)}{\partial \theta} \right)^2 = \int dx\, \frac{1}{p(x|\theta)} \left(\frac{\partial p(x|\theta)}{\partial \theta} \right)^2 \ . \tag{22}$$

The error in the parameter θ is related to the estimator T in the following way:

$$\delta T \equiv \frac{T}{|d\langle T \rangle_\theta / d\theta|} - \theta \ . \tag{23}$$

Note that this is not the same as ΔT, as δT is related to the *actual value* of θ. The derivative removes the local difference in the units of the estimator and the parameter [4]. Note that a bias in the estimator T shows up as a non-zero $\langle \delta T \rangle_\theta$.

In order to derive the Cramér-Rao bound, we need to relate the error in θ (given by δT) to the variance of the estimator ΔT. Using $\Delta T = T(x) - \langle T \rangle_\theta$, we find

$$\langle (\Delta T)^2 \rangle_\theta = \left| \frac{d\langle T \rangle_\theta}{d\theta} \right|^2 \left(\left\langle (\delta T)^2 \right\rangle_\theta - \langle \delta T \rangle_\theta^2 \right) \ . \tag{24}$$

The Cramér-Rao bound then follows immediately:

$$\langle (\delta T)^2 \rangle_\theta \geq \frac{1}{N\,F(\theta)} + \langle \delta T \rangle_\theta^2 \geq \frac{1}{N\,F(\theta)} \ . \tag{25}$$

Furthermore, the Fisher information can be expressed in terms of the statistical distance ds:

$$F(\theta) = \int dx \, \frac{1}{p(x|\theta)} \left(\frac{\partial p(x|\theta)}{\partial \theta} \right)^2 = \left(\frac{ds}{d\theta} \right)^2 \ . \tag{26}$$

So the Fisher information is the square of the derivative to θ of the statistical distance.

Finally, let's return to the distinguishability criterion between two probability distributions stated on page 4. If we take the variance of an unbiased estimator to be $\Delta\theta$ and we consider two probability distributions with (finite but small) distance $(\Delta s)^2$, the Cramér-Rao bound becomes

$$(\Delta\theta)^2 \equiv \langle (\delta T)^2 \rangle_\theta \geq \frac{1}{NF(\theta)} \simeq \frac{1}{N} \left(\frac{\Delta s}{\Delta \theta} \right)^{-2} \qquad \text{or} \qquad N(\Delta s)^2 \gtrsim 1 \ , \tag{27}$$

which we set out to prove.

Acknowledgments

I thank Sam Braunstein for valuable comments on these lecture notes.

References

[1] For a general discussion on covariance and contravariance, see the Wikipedia entry on covariance: `http://en.wikipedia.org/wiki/Covariant`.

[2] G. Weinreich, *Geometrical vectors*, Chicago Lectures in Physics, The University of Chicago Press (1998).

[3] S.L. Braunstein, *Geometry of quantum inference*, Phys. Lett. A **219**, 169 (1996).

[4] S.L. Braunstein and C.M. Caves, *Statistical distance and the geometry of quantum states*, Phys. Rev. Lett. **72**, 3439 (1994).

[5] See also the Wikipedia entry on the Schwarz inequality:
`http://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality`.

[6] T.M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley & Sons Inc. (2006).