The simplex and probability

P.G.L. Porta Mana
Kavli Institute, Trondheim <pgl@portamana.org>

11 January 2019; updated 17 March 2020

A collection of geometric and mathematical facts about simplices of any dimension and about probability distributions on them. This collection is regularly updated

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

I collect here some geometric and mathematical facts about simplices of any dimensions, and about distributions of probability on them. I often use them in my research and thought they can be useful to other people. I'll update this collection regularly. Almost no explicit proofs are given, although sometimes I give the idea of the proof. Please let me know if you find any mistakes or if you think some other fact should be included.

1 Simplex

Consider $N + 1$ mutually exclusive and exhaustive propositions. The distributions of their relative frequencies form an N-dimensional simplex. Label the propositions $\{0, \dots, N\}$, denote a relative-frequency distribution by $(q_0, \dots, q_N) =: \boldsymbol{q}$. In the rest of this memo it's always implicitly assumed that $q_i \geqslant 0$, also in the integration domains. Denote by

$$\Delta_N \coloneqq \{(x_0, \dots, x_N) \mid x_i \geqslant 0, \textstyle\sum_i x_i = 1\} \tag{1}$$

the N-dimensional simplex, or ND-simplex, which can be thought of as embedded in the non-negative orthant of $\mathbf{R}^{N+1}$. It is a convex space, and therefore part of an affine space (Porta Mana 2019).

2 Geometric properties

Let's examine some geometric properties of the ND-simplex, embedded as a symmetric $(N + 1)$-hyperhedron (triangle, tetrahedron, and so on) in N-dimensional Euclidean space – that is, a space with a flat metric. Suppose its edges have length L; note that when it's naturally embedded in the nonnegative orthant, then $L = \sqrt{2}$.

- Its hypervolume is

$$\left(\frac{L}{\sqrt{2}}\right)^{N} \frac{\sqrt{N+1}}{N!} \ . \tag{2}$$

- The hyperarea of a facet (equal to the hypervolume of one less dimension) is

$$\left(\frac{L}{\sqrt{2}}\right)^{N-1} \frac{\sqrt{N}}{(N-1)!} \ . \tag{3}$$

- The distance between the centre and a vertex is

$$\frac{L}{\sqrt{2}}\sqrt{\frac{N}{N+1}} \ , \tag{4}$$

which is also the radius of its circum-hypersphere.

- The distance between the centre and the centre of a facet is

$$\frac{L}{\sqrt{2}}\frac{1}{\sqrt{N\,(N+1)}} \ , \tag{5}$$

which is also the radius of its in-hypersphere.

- The distance between a vertex and its opposite facet is therefore

$$\frac{L}{\sqrt{2}}\sqrt{\frac{N+1}{N}} \ . \tag{6}$$

- The distance between the centre and the origin in the embedding orthant is

$$\frac{L}{\sqrt{2}}\frac{1}{\sqrt{N+1}} \ . \tag{7}$$

Note the following curious facts as N grows:

- the hypervolume is smaller and smaller compared to that of the embedding hypercube;

- the ratio of the radii of circum- and in-hypersphere is N, therefore the ratio between their hypervolumes is N^N: most of the hypervolume of the ND-simplex is ‘at the corners’;

- the centre becomes closer and closer to the origin of the embedding orthant, with respect to the size of the edges;

- the centre becomes closer and closer to the facets with respect to the distance to the edges;
- almost all of the segment from a vertex to its opposite facet is between the vertex and the centre.

3 Densities

As basic density (that is, volume element) we can take either

$$\mathrm{d}q_1 \cdots \mathrm{d}q_N \ , \qquad (q_1, \dots, q_N) \in \Delta_N \ , \tag{8}$$

or

$$\mathrm{d}q_0 \cdots \mathrm{d}q_N \ \delta(1 - \textstyle\sum \boldsymbol{q}) \ , \qquad (q_0, \dots, q_N) \in [0, +\infty[^{N+1} \ , \tag{9}$$

which is borrowed from a Euclidean volume element of the embedding space. The latter leads to more symmetric formulae. The two densities are equivalent, and their integration gives $1/N!$, as can be proven inductively (Δ_k is the base of Δ_{k+1}: multiply its k-volume by a unit height and divide by $k+1$) or as shown in Jaynes (2003 § 18.10). Let's denote either density by $\mathrm{d}\boldsymbol{q}$. When (8) is intended, any q_0 that appears in the integral must be understood as $q_0 \equiv 1 - \sum_{i=1}^{N} q_i$.

4 Flat prior

The N-simplex has a natural convex structure. Thus the ratio of two N-volumes is well-defined. There's only one normalized density that assigns the same degree of belief to any two N-volumes having unit ratio:

$$N! \, \mathrm{d}\boldsymbol{q} \ , \tag{10}$$

called the flat prior.

5 Jeffreys prior

It is also possible to embed the N-simplex into a hyperspherical surface in $[0, +\infty[^{N+1}$ via

$$(q_1, \dots, q_N) \mapsto (x_0, x_1, \dots, x_N) = \left(\sqrt{1 - q_1 - \cdots - q_N}, \sqrt{q_1}, \dots, \sqrt{q_N}\right) . \tag{11}$$

The normalized density induced by the Euclidean one is in this case

$$\Gamma\Bigl(\frac{N+1}{2}\Bigr)\,\frac{\mathrm{d}\boldsymbol{q}}{\prod_{i=0}^{N}\sqrt{\pi q_i}} \tag{12}$$

6 Entropic prior

The relative-frequency distribution $\boldsymbol{q}$ in m observations can be obtained in $\binom{m}{m\boldsymbol{q}}$ ways, where

$$\binom{m}{m\boldsymbol{q}} \coloneqq \frac{m!}{(mq_0)!\cdots(mq_N)!} \approx \exp[m\mathrm{H}(\boldsymbol{q})] \tag{13}$$

is the multinomial coefficient. If we have equal beliefs in the occurrence of these ways and m is very large, our belief about the relative frequency $\boldsymbol{q}$ can be approximated by the density

$$\frac{\exp[m\mathrm{H}(\boldsymbol{q})]}{\int\mathrm{d}\boldsymbol{q}\;\exp[m\mathrm{H}(\boldsymbol{q})]}\,\mathrm{d}\boldsymbol{q}\;, \tag{14}$$

called the entropic prior.

7 Metrics

A density on the simplex doesn't induce any canonical density on a lower-dimensional subset. One way to induce a density on every lower-dimensional subset is to equip the simplex with a metric (Choquet-Bruhat et al. 1996 ch. V; Bossavit 1991 ch. 4). We can define a metric either in terms of intrinsic properties of the simplex or by embedding the simplex in a metric space.

7.1 Flat metric

The flat metric is the one that respects the convex structure of the simplex, in the sense that every two parallel d-dimensional subsets whose d-volumes are in a ratio of one-to-one – this can be calculated using only the convex structure (Porta Mana 2019) – are given equal d-volumes by the metric. The metric also allows the comparison of non-parallel d-volumes, something that can't be done with the convex structure alone. This metric can also be obtained by embedding the N-simplex into the Euclidean space $[0,+\infty[^{N+1}$ via

$$(q_1,\dots,q_N) \mapsto (x_0,x_1,\dots,x_N) = \bigl(1-q_1-\dots-q_N, q_1,\dots,q_N\bigr) \tag{15}$$

and pulling-back the metric.

The embedding above has tangent map

$$\begin{pmatrix} -\boldsymbol{u} \\ \boldsymbol{I}_N \end{pmatrix} \tag{16}$$

with

$$\boldsymbol{u} \coloneqq \underbrace{\begin{pmatrix}1 & \dots & 1\end{pmatrix}}_{N}, \qquad \boldsymbol{I}_N \coloneqq \text{identity } N\text{-matrix} . \tag{17}$$

In coordinates $\boldsymbol{x}$ the metric is represented by the identity matrix $\boldsymbol{I}_{N+1}$. The representation of its pull-back is therefore

$$\begin{pmatrix} -\boldsymbol{u}^\intercal & \boldsymbol{I}_N \end{pmatrix} \boldsymbol{I}_{N+1} \begin{pmatrix} -\boldsymbol{u}^\intercal \\ \boldsymbol{I}_N \end{pmatrix} = \boldsymbol{I}_N + \boldsymbol{u}^\intercal \boldsymbol{u} . \tag{18}$$

This is a matrix with all unit elements outside the diagonal and 2 on the diagonal. Its determinant can be found with Sylvester's theorem (Sylvester 1851; Akritas et al. 1996):

$$\det(\boldsymbol{I}_N + \boldsymbol{u}^\intercal \boldsymbol{u}) = \det(\boldsymbol{I}_1 + \boldsymbol{u}\boldsymbol{u}^\intercal) = 1 + N . \tag{19}$$

This peculiar expression for the metric comes from the fact that q_0 is a function of the other q_i. If we use $(q_0,q_1,\dots,q_N)$ as fictitious coordinates, multiplying by a delta for normalization (§ 1), then the metric is simply expressed by the identity matrix $\boldsymbol{I}_{N+1}$.

8 Projections

***

9 Truncated normal distributions

A truncated normal distribution on the simplex is – we would think – proportional to

$$\exp\Bigl[-\tfrac{1}{2}(\boldsymbol{q}-\boldsymbol{q}^*)^\intercal \boldsymbol{\varSigma}^{-1} (\boldsymbol{q}-\boldsymbol{q}^*)\Bigr]\,\mathrm{d}\boldsymbol{q} . \tag{20}$$

It's truncated because it's defined on the simplex, a bounded space.

Let's first recall some facts about truncated normal distributions on any bounded space, not just on the simplex:

- The normalization constant is not the one of the standard normal on $\mathbf{R}^N$, for obvious reasons of integration domain.
- The centre $\boldsymbol{q}^*$ is *not* the mean of $\boldsymbol{q}$:

$$\mathrm{E}(\boldsymbol{q}) \neq \boldsymbol{q}^* \; ! \tag{21}$$

  That's why it's better to call $\boldsymbol{q}^*$ 'centre' (owing to the symmetries of the distribution function) or 'location parameter'.
- The scale matrix $\boldsymbol{\Sigma}$ is *not* the covariance of $\boldsymbol{q}$:

$$\mathrm{E}(\boldsymbol{q}\boldsymbol{q}^\intercal) \neq \boldsymbol{\Sigma} \; ! \tag{22}$$

  That's why it's better to call $\boldsymbol{\Sigma}$ 'scale form' than 'covariance matrix'.
- The scale form $\boldsymbol{\Sigma}$ can also be negative-definite or indefinite. In the directions where its singular values are negative, the resulting distribution function in $\boldsymbol{q}$ coordinates is ∪-shaped and has maxima at the boundaries. (Positive-definiteness is only required on unbounded spaces such as $\mathbf{R}^N$: negative values would try to push the probability mass to infinity.)
- If most mass is enough away from the boundaries (the distribution function is very peaked), so that there's little difference for the integrals if we extend them outside the simplex, then the covariance matrix will be approximately equal to the scale form $\boldsymbol{\Sigma}$.

There are additional subtleties with the mathematical expression (20), though, if it's expressed in affine coordinates; that is, if $\boldsymbol{q}$ is represented by an $(N+1)$-tuple summing up to one and the formula is conceived in the positive orthant.

First of all, remember from § 8 that $\boldsymbol{u}^\intercal \boldsymbol{q} = 1$ for every point $\boldsymbol{q}$ and $\boldsymbol{q}''$ of the simplex. This means that – as long as our integrals are strictly over the simplex – we can equivalently rewrite the normal (20) as

$$\begin{gathered}\exp\Big[-\tfrac{1}{2}(\boldsymbol{q}-\boldsymbol{q}^*)^\intercal\, \hat{\boldsymbol{\Sigma}}^{-1}\,(\boldsymbol{q}-\boldsymbol{q}^*)\Big]\;\mathrm{d}\boldsymbol{q}\\ \text{with}\quad \hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{h}\boldsymbol{u}^\intercal + \boldsymbol{u}\boldsymbol{h}^\intercal\end{gathered} \tag{23}$$

with *arbitrary* $(N+1)$-vector $\boldsymbol{h}$. Note that $\boldsymbol{\Sigma}^{-1}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ have generally different (pseudo)inverses.

Second, from expression (23) we see that any truncated normal can also be equivalently written in the form

$$\exp\left(-\tfrac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{\Delta}^{-1}\boldsymbol{q}\right)\,\mathrm{d}\boldsymbol{q} \tag{24}$$

with

$$\boldsymbol{\Delta}^{-1} \coloneqq \boldsymbol{\Sigma}^{-1} - \boldsymbol{u}\boldsymbol{q}^{*\intercal}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{q}^{*}\boldsymbol{u}^\intercal + \boldsymbol{u}\boldsymbol{q}^{*\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{q}^{*}\boldsymbol{u}^\intercal\ . \tag{25}$$

Also in this case $\boldsymbol{\Delta}^{-1}$ is determined but by an additional $k\boldsymbol{U}$ term which leads to a constant term that can be absorbed by the normalization constant.

These two subtleties raise two questions. If we see expressions such as (20) or (23), how do we know that its $\boldsymbol{\Sigma}^{-1}$ or $\boldsymbol{\Delta}^{-1}$ is *really* an approximation of the inverse covariance matrix? what if a $k\boldsymbol{U}$ term was actually added to it? And how do we know what the centre $\boldsymbol{q}^*$ of the distribution is in the expression (23)?

The possibility of expression (23) can be understood intuitively as follows. A truncated normal on an N-dimensional space has N location parameters and $N(N+1)/2$ scale parameters. But $\boldsymbol{\Delta}^{-1}$ in expression (23) comprises $(N+1)(N+2)/2 \equiv N(N+1)/2 + N + 1$ parameters: of the additional $N+1$, one is superfluous – leading to the freedom of a constant additive term – and the rest encode the location parameter.

10 Diverse

Suppose we have a covariance matrix – equivalent to the inverse of a quadratic form – on a simplex. In the canonical embedding orthant, this covariance matrix has a zero axis, corresponding to the plane of the embedded simplex. Taking the *pseudo*inverse leaves this axis to zero while inverting the others to give the quadratic form. This doesn't affect the results for vectors tangent to the simplex, since they have no components along that axis.

Consider the projection from a higher-dimensional simplex to the previous one via a hypergeometric distribution ('drawing without replacement'). This is done via a projection matrix $\boldsymbol{M}$, with the hypergeometric distribution as its entries.

We can pull back the quadratic form from the lower-D to the higher-D simplex, obtaining a degenerate quadratic form. This is done by multiplying the quadratic form on the left and right by $\boldsymbol{M}^\intercal$ and $\boldsymbol{M}$. This form has a covariance matrix with several infinite axes, corresponding to the kernel of the projection, and the zero axis corresponding to the direction outside of the simplex.

- The projection matrix has a right inverse (pseudoinverse, and different from the transpose).
- The projection matrix maps the uniform distribution to the uniform distribution; the inverse projection also maps the uniform distribution to the uniform distribution.
- The previous fact implies that the projection and its pseudoinverse map the outside axes of covariance matrices in the two simplices to each other.
- We can 'add thickness' to the axis outside of the simplex by adding a constant term to *all* entries of the covariance matrix.
- The projection and its pull-back map the outside axis to itself.
- Therefore we can add a constant term to the low-D covariance matrix or to its pull-back.
- The pseudoinverse of the pulled-back quadratic form – that is, its covariance matrix – can be obtained by multiplying original covariance matrix by the pseudoinverses of the projectors. This, however, can only be done as long as we leave the axis outside the simplex equal to zero (so that the pseudoinverse doesn't touch it).
- The pseudoinverse maps a spher

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Akritas, A. G., Akritas, E. K., Malaschonok, G. I. (1996): *Various proofs of Sylvester's (determinant) identity*. Math. Comput. Simulat. **42**$^{4–6}$, 585–593. http://www.inf.uth.gr/~akritas/articles/38.pdf.

Bossavit, A. (1991): *Differential Geometry: for the student of numerical methods in Electromagnetism*. https://www.researchgate.net/publication/200018385_Differential_Geometry_for_the_student_of_numerical_methods_in_Electromagnetism.

Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1996): *Analysis, Manifolds and Physics. Part I: Basics*, rev. ed. (Elsevier, Amsterdam). First publ. 1977.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Porta Mana, P. G. L. (2019): *Affine and convex spaces: blending the analytic and geometric viewpoints*. Open Science Framework doi:10.17605/osf.io/jw9f6, arXiv:1104.0032. First publ. 2011.

Sylvester, J. J. (1851): *On the relation between the minor determinants of linearly equivalent quadratic functions*. Phil. Mag. **1**^{4}, 295–305.