November 14, 2003

Hyperplane Priors

V. Dose

Centre for Interdisciplinary Plasma Science, Max-Planck-Institut für Plasmaphysik, EURATOM Association, Boltzmannstrasse 2, D-85748 Garching bei München, Germany

Abstract. The requirement of transformation invariance of a probability distribution is employed to derive prior probabilities for the coefficients of the equation describing a hyperplane. In two dimensions, this is a straight line, in three dimensions an ordinary plane etc. We treat the general case of n dimensions and propose a procedure to normalize the resulting distributions in order to make them proper and appropriate for model comparison problems.

INTRODUCTION

Fitting a straight line in two dimensions to a set of experimental data is one of the most important traditional regression problems in physics [1, 2]. The usual procedure employs the least squares method meaning that in a Bayesian sense the prior probability for the coefficients of the straight line is taken flat, improper. This choice is often a reasonable approximation but in a strict sense unacceptable since it violates the requirement of transformation invariance because a flat prior in a given system of coordinates does in general not remain flat after translations and rotations of that system. If the prior probability is uninformative, then we must insist that it remain invariant under such basic transformations. As a first introductory example we shall derive the prior distribution for a straight line in two dimensions parallel to one of the axes of the coordinate system. The only sensible transformation in this case is a translation. Our second introductory example will be the prior probability for the coefficient of a straight line in two dimensions passing through the origin. The only sensible coordinate transform for this case is a rotation of the coordinate system. We show that the resulting prior distribution satisfies intuition. These two special elementary cases constitute the elements of the general case of an $(n-1)$ dimensional hyperplane in n-dimensional space.

THE TRANSFORMATION INVARIANCE EQUATION

Consider a probability density of a set of parameters $\vec{a}, p(\vec{a})$. $p(\vec{a})da_1\cdots da_n$ is then an element of probability mass whose value must be independent from the system of coordinates which we use to evaluate its numerical value. Hence, for a different system of coordinates $\vec{a}'$ we must require that

$$p(\vec{a})da_1\cdots da_n = p(\vec{a}')da_1'\cdots da_n' \tag{1}$$

yielding the functional equation

$$p(\vec{a}) = p(\vec{a}') \cdot \left| \frac{\partial(a_1' \cdots a_n')}{\partial(a_1 \cdots a_n)} \right| \tag{2}$$

where the determinant on the right-hand side is the Jacobian of the transformation $\vec{a} \to \vec{a}'$. Since any finite transformation $\vec{a} \to \vec{a}'$ can be constructed from an appropriate sequence of infinitesimal transformations it is sufficient to consider a single infinitesimal transformation. We denote the infinitesimal transformation which maps $\vec{a}$ onto $\vec{a}'$ as $T_\varepsilon(\vec{a})$. Equation (2) becomes then

$$p(\vec{a}) = p(T_\varepsilon(\vec{a})) \cdot \left| \frac{\partial T_\varepsilon(\vec{a})}{\partial \vec{a}} \right| \quad . \tag{3}$$

The right-hand side of (3) is a function of ε, the left hand side is independent of ε. Hence, by differentiating (3) by ε we obtain

$$\frac{\partial}{\partial \varepsilon} \left\{ p(T_\varepsilon(\vec{a})) \left| \frac{\partial T_\varepsilon(\vec{a})}{\partial \vec{a}} \right| \right\}_{\varepsilon=0} = 0 \tag{4}$$

the general functional equation which satisfies the requirement of transformation invariance. Well known special cases are obtained for $\vec{a}$ scalar and subject to the similarity transformation $a' = a(1+\varepsilon)$

$$\frac{\partial}{\partial \varepsilon} \{ p(a(1+\varepsilon))(1+\varepsilon) \}_{\varepsilon=0} = 0 \quad , \tag{5}$$

$$p'(a) \cdot a + p(a) = 0 \to p(a) = \frac{1}{a} \tag{6}$$

and, for a subject to a translation $a' = a + \varepsilon$,

$$p'(a) = 0 \to p(a) = \text{const.} \tag{7}$$

The former is the well known Jeffreys prior for a scale variable, the latter the improper prior for a location variable.

COORDINATE TRANSFORMS

The general problem which we are going to solve is to find a prior distribution for the coefficients $\vec{a}$ of the linear form

$$a_1 x_1 + a_2 x_2 + \cdots a_n x_n + 1 = 0 \quad . \tag{8}$$

In contrast to the examples of scale and location variable priors where the transform concerned directly the parameters, we shall now investigate the consequences of a transformation applied to $\vec{x}$. First we investigate one of the two simplest cases namely

$$a_1 x_1 + 1 = 0 \quad . \tag{9}$$

In two dimensional space this equation describes a straight line parallel to the x_2 axis. The only sensible transformation on x_1 is in this case a translation $x_1' = x_1 + \varepsilon$. In the transformed coordinate system we have

$$(x_1 + \varepsilon)a_1' + 1 = 0 \quad . \tag{10}$$

Ordering terms we get

$$x_1 \cdot \frac{a_1'}{1 + \varepsilon a_1'} + 1 = 0 \quad . \tag{11}$$

Comparison with (9) yields the transformation in a which is implied by the translation of x_1

$$a_1 = \frac{a_1'}{1 + \varepsilon a_1'} \rightarrow a' = \frac{a}{1 - a\varepsilon} \approx a + \varepsilon a^2 \tag{12}$$

where the latter equality results from a neglect of all terms of order ε^2 and higher. This is in line with the infinitesimal character of the transformation. The Jacobian of the transformation is readily calculated to be $(1 + 2a\varepsilon)$. Substituting these results into (4) we arrive at

$$\begin{aligned} \frac{d}{d\varepsilon}\{p(a + \varepsilon a^2)(1 + 2a\varepsilon)\}_{\varepsilon=0} = 0 \quad , \\ p'a^2 + 2ap = 0 \quad . \end{aligned} \tag{13}$$

Assuming $a \neq 0$, this differential equation may be simplified and easily solved to give

$$p(a) = \frac{1}{a^2} \quad . \tag{14}$$

This distribution is normalizable, because of the restriction $a \neq 0$. The question which cutoff a_{min} to choose will be addressed after treating the general case.

The second elementary example which we will treat, is a straight-line passing through the origin. The equation for this case is

$$ax_1 + x_2 = 0 \tag{15}$$

The only sensible transformation of the $\vec{x}$-coordinate system is in this case a rotation. Let φ be the angle of rotation between the $\vec{x}'$ and the $\vec{x}$ coordinate system. The transformation is then

$$\begin{aligned} x_1' &= x_1 \cos\varphi - x_2 \sin\varphi \approx x_1 - \varepsilon x_2 \quad , \\ x_2' &= x_1 \sin\varphi + x_2 \cos\varphi \approx \varepsilon x_1 + x_2 \quad . \end{aligned} \tag{16}$$

We substitute (16) into (15) and obtain

$$a'(x_1 - \varepsilon x_2) + \varepsilon x_1 + x_2 = 0 \quad . \tag{17}$$

Ordering in terms of x_1 and x_2 yields the implied transformation on a

$$a' = \frac{a+\varepsilon}{1-a\varepsilon} \approx a + \varepsilon(1+a^2) \quad . \tag{18}$$

The Jacobian of this transformation is $(1+2a\varepsilon)$ yielding the functional equation

$$\frac{\partial}{\partial \varepsilon}\{p(a+\varepsilon(1+a^2))(1+2a\varepsilon)\}_{\varepsilon=0} = 0 \tag{19}$$

and the associated differential equation

$$p'(a^2+1) + 2ap = 0 \tag{20}$$

which is readily solved to yield

$$p(a) = \frac{1}{\pi}\frac{1}{1+a^2} \quad . \tag{21}$$

The factor $1/\pi$ arises of course not from the solution of (20) but from subsequent normalization. This result can easily be shown to conform with intuition. Noting that the meaning of a is $a = -tg\psi$, where ψ is the angle of the straight line with respect to the positive x_1-axis, the corresponding distribution of ψ is obtained from (2) as $p(\psi) = 1/\pi$. This means that the uninformative prior for a straight line passing through the origin assumes every orientation to be equally probable.

After these two prototypic examples we are well prepared to treat the general case of an $(n-1)$ dimensional plane in n-dimensional space. The equations for the hyperplane in the primed and unprimed coordinate systems are

$$a_1 x_1 + a_2 x_2 + \cdots a_n x_n + 1 = 0 \quad , \tag{22}$$

$$a'_1 x'_1 + a'_2 x'_2 + \cdots a'_n x'_n + 1 = 0 \quad . \tag{23}$$

The general rotation in n-dimensional space may be expressed as a sequence of two-axis (j,k) rotations [5]. There exist $\binom{n}{2}$ such rotations e.g., one for $n=2$, three for $n=3$, six for $n=4$ etc. Now we perform one such rotation: $x'_i = x_i$ for all i not equal to either j or k and

$$\begin{aligned} x'_j &= x_j - \varepsilon x_k \quad , \\ x'_k &= \varepsilon x_j + x_k \quad . \end{aligned} \tag{24}$$

Substituting the primed coordinates into (23) and collecting coefficients of x_i yields the implied transformation $a'_i = a_i$ for all i not equal to either j or k and

$$\begin{aligned} a_j &= a'_j + \varepsilon a'_k \\ a_k &= a'_k - \varepsilon a'_j \end{aligned} \quad \longrightarrow \quad \begin{aligned} a'_j &= a_j - \varepsilon a_k \\ a'_k &= a_k + \varepsilon a_j \end{aligned} \tag{25}$$

The Jacobian of the transformation is

$$\frac{\partial(a_1'\ldots a_n')}{\partial(a_1\ldots a_n)} = \begin{vmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & 1 & \ldots & -\varepsilon & & \\ & & & \ddots & & & \\ & & \vdots & 1 & \vdots & & \\ & & & & \ddots & & \\ & & \varepsilon & \ldots & 1 & & \\ & 0 & & & & \ddots & \\ & & & & & & 1 \end{vmatrix} = 1+\varepsilon^2 \qquad (26)$$

and the general transformation invariance functional equation (4) leads to the partial differential equation

$$a_k \frac{\partial p}{\partial a_j} = a_j \frac{\partial p}{\partial a_k} \qquad (27)$$

from which follows that p can only be a function of $a_j^2 + a_k^2$. Since this must hold for any pair j,k we conclude that

$$p(a_1, a_2 \cdots a_n) = p(a_1^2 + a_2^2 + \cdots a_n^2) \quad . \qquad (28)$$

Now p must also satisfy invariance under the n possible translations

$$x_i' = x_i \quad \forall \quad i \neq k, \quad x_k' = x_k + \varepsilon \quad . \qquad (29)$$

We substitute (29) into (23) and obtain

$$\sum_i \frac{a_i' x_i}{1+\varepsilon a_k'} + 1 = 0 \qquad (30)$$

from which we read the implied transformation

$$a_i = \frac{a_i'}{1+\varepsilon a_k'} \quad \rightarrow \quad a_i' = a_i + \varepsilon a_i a_k \qquad (31)$$

with Jacobian

$$\frac{\partial(a_1'\ldots a_n')}{\partial(a_1\ldots a_n)} = \begin{vmatrix} 1+\varepsilon a_k & & & & \\ & \ddots & & 0 & \\ & & 1+2\varepsilon a_k & & \\ & 0 & & \ddots & \\ & & & & 1+\varepsilon a_k \end{vmatrix} = 1+(n+1)\varepsilon a_k \, . \qquad (32)$$

The unknown function p must therefore obey the second set of partial differential equations

$$a_1 a_k \frac{\partial p}{\partial a_1} + \cdots a_k^2 \frac{\partial p}{\partial a_k} \cdots + a_n a_k \frac{\partial p}{\partial a_k} + a_k (n+1) \cdot p = 0 \quad , \qquad (33)$$

Equation (33) may be simplified if we require $a_k \neq 0$ to

$$\sum_i a_i \frac{\partial p}{\partial a_i} + (n+1)p = 0 \quad . \tag{34}$$

Translation of other coordinates, say x_j, leads to exactly the same differential equation (34) however with side condition $a_j \neq 0$. We now use the previous result (28) that p can only be a function of $\rho = a_1^2 + a_1^2 + \cdots a_n^2$. Then $\partial p/\partial a_i = 2a_i dp/d\rho$ and we arrive at

$$2\sum_i a_i^2 \frac{dp}{d\rho} + (n+1)p = 0 \tag{35}$$

which is readily integrated to yield

$$p(a_1 \cdots a_n) = \frac{1}{\left\{a_1^2 + \cdots + a_n^2\right\}^{\frac{n+1}{2}}} \quad , |a_i| > 0 \quad . \tag{36}$$

Note that this general result contains our previous simple case (14) for $n = 1$ as well as the priors for the coefficients of a straight-line in two dimensions and an ordinary plane in three dimensions derived by Kendall and Moran [3].

NORMALIZATION

The condition $|a_i| > 0$ in (36) means of course that the distribution is normalizable. Denote by $r^2 = a_1^2 + \cdots a_n^2$ and by r_0 the minimum value of r which we allow. The normalization integral Z is then

$$Z = \int \mathrm{d}\Omega_n \int_{r_0}^{\infty} \frac{\mathrm{d}r}{r^2} = \frac{2\pi^{n/2}}{\Gamma(n/2)} \cdot \frac{1}{r_0} \tag{37}$$

and the normalized distribution becomes for all $n \geq 1$

$$p(a_1 \cdots a_n) = r_0 \frac{\Gamma(n/2)}{2\pi^{n/2}} \cdot \frac{1}{\left\{a_1^2 + \cdots + a_n^2\right\}^{\frac{n+1}{2}}} \quad . \tag{38}$$

The remaining question concerns of course the choice of r_0. Let $x_1^*, \ldots, x_n^*$ be a point through which the hyperplane is expected to pass, then for this point (22) holds. We now answer the question: "*What is the minimum value of r^2 given this point?*". The Lagrangian optimization problem is to find the minimum of

$$\Phi = a_1^2 + \cdots + a_n^2 + 2\lambda(a_1 x_1^* + \cdots + a_n x_n^* + 1) \quad . \tag{39}$$

The derivative with respect to a_k yields

$$\frac{\partial \Phi}{\partial a_k} = 2\tilde{a}_k + 2\lambda x_k^* = 0 \quad \rightarrow \quad \tilde{a}_k = -\lambda x_k^* \quad . \tag{40}$$

The derivative with respect to λ yields the hyperplane equation into which we substitute the solution (40) to obtain

$$-\lambda x_1^{*2} \cdots - \lambda x_n^{*2} + 1 = 0 \quad \rightarrow \quad \lambda = \left(\sum x_i^{*2}\right)^{-1} \quad . \tag{41}$$

So the coefficients $\tilde{a}_i$ become explicit and r_0^2 turns out to be

$$r_0^2 = \sum \tilde{a}_i^2 = \left(\sum x_i^{*2}\right)^{-1} \quad . \tag{42}$$

A suitable choice of r_0 can therefore be obtained from that point $\vec{x}^*$ which lies farthest from the origin of the data cloud. Clearly distances are invariant under rotations but not under translations. This suggests to dispose of the origin of the $\vec{x}$ coordinate system. The highest lower limit of r_0 is obviously obtained if the origin is chosen to lie in the centre of gravity of the data vectors $\vec{x}_j$.

LOWER DIMENSIONAL HYPERPLANES

In the last section we have derived the prior probability for an $(n-1)$ dimensional hyperplane in *n*-dimensional space. Though we believe that this is, from a practical point of view, the analysis of experimental data, the by far most important case it remains a matter of principle to show how to deal with lower, $(n-k)$ dimensional hyperplanes. The key to this problem is the parameter representation of a $(k \leq n-1)$ dimensional vector space. The vector $\vec{x}$ representing a point of this space is written as

$$\vec{x} = \vec{r} + \lambda_1 \vec{a}_1 + \cdots + \lambda_k \vec{a}_k \tag{43}$$

where $\vec{r}$ is the vector which fixes the origin of the vector space $\vec{a}_1, \ldots, \vec{a}_k$ which in turn spans the $(n-k)$ dimensional hyperplane. The vector equation (43) summarizes *n* scalar equations from which we may eliminate the arbitrary coefficients of the linear combination of $\{\vec{a}_i\}, \vec{\lambda}$. This leaves us with $(n-k)$ equations. A straight-line $(k=1)$ in three dimensions is therefore given by two equations. There are obtained by choosing one of the three equations (43) to express λ_1 in terms of $\vec{x}, \vec{r}, \vec{a}_1$ and then eliminate λ_1 from the remaining two equations. Choosing $\lambda_1 = (x_1 - r_1)/a_1$ the set of equations representing this straight-line is

$$\begin{aligned} \left(-\tfrac{a_2}{a_1}\right) x_1 \quad + \quad x_2 \qquad\qquad &= \quad \left(r_2 - \tfrac{a_2}{a_1} r_1\right) \quad , \\ \left(-\tfrac{a_3}{a_1}\right) x_1 \qquad\qquad + \quad x_3 \quad &= \quad \left(r_3 - \tfrac{a_3}{a_1} r_1\right) \quad . \end{aligned} \tag{44}$$

A straight-line in three dimensions is therefore characterized by four parameters, the quantities which we have put in brackets in (44). Introducing the definitions $a_2/a_1 = \alpha, a_3/a_1 = \beta, p = r_2 - a_2 r_1/a_1$ and $q = r_3 - a_3 r_1/a_1$ (44) becomes

$$\begin{aligned} x_2 &= \alpha x_1 + p \quad , \\ x_3 &= \beta x_1 + q \quad , \end{aligned} \tag{45}$$

and the transformation invariance prior for α, β, p, q is determined from the transformations implied on α, β, p, q by translating (3x) and rotating (3x) the $\vec{x}$-coordinate system. We realize from inspection of (45) that there is some equivalence between x_2, x_3, but operations on x_1 (either translation of x_1 or rotation about x_1) will yield different implied transformations. It remains to be said that the necessary calculations are simple but lengthly. They lead to a prior probability independent of p, q

$$p(\alpha,\beta) = \frac{1}{\pi}\frac{1}{(1+\alpha^2+\beta^2)^2} \quad . \tag{46}$$

This result has previously been obtained also by Kendall and Moran [3]. The discussion above leads us to suspect that it will be difficult if not impossible to derive a general result for an $(n-k)$ dimensional hyperplane due to the in-equivalence of the coordinates which results from the process of parameter elimination from the general equation (44).

SUMMARY AND CONCLUSIONS

Prior probabilities for the coefficients of an $(n-1)$ dimensional vector space have been derived for arbitrary *n*. A method has been devised to turn these distributions into proper priors which are required for a Bayesian model comparison. Important applications of these prior distributions arise in multivariate linear regression and in neural networks modeling of experimental data. A paper with an application in the latter class of problems is in preparation [4].

REFERENCES

1. Montgomery, D. C., and Peck, E. A., *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York, 1982.
2. Gull, S. F., "Bayesian data analysis: straight line fitting," in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, Kluwer Academic, Dordrecht, 1989, p.511.
3. Kendall, M. G., and Moran, P. A. P., *Geometrical Probability*, Griffin, London, 1963.
4. von Toussaint, U., Dose, V., and Gori, S., in preperation.
5. Remember the introduction of Euler angles in classical mechanics. See e.g. Landau, L. D. und Lifschitz E. M., *Lehrbuch der theoretischen Physik I,* Akademie Verlag, Berlin, 1962.