Entropy **2015**, *17*, 3898-3912; doi:10.3390/e17063898

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

General Hyperplane Prior Distributions Based on Geometric Invariances for Bayesian Multivariate Linear Regression

Udo von Toussaint

Max-Planck-Institute for Plasmaphysics, Boltzmannstrasse 2, 85748 Garching, Germany;
E-Mail: udo.v.toussaint@ipp.mpg.de; Tel.: +49-8932991817

Academic Editors: Frédéric Barbaresco and Ali Mohammad-Djafari

Received: 30 March 2015 / Accepted: 2 June 2015 / Published: 10 June 2015

Abstract: Based on geometric invariance properties, we derive an explicit prior distribution for the parameters of multivariate linear regression problems in the absence of further prior information. The problem is formulated as a rotationally-invariant distribution of L-dimensional hyperplanes in N dimensions, and the associated system of partial differential equations is solved. The derived prior distribution generalizes the already known special cases, e.g., 2D plane in three dimensions.

Keywords: prior probabilities; hyperplanes; geometrical probability; neural networks

1. Introduction

In the context of Bayesian probability theory, a proper assignment of prior probabilities is crucial. Depending on the domain, quite different prior information can be available. It may be in the form of point estimates provided by domain experts (see, e.g., [1] for prior distribution elicitation) or in the form of invariances (of the prior knowledge) of the system of interest, which should be reflected in the prior probability density [2]. However, especially for the ubiquitous case of the estimation of parameters of linear equation systems (like a straight line or hyperplane fitting), the latter requirement is often violated. Consider, for concreteness, the simple case of $y = ax$, a straight line through the origin, with a the parameter of interest. Here, the commonly-applied prior is constant, $p(a \mid I) = \text{const.}$, often accompanied by statements like “Since we do not have specific prior information, we chose a uniform prior on a...”. In Figure 1 on the left-hand side, 15 random samples generated from this prior distribution with $a \in [0, 50]$ are displayed. Confronted with this result, the typical response is (at least in the experience of the author) that instead, a more "uniform" prior distribution of the slopes was intended, which is often depicted like in Figure 1 on the right-hand side. This plot was generated from a prior distribution that has an equal probability density for the angle of the line to the abscissa, corresponding to

$$p\left(a \mid I\right) \sim \frac{1}{\left(1+a^2\right)^{3/2}}. \tag{1}$$

Additionally, in fact, in practice, the units of the axes are commonly chosen in such a way that extreme values of the slopes are not *a priori* overrepresented. If we generalize this requirement to more than one independent or dependent variable, then the desired prior probability should be invariant under arbitrary rotations in this parameter space. Some important special cases have been given already in [3], e.g., for a 1D line in two dimensions or a 2D plane in three dimensions. There also, the governing transformation invariance equation underlying invariant priors is derived. These special cases have since then been generalized to invariant priors for $(N-1)$-dimensional hyperplanes in N-dimensional space; see, e.g., [4]. These hyperplane priors proved to be valuable for Bayesian neural networks [5], where the specific properties of the prior density favored node-pruning instead of simple edge pruning of standard (quadratic) weight regularizers. This is especially helpful for a Bayesian approach to fully-connected deep convolutional networks; see e.g., [6,7].

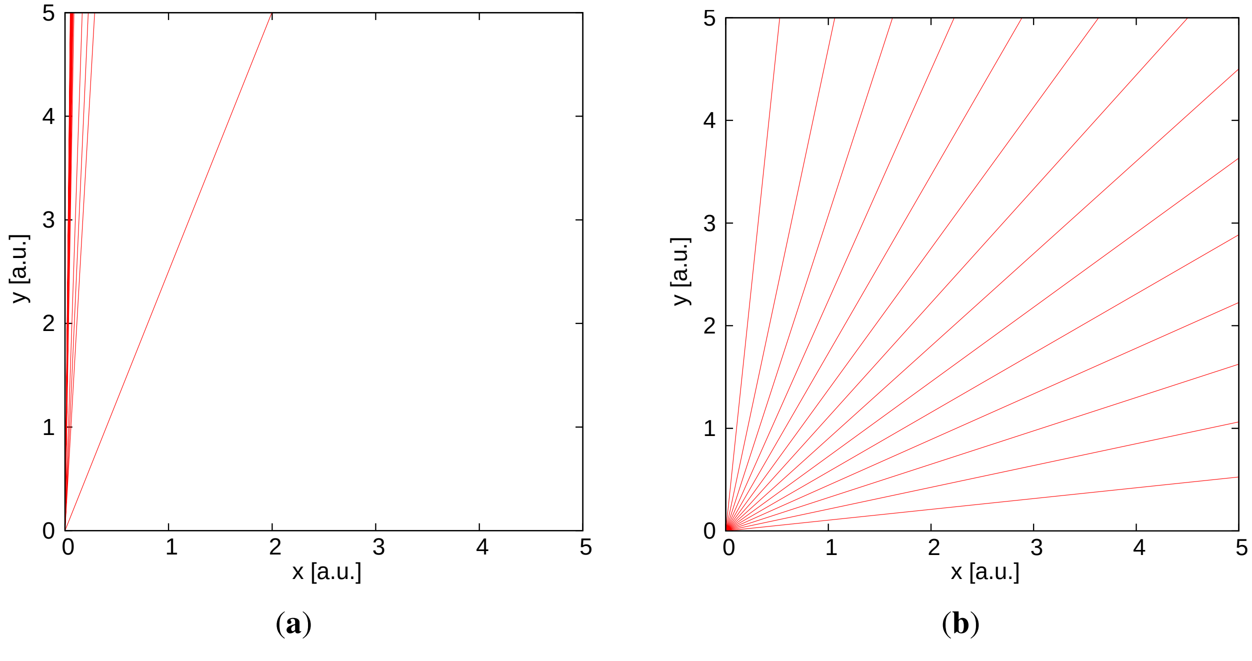

Figure 1. Comparison of two different priors. (**a**) 15 random samples drawn from $p\left(a \mid I\right) = 1/50$, *i.e.*, a uniform distribution in the slope with $0 \leq a \leq 50$. (**b**) the density $p\left(a \mid I\right) \sim \left(1+a^2\right)^{-3/2}$, corresponding to a distribution uniform in the angle, is visualized by 15 samples.

Nevertheless, the general case of prior probability densities for L-dimensional hyperplanes in N-dimensions $(N > L)$ in a suitable parameterization has not been available so far. It has even been conjectured that it is impossible to derive a general solution [8]. Luckily, this conjecture has been too pessimistic, and an explicit formula for the prior density, which can directly be applied to linear regression problems, is derived below.

It should be pointed out that multivariate regression is of course a longstanding topic in Bayesian inference. with classical contributions, e.g., by Box and Tiao [9], Zellner [10] or West [11]. However, the standard approach is based on the use of conjugate priors (instead of invariance priors), mostly for computational convenience [12]. In contrast, the subsequently derived prior distribution is determined by the basic desideratum of consistency if the available prior information is invariant under the considered transformations (*i.e.*, rotations). Whether this invariance holds depends on the considered problem and must not be assumed without further consideration (similar to the case of flat priors for the coefficients). For example, the assumption of rotation invariance may not be suitable for covariates with different underlying units (e.g., m^2, kg).

2. Problem Statement

In standard notation, a multivariate regression model is notated as follows:

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{t}, \qquad x_i \in \mathbb{R}^L, \mathbf{A} \in \mathbb{R}^{M\times L}, \mathbf{t} \in \mathbb{R}^M \text{ and } y_i \in \mathbb{R}^M, \tag{2}$$

with:

$$\mathbf{z}_i = \mathbf{y}_i + \epsilon_i, \epsilon_i \in \mathbb{R}^M, \tag{3}$$

where $\mathbf{z}_i$ is the response vector, $\mathbf{y}_i$ the model value vector, $\mathbf{x}_i$ the vector of the L covariates for observation i, $\mathbf{t}$ the intercept vector and $\mathbf{A}$ the $M \times L$-dimensional matrix of adjacent regression coefficients. The observation noise ϵ_i of each data point is often considered as Gaussian distributed, $\epsilon_i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right)$. This regression model can also be considered as estimating the "best" L-dimensional hyperplane in an N-dimensional space, because in an N-dimensional space, an L-dimensional hyperplane is given by:

$$\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1L}x_L + t_1 \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2L}x_L + t_2 \\
y_3 &= a_{31}x_1 + a_{32}x_2 + \cdots + a_{3L}x_L + t_3 \\
&\vdots \\
y_M &= a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{ML}x_L + t_M
\end{aligned} \tag{4}$$

with $M = N - L$.

The quantity of interest is the prior probability density $F\left(\mathbf{A}\right) = F\left(a_{11}, \cdots, a_{ML}, t_1, \cdots, t_M | I\right)$ for the coefficients $a_{11}, \cdots, a_{ML}, t_1, \cdots, t_M$, which remains invariant under translations and rotations of the coordinate system.

3. Derivation

Using the transformation invariance equation derived in [3]:

$$\sum_{i=1}^{N} \frac{\partial}{\partial z_i}\left(F\left(z_1, \cdots, z_N\right) g_i\left(z_1, \cdots, z_N\right)\right) = 0 \tag{5}$$

for infinitesimal transformations of the form $z_i^{'} = z_i + \epsilon g_i (z_1, \cdots, z_N)$, we can establish a partial differential equation system for F.

3.1. Invariance under Translations

Let us first consider a translation with respect to $y_i : y_i^{'} = y_i + \epsilon$, *i.e.*, $g_i = 1, g_{j,j \neq i} = 0$. Then, the equation in the primed variables reads:

$$y_i^{'} = y_i + \epsilon = a_{i1}^{'} x_1 + a_{i2}^{'} x_2 + \cdots + a_{iL}^{'} x_L + t_i^{'} \tag{6}$$

Collecting the coefficients yields $t_i^{'} = t_i + \epsilon$, and therefore, Equation (5) results in:

$$0 + \cdots + 0 + \frac{\partial}{\partial t_i} (F(\mathbf{A}, \mathbf{t}) \cdot 1) + 0 + \cdots + 0 = 0, \tag{7}$$

which holds for any i. Therefore, $F(\mathbf{A}, \mathbf{t})$ can be a function of $\vec{a}$ only. Since $F(\mathbf{A} \mid I)$ does not depend on $\vec{t}$, the prior distribution is improper (not normalizable in $\mathbf{t}$) as long as there are no limits on the magnitude of $\mathbf{t}$.

The translation with respect to x_i results in the same conclusion.

3.2. Invariance under Rotations

The general rotation in n-dimensional space may be expressed as a sequence of rotations around rotation axes, which are perpendicular to the planes spanned by appropriately-chosen pairs of coordinate system basis vectors [13]. This is based on the fact that any orthogonal matrix, *i.e.*, rotation matrices, can be written uniquely as a product of 2 × 2 rotations. To avoid convoluted language, we denote in the following the rotation around the rotation axis that is perpendicular to the plane spanned by the linear combination of the basis vectors e_i and e_j simply as rotation in the $x_i x_j$-plane.

3.2.1. Rotation in the $x_i x_j$-Plane

Now, we perform one such infinitesimal 2 × 2-rotation for independent parameters around an arbitrary rotation axis perpendicular to the plane spanned by e_i and e_j, preserving all other coordinates: $x_k^{'} = x_k \quad \forall \quad k \neq (j, i)$ and

$$\begin{aligned} x_i^{'} &= x_i - \epsilon x_j, && (8) \\ x_j^{'} &= \epsilon x_i + x_j. && (9) \end{aligned}$$

Substituting the primed coordinates into Equation (5) yields the implied transformations:

$$\begin{aligned} a_{ki}^{'} &= a_{ki} - a_{kj}\epsilon, && (10) \\ a_{kj}^{'} &= a_{kj} + a_{ki}\epsilon, && (11) \\ t_k^{'} &= t_k && (12) \end{aligned}$$

and, therefore, the partial differential equation:

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{ki}} (F(\mathbf{A}) \cdot (-a_{kj})) + \sum_{k=1}^{M} \frac{\partial}{\partial a_{kj}} (F(\mathbf{A}) \cdot (a_{ki})) = 0. \tag{13}$$

3.2.2. Rotation in the $y_i y_j$-Plane

Now, we perform one such rotation in the plane of two dependent parameters e_i and e_j; thus $y_k^{'} = y_k \quad \forall \quad k \neq (j, i)$ and:

$$\begin{aligned} y_i^{'} &= y_i - \epsilon y_j, & (14) \\ y_j^{'} &= \epsilon y_i + y_j. & (15) \end{aligned}$$

Substituting the primed coordinates into Equation (5) yields the implied transformations:

$$\begin{aligned} a_{ik}^{'} &= a_{ik} - a_{jk}\epsilon, & (16) \\ a_{jk}^{'} &= a_{jk} + a_{ik}\epsilon, & (17) \\ t_i^{'} &= t_i - t_j\epsilon, & (18) \\ t_j^{'} &= t_j + t_i\epsilon, & (19) \\ t_k^{'} &= t_k & (20) \end{aligned}$$

and, therefore, the partial differential equation:

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{ik}} \left(F\left(\mathbf{A}\right) \cdot \left(-a_{jk}\right)\right) + \sum_{k=1}^{M} \frac{\partial}{\partial a_{jk}} \left(F\left(\mathbf{A}\right) \cdot \left(a_{ik}\right)\right) = 0. \tag{21}$$

3.2.3. Rotation in a Plane Spanned by $x_i y_j$-Axes

Performing a rotation in the xy-plane, we obtain:

$$\begin{aligned} x_i^{'} &= x_i - \epsilon y_j, & (22) \\ y_j^{'} &= \epsilon x_i + y_j. & (23) \end{aligned}$$

which yields (see the Appendix):

$$\begin{aligned} a_{ji}^{'} &= a_{ji} + \left(1 + a_{ji}^2\right)\epsilon, & (24) \\ a_{kl}^{'} &= a_{kl} + \left(a_{jl} a_{ki}\right)\epsilon, & (25) \\ t_k^{'} &= t_k + \left(a_{ki} t_j\right)\epsilon & (26) \end{aligned}$$

and therefore:

$$\sum_{k=1}^{M} \sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} \left(F \cdot \left(a_{jl} a_{ki}\right)\right) + \frac{\partial}{\partial a_{ji}} F + F \cdot a_{ji} = 0. \tag{27}$$

4. The PDE System

The translation invariance of Equation (5) excludes a dependence of F on $t_1, \cdots, t_M$, so F is of the form $F\left(a_{11}, \cdots, a_{ML} | I\right)$. Rotation invariance with respect to the y-axis requires F to fulfill the homogeneous, linear system of first order partial differential equations $\left(i, j \in [1, M], i \neq j\right)$ (*i.e.*, Equation (21)):

$$\sum_{k=1}^{L} \frac{\partial}{\partial a_{jk}} \left(F \cdot a_{ik}\right) - \sum_{k=1}^{L} \frac{\partial}{\partial a_{ik}} \left(F \cdot a_{jk}\right) = 0 \tag{28}$$

and similar for rotations around the x-axis $(i, j \in [1, L], i \neq j)$ (Equation (13)):

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{kj}} (F \cdot a_{ki}) - \sum_{k=1}^{M} \frac{\partial}{\partial a_{ki}} (F \cdot a_{kj}) = 0. \tag{29}$$

Rotations around an axis perpendicular to a plane given by an x,y-pair require the probability distribution to obey also $(i \in [1, L], j \in [1, M])$:

$$\sum_{k=1}^{M} \sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} (F \cdot (a_{jl} a_{ki})) + \frac{\partial}{\partial a_{ji}} F + F \cdot a_{ji} = 0. \tag{30}$$

Using the product rule, the double sum can be rewritten as

$$\sum_{k=1}^{M} \sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} (F \cdot (a_{jl} a_{ki})) = \sum_{k=1}^{M} \sum_{l=1}^{L} a_{jl} a_{ki} \frac{\partial}{\partial a_{kl}} F + F \cdot \sum_{k=1}^{M} \sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} (a_{jl} a_{ki}) \tag{31}$$

and the last term of the previous equation can be split into three parts and simplified:

$$\begin{aligned} & F \cdot \sum_{k=1}^{M} \sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} (a_{jl} a_{ki}) = \\ & \quad F \cdot \sum_{k=1, k \neq j}^{M} \frac{\partial}{\partial a_{ki}} (a_{ji} a_{ki}) + F \cdot \sum_{l=1, l \neq i}^{L} \frac{\partial}{\partial a_{jl}} (a_{jl} a_{ji}) + F \cdot \frac{\partial}{\partial a_{ji}} a_{ji}^2 = \\ & \quad (M-1) a_{ji} F + (L-1) a_{ji} F + 2 a_{ji} F = (M+L) a_{ji} F. \end{aligned} \tag{32}$$

Using this, Equation (30) can be written as:

$$\sum_{k=1}^{M} \sum_{l=1}^{L} a_{jl} a_{ki} \frac{\partial}{\partial a_{kl}} F + \frac{\partial}{\partial a_{ji}} F + (M+L+1) a_{ji} F = 0. \tag{33}$$

5. Solution

This system of PDEs (Equations (28), (29) and (33)) can be tackled with the theory of Lie groups, which provides a systematic, though algebraically-intensive solution strategy, which is implemented in contemporary computer algebra systems. The solutions of several test cases computed by the Maple computer algebra system (http://www.maplesoft.com/) (it proved to be superior to MATHEMATICA (www.http://www.wolfram.com/mathematica/) for the present PDE-systems) led to the conjecture that a general solution to this PDE system is given by the sum of the squares of all possible minors of the coefficient matrix:

$$\begin{aligned} & F(a_{11}, \cdots, a_{ML} | I) \\ = & \left[1 + \sum_{k=1}^{\binom{M}{P}\binom{L}{P}} \left(\det\left(A^{P,k}\right)\right)^2 + \sum_{k=1}^{\binom{M}{P-1}\binom{L}{P-1}} \left(\det\left(A^{P-1,k}\right)\right)^2 + \cdots + \sum_{k=1}^{\binom{M}{1}\binom{L}{1}} \left(\det\left(A^{1,k}\right)\right)^2 \right]^{-\frac{M+L+1}{2}} \end{aligned} \tag{34}$$

where A^n denotes a submatrix (minor) of size $n \times n$ (this notation is used at various places throughout the paper and should not be confused with the power of a matrix, which does not occur in this paper) and

$P = \text{Min}\,(M, L)$. Equation (34) does not appear unreasonable from the onset as prior density, because it preserves the underlying symmetry of the problem (permutation invariance of the parameters) and it is non-negative.

An explicit example for the case $N = 4, L = 2$ is:

$$F\,(a_{11}, a_{12}, a_{21}, a_{22}|I) = \left[1 + a_{11}^2 + a_{12}^2 + a_{21}^2 + a_{22}^2 + (a_{11} \cdot a_{22} - a_{12} \cdot a_{21})^2\right]^{-5/2}. \tag{35}$$

A two-dimensional slice of this probability density is given in Figure 2. The high symmetry of the prior distribution with respect to parameter permutations results in similar, "Cauchy"-like shapes if slices along other parameter axis are displayed.

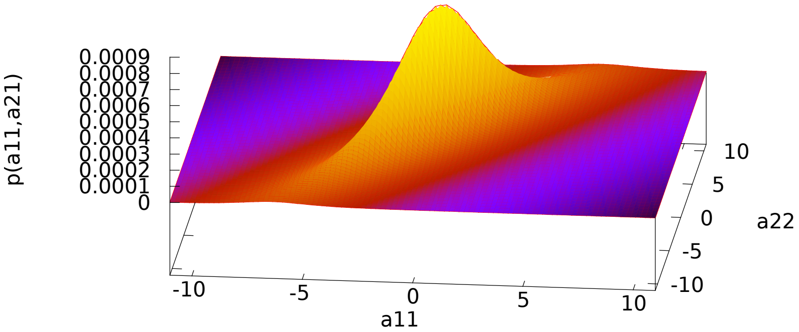

Figure 2. Probability density of $p\,(a_{11}, a_{21} \mid a_{12}, a_{22}, I)$ for $a_{12} = 3$ and $a_{22} = 5$ for the case $N = 4, L = 2$. The probability density exhibits the typical "Cauchy-"like shape with heavy tails compared to a binormal distribution. Due to the symmetry of the prior distribution, slices with respect to the other parameters display the same basic features.

For the case $N = 6, L = 3$, the solution is given by:

$$\begin{aligned}
& F\,(a_{11}, \cdots, a_{33}|I) \\
= \; & (1 + a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{21}^2 + a_{22}^2 + a_{23}^2 + a_{31}^2 + a_{32}^2 + a_{33}^2 + \\
& (a_{22}a_{33} - a_{23}a_{32})^2 + (a_{21}a_{33} - a_{23}a_{31})^2 + (a_{21}a_{32} - a_{22}a_{31})^2 + \\
& (a_{12}a_{33} - a_{13}a_{32})^2 + (a_{11}a_{33} - a_{13}a_{31})^2 + (a_{11}a_{32} - a_{12}a_{31})^2 + \\
& (a_{12}a_{23} - a_{13}a_{22})^2 + (a_{11}a_{23} - a_{13}a_{21})^2 + (a_{11}a_{22} - a_{12}a_{21})^2 + \\
& (a_{11} \cdot (a_{22}a_{33} - a_{23}a_{32}) - a_{12} \cdot (a_{21}a_{33} - a_{23}a_{31}) + a_{13} \cdot (a_{21}a_{32} - a_{22}a_{31}))^2)^{-7/2}.
\end{aligned}$$

6. Proof

6.1. Preliminaries

To prove that Equation (34) fulfills the equation system given by Equations (28), (29) and (33), we verify directly that Equation (34) solves the PDEs.

We will make repeated use of the Laplace expansion of determinants:

$$\det\,(A^n) = \sum_{j=1}^{n} a_{ij}\,(-1)^{i+j} \det\left(M_{ij}^{n-1}\right) \tag{36}$$

where the minor M_{ij}^{n-1} is the $(n-1) \times (n-1)$-matrix derived from the $n \times n$-matrix A^n by deletion of the i-th row and j-th column (by definition $M^0 := 1$). The cofactor matrix A_{ij}^{n-1} is defined to be:

$$A_{ij}^{n-1} = (-1)^{i+j} M_{ij}^{n-1} \tag{37}$$

and satisfies the following n-equations $(i, j, k = 1, 2, \cdots, N)$:

$$\sum_{j=1}^{n} a_{ij} \det\left(A_{kj}^{n-1}\right) = \delta_{ik} \det\left(A^n\right), \quad \sum_{i=1}^{n} a_{ij} \det\left(A_{ik}^{n-1}\right) = \delta_{jk} \det\left(A^n\right). \tag{38}$$

Further useful is the following form of the Laplace expansion, taking into account index shifts of a previous deletion of row k and column i of an $(n+1)$-matrix A^{n+1}, resulting in the minor M_{ki}^n:

$$\det\left(M_{ki}^n\right) = \sum_{l=1, l \neq i}^{n+1} a_{jl} \left(-1\right)^{(l'+j')} \det\left(M_{(jk)(li)}^{n-1}\right) \tag{39}$$

where $M_{(jk)(li)}^{n-1}$ is the minor given by deletion of the j-th and k-th row and the l-th and i-th column. l' and j' are defined as:

$$\begin{aligned} l' = l \quad \forall \quad (l < i) \quad \text{and} \quad l' = l - 1 \quad \forall \quad (l > i) \\ j' = j \quad \forall \quad (j < k) \quad \text{and} \quad j' = j - 1 \quad \forall \quad (j > k). \end{aligned} \tag{40}$$

In the following, we face the problem of possibly too heavy of a nomenclature, because we need summation indices, while we also need to keep track of the original indices underlying the entries in the minors, where some rows and columns have been deleted, although the relative order is preserved. The mapping could be expressed, e.g., as $a_{i(i')j(j')}$ with $i', j' \in [1, m]$ and $i\,(:) \in [1, M]$ and $j\,(:) \in [1, L]$. To avoid this cumbersome notation, we implicitly assume from now on (up to the Conclusion Section) this mapping for all summations that are indexed by either k or l. Therefore:

$$\sum_{k=1}^{m} a_{ki} \det\left(A_{kj}^{m-1}\right) \quad \text{has to be read as} \quad \sum_{k'=1}^{m} a_{k(k')i} \det\left(A_{k(k')j}^{m-1}\right). \tag{41}$$

6.2. $x_i x_j$- and $y_i y_j$-Rotations

We now verify that Equation (34) solves Equation (29). It is obvious that only those determinants of Equation (34) that contain column i or column j have the potential to provide non-zero contributions in Equation (29): if column j is missing, the derivative in the first term is zero. If, instead, column i is missing, then the derivative in the second term of Equation (29) yields zero. To proceed, we introduce $H(\mathbf{A})$ via:

$$F(\mathbf{A}) = H(\mathbf{A})^{-\frac{L+M+1}{2}}. \tag{42}$$

It is noteworthy that $H(\mathbf{A})$ has a very simple form: it is given by a sum of positive terms. This almost decouples the problem, and we can largely proceed on a term-by-term basis. Using the equality:

$$\frac{\partial}{\partial a_{pq}} \left(\det\left(A^m\right)\right)^2 = 2 \det\left(A^m\right) \det\left(A_{pq}^{m-1}\right) \tag{43}$$

the left-hand side of Equation (29) transforms to $(i, j \in [1, L], i \neq j)$:

$$-(M+L+1)\, H(a)^{-\frac{L+M+3}{2}} \det(A^m) \cdot \left(\sum_{k=1}^{m} a_{ki} \det\left(A_{kj}^{m-1}\right) - \sum_{k=1}^{m} a_{kj} \det\left(A_{ki}^{m-1}\right) \right) \tag{44}$$

and using Equation (38), we obtain:

$$-(M+L+1)\, H(a)^{-\frac{L+M+3}{2}} \det(A^m) \cdot \left(\delta_{ij} \det(A^m) - \delta_{ij} \det(A^m)\right) = 0 \tag{45}$$

and, therefore, Equation (34) solves Equation (29). The calculation is similar for Equation (28) and yields the result that Equation (34) solves also the system Equation (28).

6.3. $(x_i y_j)$-Rotations

The verification of the successful solution of Equation (33) by Equation (34) requires some more steps. As before, Equation (33) can be written as:

$$-\frac{M+L+1}{2} \left(\sum_{k=1}^{M} \sum_{l=1}^{L} a_{jl} a_{ki} H(\mathbf{A})^{-\frac{L+M+3}{2}} \frac{\partial H(\mathbf{A})}{\partial a_{kl}} + H(\mathbf{A})^{-\frac{L+M+3}{2}} \frac{\partial H(\mathbf{A})}{\partial a_{ji}} - 2 a_{ji} H(\mathbf{A})^{-\frac{L+M+1}{2}} \right) = 0 \tag{46}$$

and after multiplication with $H(\mathbf{A})^{\frac{L+M+3}{2}}$ as:

$$\begin{aligned} & -(M+L+1) \cdot \\ & \left(\sum_{m=1}^{P} \sum_{r=1}^{\binom{M}{m}\binom{L}{m}} \left(\sum_{k=1}^{m} \sum_{l=1}^{m} a_{jl} a_{ki} \det(A^{m,r}) \det\left(A_{kl}^{m-1,r}\right) + \det(A^{m,r}) \det\left(A_{ji}^{m-1,r}\right) \right) - a_{ji} H(\mathbf{A}) \right) \\ = \; & 0 \end{aligned} \tag{47}$$

6.3.1. Matrices with Either Row j or Column i

The inner double sum can be simplified for all matrices containing either row j or column i (*i.e.*, all matrices of size $P \times P$ and all matrices $A^{m,r}$ of size $m \times m, m \in (1, 2, \cdots, P-1)$ with label $r = 1, 2, \cdots, \binom{M}{m}\binom{L}{m} - \binom{M-1}{m}\binom{L-1}{m}$) using the Laplace expansion (here, the expansion with respect to row j is shown):

$$\begin{aligned} & \sum_{k=1}^{m} \sum_{l=1}^{m} a_{jl} a_{ki} \det(A^{m,r}) \det\left(A_{kl}^{m-1,r}\right) \\ = \; & \det(A^{m,r}) \sum_{k=1}^{m} a_{ki} \sum_{l=1}^{m} a_{jl} \det\left(A_{kl}^{m-1,r}\right) \\ = \; & \det(A^{m,r}) \sum_{k=1}^{m} a_{ki} \delta_{jk} \det(A^{m,r}) = a_{ji} \left(\det(A^{m,r})\right)^2 \end{aligned} \tag{48}$$

which cancels the corresponding determinant of $H(\mathbf{A})$ in the last term of Equation (48).

6.3.2. Matrices with Neither Row j nor Column i

The basic idea is to show that $\binom{M-1}{m}\binom{L-1}{m}$-matrices with neither row j nor column i, $m \in (1, 2, \cdots, P-1)$, cancel with the contributions of the corresponding matrices including row j and column i of size $(m+1) \times (m+1)$ of the second term.

Please note that there is a one-to-one correspondence of minors of size $m \times m$ without the j-th row and i-th column and the matrices of size $(m+1) \times (m+1)$ with row j and column i in the second term, therefore allowing one to label both with the same index r. After division by $-(M+L+1)$, the remaining terms of Equation (48) are (taking into account that the labeling of the rows and columns of the matrices of size $(m+1) \times (m+1)$ and $(m) \times (m)$ must be consistent):

$$\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{ji}^{m,r}\right)\det\left(A_{(jk)(il)}^{m-1,r}\right)+\det\left(A^{m+1,r}\right)\det\left(A_{ji}^{m,r}\right)-a_{ji}H_{ji}(\mathbf{A})=0 \tag{49}$$

with H_{ji} now only containing determinants with neither row j nor column i. If we now only consider the relevant term of H_{ji}, we can write:

$$\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{ji}^{m,r}\right)\det\left(A_{(jk)(il)}^{m-1,r}\right)+\det\left(A^{m+1,r}\right)\det\left(A_{ji}^{m,r}\right)-a_{ji}\det\left(A_{ji}^{m,r}\right)^{2}=0. \tag{50}$$

The equation is trivially true if $\det\left(A_{ji}^{m,r}\right)=0$; otherwise, we can divide by $\det\left(A_{ji}^{m,r}\right)$ and obtain:

$$\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{(jk)(il)}^{m-1,r}\right)+\det\left(A^{m+1,r}\right)-a_{ji}\det\left(A_{ji}^{m,r}\right)=0. \tag{51}$$

Replacing the various cofactors by the corresponding minors (*cf.* Equations (36) and (37)) yields:

$$\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\left(-1\right)^{(i+j+k'+l')}\det\left(M_{(jk)(il)}^{m-1,r}\right)+\det\left(A^{m+1,r}\right)-a_{ji}\left(-1\right)^{(i+j)}\det\left(M_{ji}^{m,r}\right)=0 \tag{52}$$

and after replacing $\det\left(A^{m+1,r}\right)$ by its Laplace expansion together with multiplication by $\left(-1\right)^{(i+j)}$, the equation reads:

$$\begin{aligned}\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\left(-1\right)^{(k'+l')}\det\left(M_{(jk)(il)}^{m-1,r}\right)+\\ \left(-1\right)^{(i+j)}\sum_{k=1} a_{ki}\left(-1\right)^{(i+k)}\det\left(M_{ki}^{m,r}\right)-a_{ji}\det\left(M_{ji}^{m,r}\right)=0\end{aligned} \tag{53}$$

and can be simplified to:

$$\sum_{k=1,k\neq j}^{m+1}\sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\left(-1\right)^{(k'+l')}\det\left(M_{(jk)(il)}^{m-1,r}\right)+\sum_{k=1} a_{ki}\left(-1\right)^{(j+k)}\det\left(M_{ki}^{m,r}\right)-a_{ji}\det\left(M_{ji}^{m,r}\right)=0 \tag{54}$$

because $\left(-1\right)^{2i}$ equals one in the second term. Therefore, the third term cancels with the second term for $k=j$, and the remaining equation is given by:

$$\sum_{k=1,k\neq j}^{m+1} a_{ki}\left(-1\right)^{k'}\sum_{l=1,l\neq i}^{m+1} a_{jl}\left(-1\right)^{l'}\det\left(M_{(jk)(il)}^{m-1,r}\right)+\sum_{k=1,k\neq j} a_{ki}\left(-1\right)^{(j+k)}\det\left(M_{ki}^{m,r}\right)=0. \tag{55}$$

Using Equation (39) together with the definition Equation (40), the inner sum of the first term of Equation (55) can be rewritten as:

$$\sum_{l=1,l\neq i}^{m+1} a_{jl}\,(-1)^{l'}\det\left(M_{(jk)(il)}^{m-1,r}\right) = (-1)^{j}\,M_{ki}^{m,r} \quad \forall\ (j<k) \tag{56}$$

and:

$$\sum_{l=1,l\neq i}^{m+1} a_{jl}\,(-1)^{l'}\det\left(M_{(jk)(il)}^{m-1,r}\right) = (-1)^{j-1}\,M_{ki}^{m,r} \quad \forall\ (j>k)\,. \tag{57}$$

Splitting the summation over k into two parts $(k<j)$ and $(k>j)$ and inserting the definition for k', we obtain:

$$\begin{aligned} &\sum_{k=1,k<j} a_{ki}\,(-1)^{(j+k)}\det\left(M_{ki}^{m,r}\right) + \sum_{k>j} a_{ki}\,(-1)^{(j+k)}\det\left(M_{ki}^{m,r}\right) \ + \\ \sum_{k=1,k<j} a_{ki}\,(-1)^{((j-1)+k)}&\det\left(M_{ki}^{m,r}\right) + \sum_{k>j} a_{ki}\,(-1)^{(j+(k-1))}\det\left(M_{ki}^{m,r}\right) \ = \ 0 \end{aligned} \tag{58}$$

where the first two terms cancel the last two terms.

Summarizing the previous approach, we have shown that for an arbitrary $n\times n$-determinant, the first and third term of Equation (48) almost cancel. Only determinants not containing the j-th row and the i-th column remain. These remaining contributions are canceled by the $(n+1)$-order determinant (required to contain the matrix element a_{ji}) of the second term in Equation (48). This schema can be repeated down to $n=1$, and the last step $(n=0)$ is easily explicitly calculated. This finishes our derivation.

7. Relation to Previously-Derived Special Cases

The underlying equation systems of the special case of an $(n-1)$-dimensional hyperplane in an n-dim space used in [8] and in this paper differ slightly due to a different parameterization, and therefore, the derived priors appear on first glance to be different, although they are identical, as will be shown below.

For probability density functions in different coordinate systems, the following equation holds:

$$p\left(\vec{a}\right)d\vec{a} = p\left(\vec{b}\left(\vec{a}\right)\right)\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| d\vec{a}, \tag{59}$$

where $|\cdots|$ denotes the absolute value of the Jacobi determinant:

$$\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| = \left|\det\begin{pmatrix} \frac{\partial b_1}{\partial a_1} & \frac{\partial b_1}{\partial a_2} & \cdots & \frac{\partial b_1}{\partial a_n} \\ \vdots & & & \vdots \\ \frac{\partial b_n}{\partial a_1} & \frac{\partial b_n}{\partial a_2} & \cdots & \frac{\partial b_n}{\partial a_n} \end{pmatrix}\right|. \tag{60}$$

The equation describing the (n-1)-dim hyperplane in an n-dim space in this paper is given by:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1(n-1)}x_{n-1} + t_1 \tag{61}$$

and results in the following prior:

$$p\left(a_{11}, a_{12}, \cdots, a_{1(n-1)}, t_1\right) = \left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right)^{-\frac{n+1}{2}}. \tag{62}$$

In [8], the corresponding hyperplane equation reads:

$$0 = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + 1 \tag{63}$$

with prior distribution:

$$p\left(b_1, b_2, \cdots, b_n\right) = \left(\sum_{i=1}^{n} b_i^2\right)^{-\frac{n+1}{2}}, \text{with} \quad \sum_{i=1}^{n} b_i^2 > R_0^2. \tag{64}$$

The latter constraints yield a proper (normalizable) prior. The relation of the two different parameterizations is given by:

$$b_i = \frac{a_{1i}}{t_1} \quad \forall i \neq n \qquad \text{and} \qquad b_n = -\frac{1}{t_1} \tag{65}$$

which yields the Jacobian:

$$\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| = \left|\det\begin{pmatrix} \frac{1}{t_1} & 0 & 0 & \cdots & 0 & -\frac{a_{11}}{t_1} \\ 0 & \frac{1}{t_1} & 0 & \cdots & 0 & -\frac{a_{12}}{t_1} \\ \vdots & & \ddots & \cdots & & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{t_1} & -\frac{a_{1(n-1)}}{t_1} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{1}{t_1^2} \end{pmatrix}\right| = \frac{1}{t_1^{n+1}} \tag{66}$$

Using this result and Equation (65), we can write:

$$p\left(\vec{b}\left(\vec{a}\right)\right)\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| d\vec{a} = \frac{1}{\left(\sum_{i=1}^{n-1}\left(\frac{a_{1i}}{t_1}\right)^2 + \frac{1}{t_1^2}\right)^{\frac{n+1}{2}} t_1^{n+1}} d\vec{a} = \frac{1}{\left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right)^{\frac{n+1}{2}}} d\vec{a} \tag{67}$$

which shows the equivalence of the two priors (Equations (62) and (64)). The requirement of $\sum b_i^2 > R_0^2$ leads to:

$$R_0^2 \leq \sum_{i=1}^{n} b_i^2 = \sum_{i=1}^{n-1}\left(\frac{a_{1i}}{t_1}\right)^2 + \frac{1}{t_1^2} = \frac{1}{t_1^2}\left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right). \tag{68}$$

In the case of all $a_{1i} = 0$, we obtain:

$$t_1^2 \leq \frac{1}{R_0^2} \tag{69}$$

which means that the lower limit R_0^2 corresponds to an upper limit of t_1^2.

8. Practical Hints

In the worst case, the hyperplane prior has an exponentially-increasing number of determinants with increasing dimension. The total number of individual determinants for an N-dimensional plane in a $2N$-dimensional space is given by:

$$\sum_{k=0}^{N} \binom{N}{k}^2 = \binom{2N}{N} \tag{70}$$

which is already 70 for a 4D hyperplane in an 8D space. Therefore, it is advantageous to compute the determinants using iteratively the Laplace expansion, starting from small determinants, storing the determinants of the previous step. This requires the storage of at most $\binom{N}{N/2}^2$ terms. As a proposal density for Markov chain Monte Carlo (MCMC) sampling methods (e.g., rejection sampling), the dominating multivariate Cauchy distribution is a good candidate. Source code for the set up of the PDE system and for the solution, together with a Maple script for the verification of the solution, can be obtained from the author.

9. Conclusions

This paper has derived a prior density for L-dimensional hyperplanes in N-dimensional space, based on geometric invariances. It is suited, e.g., to parameter estimation of multilinear regression problems in the absence of further prior knowledge or Bayesian model estimation for neural networks. In the latter case, the prior has to be made proper by suitable restriction of the range of the offset parameters, which depends on domain knowledge. The obtained prior density avoids the too strong weight of "large" values of the regression coefficients typically assigned by uniform priors. Being a rational function, its influence on the parameter estimates on standard problems with Gaussian uncertainties (resulting in an exponential likelihood) on the data will be limited. However, this can be different for robust estimation approaches with heavy-tailed likelihood distributions.

Appendix

In this section, the relation between the primed coefficient $a^{'}_{nm}$ and the unprimed coefficient a_{nm} is derived. A rotation perpendicular to the x_iy_j-plane relates x_i, y_j with $x^{'}_i, y^{'}_j$ by:

$$\begin{aligned} x^{'}_i &= x_i - \epsilon y_j, && \text{(A1)} \\ y^{'}_j &= \epsilon x_i + y_j. && \text{(A2)} \end{aligned}$$

and $x^{'}_k = x_k, k = 1, \cdots, L; k \neq i$ and $y^{'}_k = y_k, k = 1, \cdots, M; k \neq j$. Using this, the system Equation (5) in the transformed coordinate system reads $(n = 1, \cdots, M; n \neq j)$:

$$\begin{aligned} y_n &= a^{'}_{n1}x_1 + a^{'}_{n2}x_2 + \cdots + a^{'}_{ni}\left(x_i - \epsilon y_j\right) + a^{'}_{n(i+1)}x_{(i+1)} + \cdots + a^{'}_{nL}x_L + t^{'}_n \\ y_j + \epsilon x_i &= a^{'}_{j1}x_1 + a^{'}_{j2}x_2 + \cdots + a^{'}_{ji}\left(x_i - \epsilon y_j\right) + a^{'}_{j(i+1)}x_{(i+1)} + \cdots + a^{'}_{jL}x_L + t^{'}_j. \end{aligned} \quad \text{(A3)}$$

Solving for y_j, we obtain:

$$y_j = \frac{1}{1 + a^{'}_{ji}\epsilon}\left(t^{'}_j - x_i\epsilon + \sum_{k=1}^{L} a^{'}_{jk}x_k\right) \quad \text{(A4)}$$

and subsequently:

$$y_n = \left(t^{'}_n + \sum_{k=1}^{L} a^{'}_{nk}x_k\right) - a^{'}_{ni}\epsilon\frac{1}{1 + a^{'}_{ji}\epsilon}\left(t^{'}_j - x_i\epsilon + \sum_{k=1}^{L} a^{'}_{jk}x_k\right). \quad \text{(A5)}$$

Using the Taylor expansion $1/\left(1+a'_{ji}\epsilon\right) = 1 - a'_{ji}\epsilon + O\left(\epsilon^2\right)$ up to first order and collecting the coefficients, the previous equations yield:

$$\begin{aligned} a_{ji} &= a'_{ji} - {a'_{ji}}^{2}\epsilon - \epsilon \quad ; \quad t_j = t'_j - t'_j a'_{ji}\epsilon \\ a_{nk} &= a'_{nk} - a'_{ni}a'_{jk}\epsilon \quad ; \quad t_n = t'_n - t'_j a'_{ni}\epsilon. \end{aligned} \tag{A6}$$

First, we solve for a'_{ji}:

$${a'_{ji}}^{2}\epsilon - a'_{ji} + \epsilon + a_{ji} = 0 \rightarrow a'_{ji} = \frac{1-\sqrt{1-4\epsilon\left(\epsilon + a_{ji}\right)}}{2\epsilon} = \underline{a_{ji} + \left(1 + {a_{ji}}^{2}\right)\epsilon} + O\left(\epsilon^2\right) \tag{A7}$$

and next for a'_{jk}:

$$\begin{aligned} a_{jk} = a'_{jk}&\left(1 - a'_{ji}\epsilon\right) \rightarrow \\ a'_{jk} &= \frac{a_{jk}}{1 - a'_{ji}\epsilon} = a_{jk}\left(1 + a'_{ji}\epsilon\right) + O\left(\epsilon^2\right) \\ &= \underline{a_{jk}\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right). \end{aligned} \tag{A8}$$

A similar calculation for a'_{ni} yields:

$$a'_{ni} = \underline{a_{ni}\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right) \tag{A9}$$

which then allows one to compute a'_{nk} for index pairs with $\{nk\} \neq \{ji\}$:

$$a'_{nk} = a_{nk} + \left(a_{ni} + a_{ni}a_{ji}\epsilon\right)\left(a_{jk} + a_{jk}a_{ji}\epsilon\right)\epsilon = \underline{a_{nk} + a_{ni}a_{jk}\epsilon} + O\left(\epsilon^2\right). \tag{A10}$$

The offset variable t_j is given by:

$$\begin{aligned} t_j = t'_j&\left(1 - a'_{ji}\epsilon\right) \rightarrow \\ t'_j &= \frac{t_j}{1 - a'_{ji}\epsilon} = t_j\left(1 + a'_{ji}\epsilon\right) + O\left(\epsilon^2\right) \\ &= \underline{t_j\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right). \end{aligned} \tag{A11}$$

and the other offset variables t_n by:

$$t'_n = t_n + \left(a_{ni} + a_{ni}a_{ji}\epsilon\right)\left(t_j + t_j a_{ji}\epsilon\right)\epsilon = \underline{t_n + a_{ni}t_j\epsilon} + O\left(\epsilon^2\right) \tag{A12}$$

which concludes the derivation of Equations (24)–(26).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Gosling, J.P.; Oakley, J.E.; O'Hagan, A. Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Anal.* **2007**, *2*, 693–718.
2. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *SSC4*, 227–241.

3. Kendall, M.; Moran, P. *Geometrical Probability*; Griffin: London, UK, 1963.
4. Von der Linden, W.; Dose, V.; von Toussaint, U. *Bayesian Probability Theory: Application to the Physical Sciences*, 1st ed.; Cambridge University Press: Cambridge, UK, 2014.
5. Von Toussaint, U.; Gori, S.; Dose, V. Bayesian Neural-Networks-Based Evaluation of Binary Speckle Data. *Appl. Opt.* **2004**, *43*, 5356–5363.
6. Hinton, G.; Salakhutdinov, R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507.
7. Minh, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; *et al.* Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
8. Dose, V. Hyperplane Priors. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Williams, C.J., Ed.; American Institute of Physics: Melville, NY, USA, 2003; Volume AIP Conference Proceedings 659, pp. 350–357.
9. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Wiley: New York, NY, USA, 1992; Reprint from 1973.
10. Zellner, A. *An Introduction to Bayesian Inference in Econometrics*; Wiley: New York, NY, USA, 1971.
11. West, M. Outlier Models and Prior Distributions in Bayesian Linear Regression. *J. R. Stat. Soc. B* **1984**, *46*, 431–439.
12. O'Hagan, A. *Kendall's Advanced Theory of Statistics, Bayesian Inference*, 1st ed.; Arnold Publishers: New York, NY, USA, 1994; Volume 2B.
13. Landau, L.; Lifschitz, E. *Lehrbuch der Theoretischen Physik I*, 1st ed.; Akademie Verlag: Berlin, Germany, 1962.