

Analysis of Variance for Directions and Axes

K.G. van den Boogaart

Mathematics and Computer Sciences in Geology, Freiberg University of Mining and Technology, Germany

Gerald v.d. Boogaart, Mathematics and Computer Sciences in Geology, Freiberg University of Mining and Technology, BvCotta-Str. 2, D-09596 Freiberg, Germany, boogaart@grad.tu-freiberg.de

1. Abstract

Many data in structural geology consist of directions and axes. Different directional measurements are associated with different objects (foliation, lineation, strain-markers) and different events and times. Structural direction interrelate in a complex way. They depend on various external parameters such as paleostress and material properties. This contribution provides some ideas how to integrate these complex data into meaningful stochastic models, which allow to analyze the dependence of directions statistically.

The central idea is the generalization of multivariate analysis of variance and multiple regression to directions and symmetric directions (e.g. axes). The generalization is based on a general distribution family for directions introduced by Beran and generalized linear models. The directions are rather described by their distribution than by their expected value as the analysis of variance for real variable does. The model works for 2D-directions, 2D-axes 3D-directions, 3D-axes and 3D-orientations. Possible independent variables are categorical and real values, physical tensors and further geometric objects such as directions and axis.

These models provide one possibility of a multivariate joint analysis of geometrical, categorical and real measurements. They can be used in a way very similar to multiple regression models. A measure of goodness of fit analogous to R^2 is provided by an entropy measure telling how precisely the directions are determined by the independent variables. ANOVA tests and tables are replaced by likelihood ratio tests and corresponding entropy measures.

2. Analysis of variance and linear models

In the beginning we should clarify that this paper only develops a new method for data types common in geology. Whether the method is useful for real world geological applications must be proven by the practice in future. The method we develop is analysis of variance and regression methods for directions and axes.

Analysis of variance (also called ANOVA) and linear regression are nice and apparently everywhere used methods to explore the dependence of quantities of the real scale on other quantities of cardinal or real scale. The quantity of interest, whose dependence we explore, is called dependent variable and denoted by Y throughout the paper. The other variables, which are supposed to determine the distribution of Y are called independent variables and in this paper denoted as X_1, X_2, \dots, X_p if they are of real scale and by k, l, m if they are on a categorical scale. In classical situation we assume Y to dependent on the independent variables linearly with an additive random error ε with mean zero and constant variance σ^2 :

$$Y = a_1X_1 + a_2X_2 + \dots + a_pX_p + b_k + c_l + \dots + d_m + \varepsilon$$

Often a normal distribution is assumed for the stochastic term ε . The $a_i, b_i, c_i, d_i, i = 1, \dots$ are assumed to be unknown constant real parameters. The models are called linear models, because Y is written in a linear form of these unknown parameters. Many different statistical methods have been developed for this class of models. E.g. we can prove the dependence of Y on X_i statistically, by a test giving statistical evidence for $a_i \neq 0$. When we have estimated parameters a_i, b_i, c_i, d_i , we are able to give expected values and confidence regions for Y , for every known set of independent variables $X_1, X_2, \dots, X_p, k, l, m$. The measure R^2 provides us with a measure of the importance of the modeled influence. R^2 is defined to be the residual variance left to the randomness of ε , divided by the total variance of Y disregarding any dependence. It is thus a measure of which part of the variance could be explained by the deterministic dependence of Y on the independent variables. More complex models can be built, which model interaction of the influence owed by the independent variables:

$$Y = a_1X_1 + a_2X_2 + a_{12}X_1X_2 + \dots + b_k + c_l + d_{kl} + \dots + \varepsilon$$

This is an interaction model. It is possible to compare models with more and with less terms on the right hand side and to decide statistically, whether the additional parameters are really needed to explain the behavior of Y . In general we should use the most simple model, which is not contradicted by the data.

However this nice and powerful methods cannot directly and meaningfully be applied to the most popular types of data in structural geology, which are directions and axes. Depending on the representation of directions we use, there are two direct ways of application of ANOVA to them and we should clarify why they are not useful.

3. Directions and axes

Geological directions and axes are conventionally described by strike s and dip d . The dip is always between -90° and 90° . For directions the strike directions can vary between 0° and 360° , while for axes it can vary between 0° and 180° only, because for directions we have to decide whether they head to west or east, while for axes both headings are indistinguishable.

Thus directions and axes are described by two real numbers and we could think about simply applying multivariate analysis of variance to these two angles. However due to the discontinuity in the parameter space at strike zero we would get totally different results and artefacts, whenever the features strike north.

Directions can be represented without such discontinuities as vectors in the three dimensional space. However in this representation directions can neither be distributed normally nor have any expected value possibly resulting from linear combinations of the independent variables, because they reside on the unit sphere. The situation is even worse for axes, which are represented by a vector and its antiparallel vector in three dimensional space. Axes cannot be distributed normally and the expected value of their vector representation would always be zero and thus not useful for linear modeling.

4. Generalized linear models

The central idea to generalize linear models for nonlinear manifolds is to view the central modeling equation (1) not as an equation for the expected value, but for a parameter of the distribution of Y :

$$Y \sim N(a_1X_1 + a_2X_2 + \dots + a_pX_p + b_k + c_l + \dots + d_m, s^2)$$

This idea can be generalized for distributions other than the normal distribution. This models are then called generalized linear models in literature (MCCULLAGH, P., A. NELDER, 1989). E.g. When Y is a two stage variable we can have models like:

$$Y \sim \text{Binomial}(\exp(Z)/(1 + \exp(Z)), Z = a_1X_1 + a_2X_2 + \dots + a_pX_p + b_k + c_l + \dots + d_m$$

which is called a logit model in literature, since the log odds $\log(p/(1 - p))$ are linear in the parameters. Generalized linear models normally work well with exponential family distributions.

5. Spherical and axial distributions

In order to apply generalized linear models to directions and axes we need appropriate distribution families. The most popular exponential family distributions for directions is the Fisher distribution. In natural parameterisation it reads:

$$f_{\mathbf{v}}(\mathbf{y}) = c_{\mathbf{v}} \exp(\mathbf{v}^t \mathbf{y}),$$

with a parameter $\mathbf{v} \in \mathbb{R}^3$ and an appropriate normalization constant $c_{\mathbf{v}}$ such that f is a distribution density on the direction sphere. It is necessary to use the natural parametrisation $\mathbf{v} \in \mathbb{R}^3$ as exponential distribution family instead of the common parametrisation with a unit vector and a concentration parameter κ to apply generalized linear model theory later. The spherical Fisher distribution is an unimodal distribution of directions and has near to normal distribution shape as far as this is possible on the direction sphere. The most often used distributions for axes is given by the Bingham distribution, which has a symmetric matrix valued parameter \mathbf{F}

$$f_{\mathbf{F}}(y) = c_{\mathbf{F}} \exp(\mathbf{y}^t \mathbf{F} \mathbf{y})$$

again $c_{\mathbf{F}}$ is a normalization constant. This density is equal for \mathbf{y} and $-\mathbf{y}$ and thus a proper density for axes. The Bingham distribution can model bimodal distribution and great circle distributions of directions and a unimodal distribution and great circle distributions of axes. Both models yield a valid distribution density regardless of the parameters \mathbf{v} or \mathbf{F} . Details on the distributions can be found in (FISHER 1995).

More advanced spherical exponential families modeling more complex conditional distributions can be found in (Beran 1979). These families similar to the Fisher distribution, but use spherical harmonic functions \mathcal{Y}_l^n up to a fixed degree $l \leq L$ instead of the vector \mathbf{x} .

$$f_{\theta}(\mathbf{x}) = c_{\theta} \exp \left(\sum_{l=1}^L \sum_{n=0}^{2l+1} Y_l^n(\mathbf{x}) \theta_l^n \right)$$

They can be applied to axes by using only the harmonic functions with even degree l . The spherical Fisher distribution is the case $L = 1$ and the Bingham distribution is the case $L = 2$ for axes. The formulae in this paper will be written for the Fisher and Bingham distributions. But they stay valid for the general case just replacing the parameter \mathbf{v} with $\theta = (\theta_l^n)$ and the same ideas apply. Models with conditional distributions of orientations based on the Fisher-matrix distribution for orientations its generalizations are discussed in (Boogaart 2002).

6. Spherical regression models

We can now model the stochastic dependence of the directions on independent variables

modeling the parameter of their distribution as a linear function of the independent variables.

$$\mathbf{v} = \mathbf{a}_1 X_1 + \mathbf{a}_2 X_2 + \dots + \mathbf{a}_p X_p + \mathbf{b}_k + \mathbf{c}_l + \dots + \mathbf{d}_m$$

The unknown parameters \mathbf{a}_i , \mathbf{b}_i , \mathbf{c}_i , \mathbf{d}_i , $i = 1, \dots$ are now supposed to be vectors rather than numbers, since here we need to calculate all three vector components of the parameter \mathbf{v} rather than just one mean value. The direction of \mathbf{v} can be interpreted as mean direction and the length of \mathbf{v} determines the spread of the distribution around the central direction. Analogously for axes we can model a linear dependence of \mathbf{F} on the independent variables:

$$\mathbf{F} = \mathbf{A}_1 X_1 + \mathbf{A}_2 X_2 + \dots + \mathbf{A}_p X_p + \mathbf{B}_k + \mathbf{C}_l + \dots + \mathbf{D}_m$$

The unknown parameters \mathbf{A}_i , \mathbf{B}_i , \mathbf{C}_i , \mathbf{D}_i , $i = 1, \dots$ are now supposed to be symmetric matrices rather than numbers. The first eigenvector of \mathbf{F} specifies the modal direction and the other two determine the anisotropy of the bellshaped double mode. The eigenvalues determine the density in the three eigendirections of \mathbf{F} .

7. Spherical regressors and directions

Natural directions, vectors, axes, and tensors are good candidates for independent variables determining the distributions of the observed directions and axes. Examples for such independent variables are moving directions, relative displacement, main axes of stress field or the strain tensor.

Some of them can be used as regressors in a canonical way. First of all directions can be represented as real numbers, the entries x_i of the vector $\mathbf{x} = (x_i)_{i=1, \dots, 3}$ and thus be used as regressors. Thus we need an additional matrix valued parameter $\mathbf{M} \in \mathbb{R}^{3 \times 3}$:

$$\mathbf{v} = \dots + \mathbf{M}_k \mathbf{x}_k + \dots$$

However a useful assumption would be that the influence induced by \mathbf{x} is symmetric around \mathbf{x} and thus $\mathbf{M}\mathbf{x} \parallel \mathbf{x}$, which implies $\mathbf{M}_k = A_k \mathbf{I}$, $A_k \in \mathbb{R}$. Thus, the symmetric influence of a direction can be modeled with only one additional real parameter $A_k \in \mathbb{R}$ in the model:

$$\mathbf{v} = \dots + A_k \mathbf{x}_k + \dots$$

Here the regressor \mathbf{x} can be either the vectorial representation of a direction $\mathbf{x} \in S_2$ or a full vector $\mathbf{x} \in \mathbb{R}^3$. When we use the more complicated families of (Beran 1979) instead of Fisher distribution, we need no use the harmonic functions as regressors instead of \mathbf{x} .

8. Axial regressors and axes

With the same argument of symmetry it is not useful to have axial regressors for directions or directional regressors for axes. However a symmetric axial regressor for axes is made up by

$$\mathbf{F} = \dots + A_k \mathbf{x}\mathbf{x}^t + \dots, \quad A_k \in \mathbb{R}$$

The $\mathbf{x}\mathbf{x}^t$ is a symmetric tensor of rank 2. Other symmetric tensors ϵ_{ij} of rank 2 can be used as regressors using only one parameter in the same way.

$$\mathbf{F} = \dots + A_k (\epsilon_{ij})_{ij} + \dots, \quad A_k \in \mathbb{R}$$

When we use the more complicated families of (Beran 1979) instead of the Bingham distribution, we need no use the even harmonic functions as regressors instead of $\mathbf{x}\mathbf{x}^t$.

9. Measure of Randomness

A measure of randomness analogously to R^2 in ANOVA and linear regression is provided by the entropy (KULLBACK 1959).

$$I(f) := \oint f(\mathbf{y}) \ln f(\mathbf{y}) d\mathbf{y}$$

$d\mathbf{y}$ denotes the integration over the sphere, which can be written as $d\mathbf{y} := \frac{1}{4\pi} \cos(d) dd\theta ds$ and f the density of the distribution given by the model and its estimated parameters according to $d\mathbf{y}$. A measure varying between 0 and 1, where 1 describes perfect explanation and 0 no difference to the model of total randomness, is given by:

$$R^2 := 1 - I(f)^2$$

This measure can be used like the ordinary R^2 of linear regression. Indeed the definition is equal to R^2 , when using linear models.

10. Parameter estimation and testing

The model is a special case of an exponential distribution family, where the regression parameters are the natural parameters of the model. Thus parameter estimation and testing can be done according to the standard techniques for exponential families e.g. (WITTING 1995). The maximum likelihood estimator for the parameters can be calculated by the iteration scheme of Fisher's scoring method. Test problems testing one of these regression models as hypothesis and another model with additional parameters as alternative can be constructed as the standard likelihood ratio test.

11. Conclusions

The basic methods of real ANOVA given by modeling influences, estimating the parameters and choosing models based on model comparison tests and a measure of residual randomness given by R^2 can be applied to directions and axes by the models proposed here. Additionally to real and categorical regressors we can use directions and vectors for directions or axes and rank 2 tensors for axes respectively.

12. References

Beran, R. (1979): Exponential families for directional data, *The Ann. Statist.* 7, pp 1162-1178

Boogaart, K.G. v.d. (2002): Statistics for individual crystallographic orientation measurements, PhD-thesis, Shaker, Aachen

Fisher, N.I., T. Lewis, B.J.J. Embleton (1987): *Statistical Analysis of spherical data*, Cambridge University Press, Cambridge

Kullback, S. (1959): *Information theory and statistics*, Dover Publications, New York

McCullagh, P., A. Nelder, (1989): *Generalized Linear Models*, Second Edition, Chapman and Hall, New York

Witting, H., U. Müller-Funk (1995): *Mathematische Statistik II*, Teubner, Stuttgart