Geometry of Statistics

Khalil Iskarous
Haskins Laboratories
Good Reads
Overview

- Fisher and Pearson conceived of statistics using vector-geometric intuition.
- Rao, Durbin, and others continued that tradition.
- Subsequently, in the modern synthesis of Fisher and Pearson’s ideas, an algebraic approach replaced the geometric one.
- Advantages of Vector Geometric approach: more intuitive understanding and ease of transition into advanced statistical methods.
Overview

- This is an introduction to vector geometry of the General Linear Model (GLM), i.e. Correlation, Regression, ANOVA, ANCOVA, etc.

- This geometry is present in many applications of GLM thinking. A brief introduction of how this geometry underlies LPC (Linear Predictive Coding) and ANN/CLH (Covariant Learning Hypothesis) will be presented.
Two ways of viewing data

- Two people are measured on height and weight. Two observations and two variables.

Table 1. Hypothetical Data for Adam and Eve

<table>
<thead>
<tr>
<th></th>
<th>Height (cm)</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>182</td>
<td>85</td>
</tr>
<tr>
<td>Eve</td>
<td>164</td>
<td>52</td>
</tr>
</tbody>
</table>

Graphs showing the relationship between height and weight for Adam and Eve.
Patterns of Variability

![Graphs showing patterns of variability in weight and height for static and dynamic pictures.](image-url)
Vector Direction = Pattern of Variability

- What does pattern of variability relation mean?
- If for two variables A & B, the subjects who have low values on A have low values on B and the subjects who have high values on A have high values on B.
- A & B have same patterns of variability, and their vectors will point in the same direction.
Correlation & Regression

- Galton: We have two variables, parent height (indep. v) and child height (dep. var). We want to find how similar these variables are (Correlation), and how we can predict child height from parent height (Regression).

- Correlation is about similarity of variables and Regression is about prediction of one variable from another.
Vector Geometric Picture

- We are interested in the variables, but our plot is of the observations. How can we plot the **variables**?

- Instead of plotting a point for each observation, plot each variable, as a vector. If we have two observations: P1,C1 = (68, 69) & P2,C2 = (79, 83)

- Instead of plotting (68, 69), (79, 83), plot (68, 79) as a vector representing **parent height** and (69, 83) as a vector representing **child height**.

- If we add another observation P3,C3 = (94, 88), **parent height** is now the vector (68, 79, 94) in 3D, and **child height** is another vector (69, 83, 88) in 3D.
Problem

- For every new observation, we need a new dimension. That’s not good! We run high powered experiments with many many subjects-- let’s say 39. How can you plot a vector in 39 dimensional space?

- And why would you want to do such a thing, when the nice old scatter plot can accommodate hundreds of observation points?

- Because in correlation/regression (and rest of GLM), we are interested in the relation between variables, and their patterns of variability, not the relation between observations (cf. multidimensional scaling).

- Solution: based on intuition of 2D and 3D, imagine n-D.
Quantifying Correlation

- Take the two variables as vectors.
- Correlation: how **similar** are these two vectors?
- For vectors, vectors are **similar** if they align and dissimilar if orthogonal.
- Align: Angle $\theta$ between vecs $= 0$. Ortho: $\theta = 90$.
- Cosine of the angle **is** correlation coefficient $\rho$.

$$\rho = \cos(\theta) = 1, \quad \rho = \cos(\theta) = 0, \quad \rho = \cos(\theta) = -1$$
Geometry to Algebra

- We have just derived the formula for correlation in terms of trigonometry. Easy to implement in R or matlab, 1) multiply two vectors entry by entry, 2) sum, 3) normalize by magnitudes. Normalized inner product.

The angle, $\theta$, between two vectors $x$ and $y$ is given by

$$\cos \theta = \frac{x_1y_1 + \cdots + x_Ny_N}{\|x\| \|y\|}$$

$$= \frac{x.y}{\|x\| \|y\|}$$

($x.y = x_1y_1 + \cdots + x_Ny_N$ is termed the dot product of $x$ and $y$).

Vectors $x$ and $y$ are orthogonal if $\theta = 90^\circ$. This occurs iff $x.y = 0$. 

Friday, November 6, 2009
Vectors specified by Length & Angle

- Each of the vectors has a length and an angle.
- Length = Magnitude of Variance.
- Angle = Pattern of variability.
- E.g. Variables (2,4) and (100,200) have the same pattern of variability, but have different variances.
- E.g. Variables (1,0) and (0,1) have same variance, but totally different patterns of variability.
Correlation cares about angle/pattern of variability

- The correlation coefficient is the cosine of an angle.
- It does not care about the variance/length of the variance, only their patterns of variability.
Standardization

- Divide each variable by its length.
- Every vector now has the same length, i.e, has been standardized. All variables are now vectors whose end is on the unit circle (n-dimensional sphere), distinguished only by pattern of variability.
Interlude: Vector Addition

- Tail on head Addition
Regression

- We want to predict y from x. What does that mean?
- There is a part of variable y that is explained by variable x: \( \hat{y} \). And there is a part of variable y that is not predicted by x: e. We want the former.
- E.g.: we want to decompose children’s height into a component predictable from parent and error.
- Geometrically, \( y = \hat{y} + e \)
- How do we get \( \hat{y} \) from x & y?
  - Project y onto x orthogonally
b coeffs of Reg.

- b regression coeffs: $\hat{y} = b\hat{x}$.
- b is coefficient of projection that tells you how to scale x to get the part of it that is most consistent with y.
- In the variables-as-axes picture, we get a meaning of b as slope and intercept, which is very useful.
- An additional meaning is added when b is interpreted as the projection coefficient. If the vectors align, they have a large b (maximal projection) and when they are orthogonal, they have a small b (0-projection).
- Therefore b somehow carries meaning about the similarity of the variables as well and is dependent on the angle of projection.
b vs. \( \rho \)

- \( b \) cares about variance of variables, i.e. the lengths of their vectors, but factors in correlation as well.

\[ slope = \rho \frac{\sigma_y}{\sigma_x} \]
Interlude: Sum of Squares

- **Sum of Squares**: The length of a vector is obtained by decomposing it into its additive components and summing their squared lengths via Pythagorean Theorem.

10. *Pythagoras’s theorem in N-space* says that

\[ \|y\|^2 = (y.U_1)^2 + \cdots + (y.U_N)^2. \]

That is,

\[(\text{length})^2=\text{sum of squared lengths of projections}.\]
Quality of Regression

Length of explained part over length of whole variable.

But we could have also gotten a different measure of quality of regression by directly dividing the length of explained part by length of unexplained part $e$.

Statistics like $t$, $F$, chi-square are all similar in measuring a ratio of explained vs. unexplained, give or take squaring, whole variable vs. unexplained in denominator, and extra standardizations of length.
**Standardized Regression**

- Since \( x \) and \( y \) are now on the unit-circle, with equal lengths, the cosine of the angle between them (correlation coefficient \( \rho \)) is equal to the regression coefficient \( b \). Both tell us about similarity of pattern of variability in the variable.
In ANOVA, we have two types of variables/vectors:

- **Data Vector**: A vector of all the data in an experiment.
- **Hypothesis Vector**: A vector that represents a hypothesis we want to test.

Geometry of ANOVA is geometry of regression with the identification: Data $\leftrightarrow$ Dependent Variable & Hypothesis $\leftrightarrow$ Independent Variable
ANOVA

- ANOVA will allow us to measure how much the hypothesis predicts the data.
- It measures the quality of the fit between Data and Hypothesis (Model) by projecting data vector onto hypothesis vector and measuring the quality of the fit between them by taking the ratio of the part of the data explained by the hypothesis and the part of the data left unexplained by the hypothesis!!!
Figure 7. Orthogonal Decomposition of the Observation Vector (two populations, completely randomized design).
ANOVA Example

- We want to test the hypothesis that scores of students in Class A: 39.2 & 40.4 are significantly different from scores of students in Class B: 45.3 & 46.3.

- Data Vector: \([39.2, 40.4, 45.3, 46.3]\) or \([\text{Level A}, \text{Level B}]\)

- Null Hypothesis Vector (Uniformity): \([1, 1, 1, 1]\), i.e. all scores are same

- Hypothesis that CA different from CB (Nonuniformity): \([1 \ 1 \ -1 \ -1]\) (really the subspace spanned by \([1 \ 1 \ 0 \ 0]\) and \([0 \ 0 \ -1 \ -1]\)) or any of a number of ways for specifying the contrast between the levels
Doing ANOVA

- Project Data Vector onto Hypotheses Vectors, which should be orthogonal, since you cannot predict nonuniformity hypothesis from uniformity null hypothesis.

- Evaluate hypotheses by similarity of data and hypotheses: compare length of projection onto hypothesis with projection onto orthogonal subspace.

- That is \( F \)!!!
Meaning of degrees of freedom: Dimensions

- Let’s take an experiment on 100 individuals. We are in 100 dimensional space. Data is a vector.

- The null hypothesis $[1...1]$ is another vector.

- Error is a vector in higher dimensional space.

- If we project the data onto each and get the magnitude of the projections to obtain $F$, we are faced with the problem that Error has so many more components, so we divide by its number of components.

- Normalize: divide the magnitude of the projections by the number of dimensions.

$$F = \frac{|\hat{y}|^2/p}{|\hat{e}|^2/(n-p)} = \frac{MS_{model}}{MS_{error}}.$$
p-metric

- There is a cone of data vectors near $(1,1,1)$ that would make the null uniformity hypothesis true. The particular p-value we choose defines cone size.
Meaning of b coefficients

- Just as in regression, we get coefficients of projection b. In ANOVA, these are the contrast-dependent coefficient “estimates” you get from R runs of lmer or aov!

- They tell you how to stretch each hypothesis vector to get the part of each hypothesis that predicts the data.
Let’s say we have a situation where there are two sources of explanation of the collected data: Hypotheses/treatments and Blocking.

By blocking uniform subjects into blocks, we can reduce variability within each block due to uniformity.

Vector Picture: Error variability broken down into block variability and true error:
GLM

- Dependent/Data Variable is the sum of independent variables (regression) or hypothesis variable (ANOVA) or both (ANCOVA) and an error component. Examination of all such models bills dow to same process:
  - 1) Project Data onto Hypothesis space and determine error preventing data from being totally explainable by hypothesis.
  - 2) Evaluate model by comparing projection onto hypothesis vs. error.
  - 3) Compare result by p-metric.

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i,
\]
\[
y = X\beta + e.
\]
Each group of variables with similar pattern of variability leads to one principal component.
Applications

- GLM thinking, through the vector picture, arises everywhere.
Linear Prediction

- Each sample in a speech waveform is highly correlated with nearby samples, otherwise speech would be noise.
- Linear **Prediction** allows us to predict an individual sample from nearby samples.

![Prediction vectors](image-url)
Linear Prediction

- Get a window of 25 ms, let’s say at 22 kHz, i.e. 550 samples. We want to predict each sample from 14 previous samples, let’s say.
- Matrix: each row is 14 consecutive samples.

Fig. 1. Prediction vectors.
Example, $p = 3$

- Signal: $[0.9593 \ 0.5472 \ 0.1386 \ 0.1493 \ 0.2575 \ 0.8407 \ 0.2543 \ 0.8143 \ 0.2435 \ 0.9293]$
Linear Prediction

- The 14th vector is of all 14th samples. We want to predict all 14th samples from 1th samples, 2th samples, etc.

- In GLM terms, we are trying to predict 14th vector dependent variable from 13 independent vectors.

- b coefficients, obtained by projection onto 13-th dimensional space are LPC’s. They contain the structure in the signal. Useful in coding: send 14 #s, instead of 550.

- The 13 vectors we want to project onto are not orthogonal.

- If we first orthogonalize them, then project onto new space of orthogonal vectors, the new b’s are the reflection coefficients 😊.
Covariant Learning

- We are given a set of sound patterns (LPC parameters) as 14-vectors, let’s say, and a set of letters as 14-vectors. And a set of supervised associations between the two sets. E.g., this sound goes with this letter.

- We want a network to “learn” the associations and to be able to generalize to new patterns.
Covariance Learning

- Geometry: 14 vectors in each space.

- Learn similarity of the 14-vectors in each of the spaces to the others, (generalized) cosines of angles basically.

- This is the knowledge that the network has of the association.

- Associative Memory: The association is not between observations, but between vectors. The “weights” of the association tell us how each entry of each of the 14-vectors relates to the 14 entries of the other type of vector.
Summary

- There is much much more. This picture is fertile and interconnects many disciplines.
- Books by Saville & Wood and by Wickens are a must.
- Papers on geometry of linear prediction by Alexander, Stroebach, Kailath.