

Inference in longitudinal data analysis through the distance-based model using R

Sandra Melo¹ Carles Cuadras² Oscar Melo³

¹Departamento de Agronomía, Universidad Nacional de Colombia

²Departamento de Estadística, Universidad de Barcelona

³Departamento de Estadística, Universidad Nacional de Colombia

III Jornadas de Usuarios de R. Escuela de Organización Industrial. Madrid, España, 2011.

Outline

- 1 Background and solution
 - Background
 - Solution
 - Construction distance based model in longitudinal data
 - Parameter estimation
 - Hypotheses in the form $H\mathbf{B} = \mathbf{0}$
- 2 Simulation and Application
- 3 Conclusions
- 4 References

Outline

- 1 Background and solution
 - Background
 - Solution
 - Construction distance based model in longitudinal data
 - Parameter estimation
 - Hypotheses in the form $H\mathbf{B} = \mathbf{0}$
- 2 Simulation and Application
- 3 Conclusions
- 4 References

Background

- From a classical perspective, longitudinal data have been analyzed using the model of univariate analysis of variance or multivariate with repeated measurement.
- In many statistical methods and data analysis is used the geometric concept of distance between individuals or populations, these methods are applied in fields such as agronomy, anthropology, biology, genetics, psychology, etc Arenas (2002).

- Boj et al (2011) developed the dbstats package. This package contains functions for distance-based prediction methods. These are methods for prediction where predictor information is coded as a matrix of distances between individuals.
- Cuadras (1996) present some additional results based on distance model (DB) for prediction of mixed variables (continuous and categorical) and explore the problem of missing information giving a solution using DB.

Outline

- 1 Background and solution
 - Background
 - **Solution**
 - Construction distance based model in longitudinal data
 - Parameter estimation
 - Hypotheses in the form $HB = 0$
- 2 Simulation and Application
- 3 Conclusions
- 4 References

Solution through this work

- we propose the extension of distance based estimation methods in approach multivariate longitudinal data, using distances in the explanatory variables with continuous response variable. We study balanced data, where the number of times each individual is measured is the same, and the times are equally spaced.
- Furthermore, simulations were performed in R, where is compares the distance-based (DB) method in multivariate approach with respect to classical MANOVA based on the AIC and BIC information criterion, by structures of autocorrelation AR (1) and compound symmetry.

- The use of this strategy to model problems of this kind produces results equally robust than traditional modeling strategies and works in cases with categorical, binary, mixed and continuous explanatory variables.
- We prove that the predictions generated are the same under the proposed and classical models, except in mixed data using Gower distance.
- Both in application and simulation, was used the generalized least squares (GLS) of library (nlme) for the fit of the models in the univariate approach and MANOVA for the multivariate approach, along with some adaptations that were made for distances under certain correlation structures. Library (cluster): daisy function used to calculate the dissimilarity matrix.

Outline

- 1 Background and solution
 - Background
 - Solution
 - Construction distance based model in longitudinal data
 - Parameter estimation
 - Hypotheses in the form $HB = 0$
- 2 Simulation and Application
- 3 Conclusions
- 4 References

Multivariate model fitting, estimation, and hypothesis test

Let y_{ij} denote the response of i -th individual to the r -th evaluation condition, where $i = 1, \dots, n$ and $r = t_1, \dots, t_m$. It is assumed that y_{ir} follows the general linear model

$$y_{ir} = v_i' \beta_r + e_{ir}$$

where $v_i = (v_{i1}, \dots, v_{ip})'$ is a vector with p known specific coefficients for the i -th individual and $\beta_r = (\beta_{1r}, \dots, \beta_{pr})'$ is a vector with p unknown parameters.

Let $e_i = (e_{it_1}, \dots, e_{it_m})'$ be the residual vector of length m corresponding to the i -th individual, with distribution $e_i \sim NM(0_m, \Sigma)$.

In order to represent the model in matrix form, we define $Y_{n \times m} = [y'_1, \dots, y'_n]'$, $V_{n \times p} = [v'_1, \dots, v'_n]'$, $\beta_{p \times m} = [\beta_1, \dots, \beta_m]$ and $\mathbf{e}_{n \times m} = [e'_1, \dots, e'_n]'$. Where Y is the data matrix for the response variable, V is a design matrix of rank $p \leq (n - m)$, β holds the unknown parameters and \mathbf{e} contains the random errors. The model can then be expressed as

$$Y = V\beta + \mathbf{e} \quad (1)$$

Construction of longitudinal model with distances

Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a set consisting of n individuals. Let $\delta_{ii'} = \delta(\omega_i, \omega_{i'}) = \delta(\omega_{i'}, \omega_i) \geq \delta(\omega_i, \omega_i) = 0$ be a distance (or dissimilarity) function defined on Ω . Suppose that the distance matrix with dimension $n \times n$, $\Delta = (\delta_{ii'})$ is Euclidean. Then there exists a configuration of points $v_1, \dots, v_n \in \mathbb{R}^p$, where $v_i = (v_{i1}, \dots, v_{ip})'$, $i = 1, \dots, n$, such that

$$\delta_{ii'}^2 = \sum_{j=1}^p (v_{ij} - v_{i'j})^2 = (v_i - v_{i'})'(v_i - v_{i'}) \quad (2)$$

Once one of the distance measures presented above has been selected, we define $A_x = -\frac{1}{2}\Delta_x^{(2)}$ and $F_x = \mathcal{H}A_x\mathcal{H}$, where $\Delta_x^{(2)} = (\delta_{ii'}^2)$ and $\mathcal{H} = I - \frac{1}{n}\mathbf{1}\mathbf{1}' = I - \frac{1}{n}J$ is the centered matrix, where $\mathbf{1}$ is a $n \times 1$ column vector of ones and $J = \mathbf{1}\mathbf{1}'$. Also, F_x is a semi-positive definite matrix Mardia (1979) of rank p .

Therefore, the spectral decomposition

$$\begin{aligned} F_x &= \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) A_x \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \\ &= U_x \Lambda_x^2 U_x' = XX' \end{aligned} \quad (3)$$

where $X = U_x \Lambda_x$ is a $n \times p$ matrix of rank p , Λ_x is the matrix of positive eigenvalues of F_x and U_x contains the standardized coordinates. Also, rows x'_1, \dots, x'_n of matrix X are the principal coordinates of F_x .

The model that we propose is

$$Y = \mathbf{1}B_0 + XB + \Xi \quad (4)$$

where $\mathbf{1}$ is a $n \times 1$ column vector of ones, Y is like in model (1), $X_{n \times s}$ is the matrix of known values with $\text{rank}(X) = s$, $B_{s \times m}$ is the matrix of unknown parameters, B_0 is the $1 \times m$ row vector and Ξ is the error term matrix of dimension $n \times m$. It is convenient to split X in two parts, $X = (X_{(k)} \quad L)$, where $X_{(k)}$ contains a subset of k columns of X and L contains the rest. Thus, a k dimensional distance based model is obtained. Such model can be expressed as

$$Y = \mathbf{1}B_0 + X_{(k)}B_{(k)} + \Xi_k \quad (5)$$

and $X_{(k)} = (X_1, \dots, X_k)$ where each X_r , $r = 1, \dots, k$ is a column of X (and each X_i is a principal component).

Outline

- 1 Background and solution
 - Background
 - Solution
 - Construction distance based model in longitudinal data
 - Parameter estimation
 - Hypotheses in the form $H\mathbf{B} = \mathbf{0}$
- 2 Simulation and Application
- 3 Conclusions
- 4 References

Parameter estimation

The model in (5) can be represented as

$$Y = \mathbf{X}\mathbf{B} + \Xi_k \quad (6)$$

where $\mathbf{X} = (\mathbf{1} \quad X_{(k)}) = (\mathbf{1}, X_1, \dots, X_k)$ and $\mathbf{B} = (B'_0 \quad B'_{(k)})'$.

The least squares estimator (LSE), of \mathbf{B} , $\widehat{\mathbf{B}}$, is the matrix minimizing the trace of

$$tr(\widehat{\Xi}'_k \widehat{\Xi}_k) = tr\left[(Y - \mathbf{X}\widehat{\mathbf{B}})'(Y - \mathbf{X}\widehat{\mathbf{B}})\right]$$

where $\widehat{\Xi}_k = Y - \mathbf{X}\widehat{\mathbf{B}}$. The LS estimated parameters \mathbf{B} verify the normal equations (NE) and are given by

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'Y \quad (7)$$

Least squares estimation

The residual matrix $\mathbf{R}_0 = (R_0(i, j))$ of dimension $m \times m$, is given by

$$\begin{aligned}\mathbf{R}_0 &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\widehat{\mathbf{B}} \\ &= \mathbf{Y}' \left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right) \mathbf{Y}\end{aligned}\quad (8)$$

On the other hand, consider model (6) with restriction

$$\mathbf{H}\mathbf{B} = \mathbf{D} \quad (9)$$

where \mathbf{H} , \mathbf{B} , and \mathbf{D} have dimensions $s \times (k + 1)$, $(k + 1) \times m$ and $s \times m$, respectively. Minimizing under the restriction using Lagrange multipliers Λ , it can be shown that

$$\widehat{\mathbf{B}}_{r_1} = \widehat{\mathbf{B}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}' \left[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}' \right]^{-1} (\mathbf{D} - \mathbf{H}\widehat{\mathbf{B}}) \quad (10)$$

Hypotheses in the form $H\mathbf{B} = \mathbf{0}$

A testable linear hypothesis of rank s given by matrix H is

$$H_0 : H\mathbf{B} = \mathbf{0} \quad (11)$$

where the rows in H are a linear combination of those in \mathbf{X} . Restricting to hypothesis (11), it can be shown that

$$\widehat{\mathbf{B}}_{r_1} = \widehat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-1}H' [H(\mathbf{X}'\mathbf{X})^{-1}H']^{-1} H\widehat{\mathbf{B}}$$

and the residual matrix is

$$\mathbf{R}_1 = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}_{r_1})' (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}_{r_1})$$

- 1 $\mathbf{R}_0 \sim W_m(\Sigma, n - k)$.
- 2 If H_0 holds, matrices \mathbf{R}_0 and $\mathbf{R}_1 - \mathbf{R}_0$ follow Wishart distribution with $\mathbf{R}_1 \sim W_m(\Sigma, n - k')$, $\mathbf{R}_1 - \mathbf{R}_0 \sim W_m(\Sigma, s)$ where $s = \text{rank}(H)$ and $k' = k - s$.
- 3 If H_0 holds, matrices \mathbf{R}_0 and $\mathbf{R}_1 - \mathbf{R}_0$ are independent.

If H_0 holds, then \mathbf{R}_0 and $\mathbf{R}_1 - \mathbf{R}_0$ are Wishart independent and

$$\Lambda_{W_1} = \frac{|\mathbf{R}_0|}{|(\mathbf{R}_1 - \mathbf{R}_0) + \mathbf{R}_0|} = \frac{|\mathbf{R}_0|}{|\mathbf{R}_1|} \sim \Lambda(m; n - k, s)$$

Thus $0 \leq \Lambda_{W_1} \leq 1$ has the Wilks' distribution. H_0 will not be rejected if Λ is not significative and it will be rejected if Λ is small and significative.

Simulation

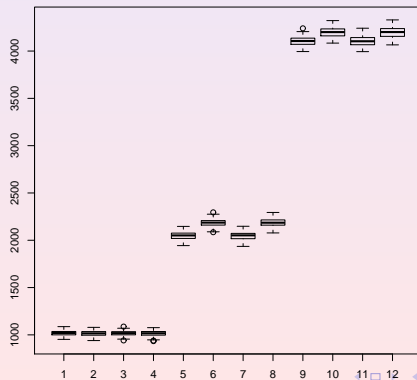
The continuous variable was sampled from a normal distribution with mean 100 and variance 100; the categorical variable was sampled from a multinomial distribution assuming three values with probabilities 0.27, 0.53, 0.2, and the binary one from a binomial distribution with $p = 0.4$. The response variable was continuous and was created from the model errors. The errors were created sampling a multivariate normal distribution with zero mean and covariance matrix Ψ_0 generated from two correlation structures: AR(1) and compound symmetry. We simulated 126 scenarios in total with $m = 4, 7, 10$, variances $\sigma^2 = 10, 50$, and correlations $\rho = -0.5, 0, 0.5, 0.9$, for sample sizes $n = 50, 100, 200$.

Simulation with compound symmetry correlation structure

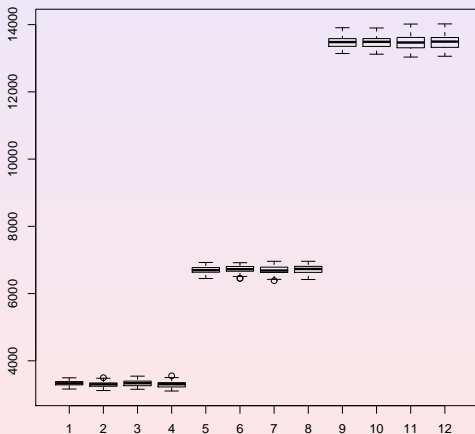
Parameters			$m = 10$			
			Distance based		Classical model	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	0	50	2502.51	2544.66	2523.18	2548.47
		100	5370.69	5419.768	5101.64	5131.09
		200	10429.75	10485.76	10241.09	10274.7
	0.5	50	2505.28	2547.43	2527.48	2552.76
		100	5363.78	5412.86	5094.74	5124.19
		200	10421.74	10477.74	10248.13	10281.74
	0.9	50	2504.25	2546.39	2527.78	2553.07
		100	5356.89	5405.97	5086.1	5115.55
		200	10418.79	10474.8	10249.33	10282.94
50	0	50	3291.77	3333.91	3327.90	3353.19
		100	6745.77	6794.84	6711.08	6740.52
		200	13472.72	13528.73	13459.97	13493.57
	0.5	50	3294.78	3336.92	3332.20	3357.48
		100	6738.33	6787.40	6704.18	6733.63
		200	13471.97	13527.98	13467.01	13500.61
	0.9	50	3293.92	3336.07	3332.50	3357.79
		100	6729.88	6778.96	6695.54	6724.99
		200	13470.7	13526.71	13468.21	13501.81

TABLE 2: Simulation with compound symmetry correlation structure

AIC criterion for distance-based and classic approaches with AR(1) and compound symmetry. $m = 4$, $\sigma^2 = 10$ and $\rho = 0.5$



The box-plots with choices $m = 10$, $\sigma^2 = 50$ and $\rho = 0.9$



Application

We applied this new approach to the study the effect of gender and exposure on the deviant behavior variable with respect to tolerance of deviant behavior for a group of children studied over a period of five years.

We initially performed an exploratory analysis of the data, where the empirical growth record can be observed, and how the changes differ significantly between youths, as Figure 1 reveals.

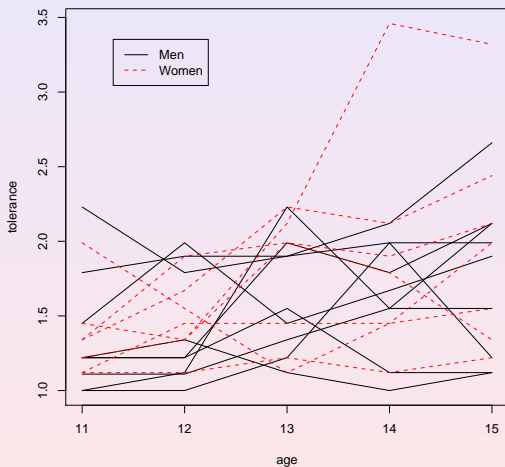
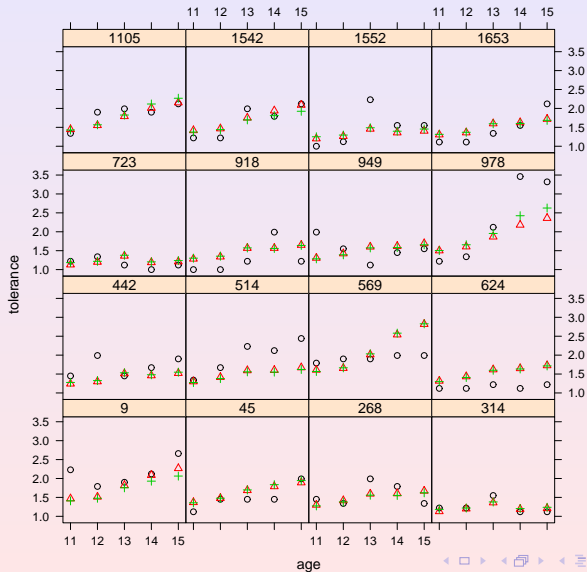


Figure: Tolerance based on age by gender



In order to illustrate the distance based approach, we fitted a MANOVA with DB to the data using Gower distance in the explicative variables (gender and exposure). The dependent variable is tolerance and the selected fit criterium was Akaike's (AIC), giving a value of -14.548 and BIC, -5.0204 . An adequate correlation structure was sought for. We also made a fit using the classic MANOVA model. The AIC and BIC values that we obtained in this case were -14.294 and -4.766 respectively, higher than with the distance based approach. However, the predictions resulted very similar with just some small differences in the last times.



Conclusions

- 1 The proposed methodology is also useful as the it provides a better fit than the classical models as additional components are added and better fit compared to the traditional approach in small sample with mixed data.
- 2 The use of this strategy is useful to predict both missing data and futures observations, which have lower errors of prediction than the traditional approach assuming normality.
- 3 We obtained similar results by both methods, but in large sample a small benefit using classical MANOVA.
- 4 A method for performing inference on the parameters of the model was presented.

References

-  Chaganty, N. R. & Mav, D. (2007), *Estimation methods for analyzing longitudinal data occurring in biomedical research*, Computational Methods in Biomedical Research 12, 371-400.
-  Cuadras, C. & Arenas, C. (1990), *A Distance Based Regression Model for Prediction with Mixed Data*, Communications in Statistics A. Theory and Methods, 19, 2261-2279.
-  Molenberghs, G. & Verbeke, G. (2005), *Models for discrete longitudinal data*, Springer, New York.
-  Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis-modeling change and event occurrence*, Oxford University Press, New York.

THANKS