Model validation and vec operators

Cite as: AIP Conference Proceedings **2040**, 110003 (2018); https://doi.org/10.1063/1.5079167
Published Online: 30 November 2018

Cristina Dias, Carla Santos, Maria Varadinov, and João Tiago Mexia

Model Validation and Vec Operators

Cristina Dias[1, a)], Carla Santos[2, b)] Maria Varadinov[3, c)] and João Tiago Mexia[4, d)]

[1]*Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Portalegre, Campus Politécnico, nº 10 7300-555 Portalegre, and CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Portugal*
[2] *Departamento de Matemática e Ciências Físicas do Instituto Politécnico de Beja, A R. de Pedro Soares, 7800-295 Beja, e CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Portugal*
[3]*Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Portalegre, and Coordenação Interdisciplinar para a Investigação e a Inovação(C3i),Campus Politécnico, nº 10,* 7300-555 Portalegre e
[4]*Departamento de Matemática da Faculdade de Ciências e Tecnologia e CMA–Centro de Matemática e Aplicações da Universidade Nova de Lisboa 2829- 516 Caparica, Portugal .*

[a)]Cristina Dias: cpsd@estgp.pt
[b)]Carla Santos: carla.santos@ipbeja.pt
[c)]Maria Varadinov: dinov@estgp.pt
[d)]João Tiago Mexia: jtm@fct.unl.pt.

Abstract. We use the **vec** and other relater operators to carry out inference for structured families of symmetric stochastic matrices $\mathbf{M}$. These are obtained through the sum of the respective mean matrix and a symmetric stochastic matrix with null mean. We consider that the **vec** operator of the matrix **E** is normal homoscedastic. The matrices on these families correspond to the treatments of a base design, and the inference is centered on model validation and the action of the factors in the base model on mean matrices.

INTRODUCTION

The models that we consider are based in the spectral decomposition of the mean matrices $\boldsymbol{\mu}$. In this work we propose a new formulation introducing vec operators, which simplify for symmetric stochastic matrices the adjustment and validation. This validation being new, give us a theoretical support for the use of rank one symmetric stochastic matrices. These vectorial operators, besides presenting themselves as an important part in the new formulation for these models, also facilitate the presentation of these results. The models for symmetric stochastic matrices are the basis for inference for isolated matrices and for structured families of matrices see, [7], [4], [1] and [3].

In these families the matrices, all of the same order, correspond to the treatments of base models. Since the matrices have all the same order, we are in the balanced case where we have the same number of degrees of freedom for the error for each treatment.

The ANOVA and related techniques are, in the balanced case, are robust techniques for heteroscedasticity and even more for non-normality, see [8] and [5].

The symmetric $n \times n$ matrix $\mathbf{W} = \lfloor w_{ij} \rfloor$ has, besides its vec the semi_vec $\mathbf{S}(\mathbf{M})$ with components $w_{1,1},\ldots,w_{n,n}, w_{2,3},\ldots,w_{2,n},\ldots,w_{n-1,n}$. We use the semi_vec to validate models for symmetric stochastic matrices. When the predominance of the first eigenvalue is very large, we can adopt a model of degree one of the form

International Conference of Computational Methods in Sciences and Engineering 2018 (ICCMSE 2018)
AIP Conf. Proc. 2040, 110003-1–110003-4; https://doi.org/10.1063/1.5079167
Published by AIP Publishing. 978-0-7354-1766-3/$30.00

$$\mathbf{M} = \boldsymbol{\mu} + \overline{\mathbf{E}} \, ,$$

where $\boldsymbol{\mu}$ is a mean matrix and $\overline{\mathbf{E}}$ is a symmetric stochastic matrix with null mean, such that $\overline{\mathbf{E}} = \frac{1}{2}(\mathbf{E} + \mathbf{E}^t)$ with $vec(\mathbf{E}) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n^2})$ normal homoscedastic.

These models have been used successfully in several applications, namely the first phase of the STATIS methodology, the inter-structure. Since Hilbert-Schmidt matrices are matrices of cross products, we can use them for the latter, changing to higher-grade models if degree one models do not fit. In what follows we will base ourselves on presenting the main results that we have for this models.

Simulations presented in the next section show that when the first eigenvalue is sufficiently dominant we can conclude that the mean matrix as rank one and that the first eigenvalue and eigenvector can be used to estimate the sole non null eigenvalue of the mean matrix and the corresponding eigenvector.

We intend to extend our treatment to structured families in which for each treatment of a base design. These treatments correspond to the level combinations of the factor in base design.

ADJUSTMENT OF A SINGLE MODEL

The model for **M** has rank given by the characteristic of the mean matrix $\boldsymbol{\mu}$, with $\boldsymbol{\mu} = \lambda \boldsymbol{\alpha}\boldsymbol{\alpha}^t$.

Let $\theta_1 \geq ... \geq \theta_n$ be the eigenvalues and $\boldsymbol{\gamma}_1,...,\gamma_n$ the eigenvectors of $\mathbf{M}$. Suppose that $\theta_1 >> \theta_2$, now admitting that $\theta_i \approx 0$, $i = 2,...,n$. With

$$\|\mathbf{M}\|^2 = \sum_{i=1}^{n} \theta_i^2 \, ,$$

we will have

$$c_i = \frac{\theta_1^2}{\sum_{l=i+1}^{k} \theta_i^2} \, ,$$

if $\mathbf{c}_i$ is large we can take $(\theta_i, \boldsymbol{\gamma}_i), i = 1,...,n$ as estimators of $(\lambda_i, \alpha_i), i = 1,...,n$, see [2], where $(\theta_i, \boldsymbol{\gamma}_i), i = 1,...,n$, are the pairs of eigenvalues and eigenvectors for **M**. When the predominance of the first eigenvalue is very large, we can adopt a model of degree one of the form

$$\mathbf{M} = \lambda \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathbf{t}} + \overline{\mathbf{E}} \, \mathbf{,}$$

where $\overline{\mathbf{E}}$ is a symmetric stochastic matrix with null mean vector and variance–covariance matrix $\sigma^2 \mathbf{I}_{n^2}$ that is $vec(\mathbf{E}) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n^2})$. A simulation study can be used to give additional validation to these models. Simulations show that, when the preponderance of the fixed component of the model is much larger than one, the first eigenvalue θ_1 and the first eigenvector $\boldsymbol{\gamma}_1$ of **M** are good estimators of λ and $\boldsymbol{\alpha}$, respectively.

The values of k chosen were: 6, 8, 10, 12 and the values for c_i were: 12.5, 50, 200 and 400 (the simulation programs were developed using the R application: http://www.r-project.org).

In the construction of the matrices $\mathbf{M}$ it was taken $\boldsymbol{\alpha} = \boldsymbol{\delta}_1$. For each pair (k, c_i), 1000 matrices **E** were generated from which we obtained the matrices $\mathbf{M} = \lambda \boldsymbol{\delta}_1 \boldsymbol{\delta}_1^{\mathbf{t}} + \overline{\mathbf{E}}$, with $\lambda = \sqrt{2c_i}$ and $\boldsymbol{\alpha} = \boldsymbol{\delta}_1$. Being $\boldsymbol{\gamma}_1$ the first eigenvector of $\mathbf{M}$, since $\|\boldsymbol{\delta}_1\|^2 = \|\boldsymbol{\gamma}_1\|^2 = 1$, we will have

$$\mathbf{Z} = \|\boldsymbol{\alpha} - \boldsymbol{\gamma}\|^2 = \sum_{i=1}^{k} (\boldsymbol{\alpha}_i - \boldsymbol{\gamma}_i)^2 = (\boldsymbol{\alpha}_1 - 1)^2 + \sum_{i=2}^{k} \boldsymbol{\alpha}_i^2 \approx 0$$

Thus, in Table 1 we present results for respective mean values and standard deviation for $\mathbf{Z}$.

TABLE 1. Mean values and standard deviations for **Z**.

$\mathbf{c}_i$	$k = 6$		$k = 8$		$k = 10$		k = 12	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
12.5	0.673	0.283	0594	0.291	0.499	0.294	0.446	0.284
50	0.940	0.060	0.909	0.117	0.877	0.129	0.842	0.158
200	0.987	0.009	0.982	0.010	0.77	0.012	0.971	0.014
400	0.997	0.002	0.996	0.002	0.994	0.003	0.993	0.003

From the analysis of the Table 1 we see that, when $c_i > 200$ we can use θ_1 and $\boldsymbol{\gamma}_1$ as estimators of λ and $\boldsymbol{\alpha}$ respectively. Let us put

$$\overline{\mathbf{M}}_h = \mathbf{M} - \sum_{l=1}^{h} \theta_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^t ,$$

where $m_h = \mathbf{S}(\overline{\mathbf{M}}_h)$. If $k \leq h$ we will have $\overline{\mathbf{M}}_h \approx \mathbf{E}$ and so the components of $\mathbf{S}(\overline{\mathbf{M}}_h)$ can be assumed to be normal with null mean values. With $Y_1,..., Y_g$ $(g = n(n-1))$ the components of $\mathbf{S}(\mathbf{M}_h)$ we are then led to test

$$H_{0k}: \quad Y_1,\ldots,Y_g \text{ i.i.d.} \sim N(0,\sigma^2),$$

using the statistics

$$\mathcal{F} = \frac{(g-1)\ (\sum_{i=1}^{g} Y_i)^2}{g\sum_{i=1}^{g} Y_i^2 - (\sum_{i=1}^{g} Y_i)^2},$$

which when the hypothesis holds, has the central F distribution with 1 and g-1 degrees of freedom, see [6]. When H_0 thus not hold, the $Y_1,..., Y_g$ will have mean values $\mu_1,\ldots,\mu_g$ and the numerator and denominator of F will have the non-centrality parameter $\delta_1 = \frac{1}{\sigma^2} \ (\frac{1}{g} \sum_{i=1}^{g} \mu_i)^2)$, and $\delta_2 = \frac{1}{\sigma^2} \ (\sum_{i=1}^{g} \mu_i^2 - \frac{1}{g}(\sum_{i=1}^{g} \mu_i)^2)$, hence, there will be alternatives in which δ_1 predominates over δ_2 (δ_2 predominates over δ_1 and in which *F* tends to take larger (smaller) values than when H_0 holds. A numerical study showed that, with $f_{p,1,g-1}$ the quantile for probability *p* of the central F distribution with 1 and g-1 degrees of freedom, the acceptance region [$f_{\frac{p}{2},1,g-1}; f_{1-\frac{p}{2},1,g-1}$] leads to p level tests with power increasing rapidly with δ_1 and δ_2 , see [6].

We test this hypothesis for increasing values of k. With $\overline{k}$ the smallest value for which the hypothesis is not rejected we will have the degree of the adjusted model given by

$$\mathbf{M}^* = \sum_{i=1}^{\overline{k}} \theta_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^t .$$

We can now complete the adjustment testing hypothesis that $\gamma_{\bar{k}+1},\ldots,\gamma_n$, $\bar{k}=1,\ldots,n-1$, are normal homoscedastic with null mean vectors independent between them self's, and from $\gamma_{\bar{k}}$, which will have a mean vector $\gamma_{\bar{k}} \neq 0$, using the test statistic

$$\mathcal{F}_{\bar{k}} = (n-\bar{k})\frac{\left\|\gamma_{\bar{k}}\right\|^2}{\sum_{h=\bar{k}+1}^{n}\left\|\gamma_h\right\|^2}.$$

This F test has n and $(n-\bar{k})\,n$ degrees of freedom and non-centrality parameter $\delta_h = \frac{1}{\sigma^2}\left\|\bar{\gamma}_k\right\|^2$, and the p level critical value $F_{1-p,n,(n-\bar{k})n}$.

FINAL REMARKS

We showed how the semi_vec can be used to determinate the number of non-null eigenvalues of a symmetric matrix $\mathbf{M} = \boldsymbol{\mu} + \bar{\mathbf{E}}$, with mean matrix $\boldsymbol{\mu}$, $\bar{\mathbf{E}} = \frac{1}{2}(\mathbf{E}+\mathbf{E}^t)$ and vec($\mathbf{E}$) normal homoscedastic. Moreover once the characteristic of $\boldsymbol{\mu}$ was determined we obtained, using adjusted eigenvalues and eigenvectors, an adjusted model for $\mathbf{M}$. In the future we will extend this treatments to find the least value $\bar{k}$ of non null eigenvalues in order to have an adjusted model $\mathbf{M}^* = \sum_{i=1}^{\bar{k}} \theta_i \gamma_i \gamma_i^t$.

ACKNOWLEDGMENTS

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

REFERENCES

1. A. Areia, M. M. Oliveira, & J. T. Mexia, *Modelling the Compromise Matrix in STATIS*, *Methodology,* 2011, 5, pp. 277-288.
2. C. Dias, "Modelos e Famílias de Modelos para Matrizes Estocásticas Simétricas", Ph.D. Thesis, Évora University, 2013.
3. C. Dias, S. Santos, M. Varadinov & J.T. Mexia, "ANOVA Like Analysis for Structured Families of Stochastic Matrices", AIP Proceedings of the 12 International Conference of computational Methods in Sciences and Engineering (ICCMSE 2016), edited by American Institute of Physics, 2016, 1790, pp. 140006-1-140006-3
4. E. E. Moreira, "Famílias Estruturadas de Modelos com Modelo Base Ortogonal:Teoria e Aplicações", Ph.D. Thesis, Faculty of Science and Technology, New University of Lisbon, 2008.
5. H. Scheffé, *The Analysis of Variance*, New York: John Wiley & Sons, 1959.
6. J. T. Mexia, Controlled heteroscedasticity, quotient vector spaces and *F* tests for hypotheses on mean vectors, Trabalhos de Investigação, nº2, Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 1989.
7. M. M. Oliveira and J. T. Mexia, AIDS in Portugal: endemic versus epidemic forecasting scenarios for mortality, *International Journal of Forecasting*, 2004, 20, pp.131–137.
8. P. K. Ito, *Robustness of Anova and Macanova Test Procedures*, P. R. Krishnaiah (ed), Handbook of Statistics1, Amsterdam: North Holland, 1980, pp. 199-236.