University of São Paulo
"Luiz de Queiroz" College of Agriculture

On testing genetic covariance via the mean cross-products ratio

Anderson Rodrigo da Silva

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agronomic Experimentation

Piracicaba
2015

Anderson Rodrigo da Silva
Agronomist

On testing genetic covariance via the mean cross-products ratio

Advisor:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agronomic Experimentation

Piracicaba
2015

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Silva, Anderson Rodrigo da
On testing genetic covariance via the mean cross-products ratio / Anderson Rodrigo da Silva. - - Piracicaba, 2015.
92 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura “Luiz de Queiroz”.

1. Correlação genética 2. MANOVA 3. Simulação de Monte Carlo 4. Bootstrap 5. Lambda de Wilks I. Título

CDD 519.535
S586t

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

ACKNOWLEDGMENTS

To my family, Bárbara (wife), Ana (mother), José Luiz (father) and Amanda (sister), and friends for all kind of support, love and patience.

To my advisor, Professor Carlos Tadeu, for all the directions on this work.

To Professor Cláudio Lopes, Department of Genetics of the "Luiz de Queiroz" College of Agriculture - University of São Paulo, for presenting us the problem that has originated this study.

I am grateful to my supervisors in the University of Exeter (UK), Dr. Trevor Bailey and Dr. Wojtek Krzanowski, for all the meetings and their helpful comments.

To Dr. Kent Eskridge, University of Nebraska-Lincoln (USA), for reading my project and giving me directions on the Wilks' Lambda approach.

A word of thanks to Professor Maurício Mota, Department of Mathematics and Statistics of the Federal University of Ceará, for giving some attention on the exact distribution of two independent Gaussian variables.

To the "Luiz de Queiroz" College of Agriculture - University of São Paulo for welcoming me as a Ph.D. student.

To CAPES and CNPq for the scholarships and financial support in Brazil and overseas.

Loads of thanks to the Goiano Federal Institute for allowing me the time off needed to finish this thesis. And to my job's colleagues over there.

To all the authors cited in this thesis. Their findings are the basis of this work.

Finally, I would like to thank Robert Gentleman, Ross Ihaka and all the core group that have developed the most amazing statistical software ever, R (of course)!

The only relevant test of the validity of a hypothesis is comparison of prediction with experience.

Milton Friedman

CONTENTS

RESUMO

Teste da covariância genética via razão de produtos cruzados médios

Quando um fator genético está sendo estudado em mais de uma variável de resposta, estimativas das covariâncias genéticas são essenciais, especialmente para programas de melhoramento. Em uma análise de covariância genética, produtos cruzados médios devido ao efeito genético, a partir do qual é obtida a covariância genética, e devido ao efeito residual são obtidos. Estocasticamente, quantificar a magnitude da variação conjunta de duas variáveis resposta devido ao efeito genético em relação à variação devida ao efeito residual pode permitir realizar inferências sobre a covariância genética associada. Neste estudo são apresentados testes de significância para a covariância genética de duas formas: testes que levam em conta os efeitos genéticos e ambientais (ou residuais) e testes que consideram apenas a informação genética. A primeira forma refere-se testes baseados na razão de produtos cruzados médios via bootstrap não paramétrico e simulação de matrizes Wishart pelo método de Monte Carlo. A segunda maneira de testar a covariância genética refere-se a testes com base em uma adaptação das estatísticas de Wilks e Pillai para avaliar a independência de dois conjuntos de variáveis. Para o primeiro tipo de testes, as distribuições empíricas sob a hipótese nula, ou seja, covariância genética nula, foram construídas e analisadas graficamente. Além disso, foi feito um estudo analítico da distribuição da razão de produtos cruzados médios obtidos a partir de variáveis normalmente distribuídas com média zero e variância finita. Escrever algoritmos computacionais em linguagem R para realizar os testes propostos também foi um dos objetivos deste estudo. Apenas sob certas condições a função de densidade de probabilidade do produto de duas variáveis aleatórias gaussianas aproxima-se da curva normal. Por conseguinte, o estudo da distribuição da razão de produtos cruzados médios como um quociente de duas variáveis gaussianas não é adequado. Os testes baseados na razão de produtos cruzados médios estão relacionados tanto com o valor da covariância genética quanto com a magnitude desta em relação à covariância residual. Ambas as abordagens (bootstrap e simulação) mostraram-se mais sensíveis do que os testes baseados apenas nas informações genéticas. O desempenho dos testes baseados na razão de produtos cruzados médios está relacionado à qualidade dos dados originais em termos das pressuposições da MANOVA, e a estatistica de teste não depende da estimação da matriz de covariâncias genéticas $\mathbf{\Sigma}_G$. A adaptação das estatísticas de Wilks e Pillai pode ser usada para testar a covariância genética. As aproximações à distribuição χ^2_1 foi verificada. A precisão de suas inferências está relacionada a qualidade da matriz $\mathbf{G}$.

Palavras-chave: Correlação genética; MANOVA; Simulação de Monte Carlo; Bootstrap; Lambda de Wilks

ABSTRACT

On testing genetic covariance via the mean cross-products ratio

When a genetic factor is being studied for more than one response variable, estimates of the genetic covariances are essential, specially in breeding programs. In a genetic covariance analysis, genetic and residual mean cross-products are obtained. Stochastically, to quantify the magnitude of the joint variation of two response variables due to genetic effect with respect to the variation due to residual effect may allow one to make inferences about the significance of the associated genetic covariance. In this study it is presented tests of significance for genetic covariance upon a twofold way: tests that take into account the genetic and environmental effects and tests that only consider the genetic information. The first way refers to tests based on the mean cross-products ratio via nonparametric bootstrap resampling and Monte Carlo simulation of Wishart matrices. The second way of testing genetic covariance refers to tests based on adaptation of Wilks' and Pillai's statistics for evaluating independence of two sets of variables. For the first type of tests, empirical distributions under the null hypothesis, i.e., null genetic covariance, were built and graphically analyzed. In addition, the exact distribution of mean cross-products ratio obtained from variables normally distributed with zero mean and finite variance was examined. Writing computational algorithms in R language to perform the proposed tests was also an objective of this study. Only under certain conditions does the probability density function of the product of two random Gaussian variables approximate a normal curve. Therefore, studying the distribution of a mean cross-products ratio as a quotient of two Gaussian variables is not suitable. Tests based on mean cross-products ratio are related to both the value of the genetic covariance and the magnitude of the latter relative to the residual covariance. And both approaches (bootstrap and simulation) are more sensitive than the tests based only on genetic information. The performance of the tests based on mean cross-products ratio is related to the quality of the original data set in terms of the MANOVA assumptions, and the test statistic does not depend on the estimation of the matrix of genetic covariances $\mathbf{\Sigma}_G$. The adaptation of Wilks' and Pillai's statistics can be used to test the genetic covariance. Their approximations to a χ^2_1 distribution were checked and the accuracy of their inferences is related to the quality of $\mathbf{G}$.

Keywords: Genetic correlation; MANOVA; Monte Carlo simulation; Bootstrap; Wilks' Lambda

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In quantitative genetics, there are many important parameters when strategies of conservation of natural genetic variability are determined (CARLINI-GARCIA et al., 2001). Often, testing the significance of these parameters is an important issue. However, during the estimation process the standard errors are not always supplied and there is little information about the actual distribution of the estimates. Besides, obtaining explicit expressions for these standard errors is not an easy task, since the estimators are usually ratios of random variables with unknown distributions. The authors used bootstrap resampling to estimate the probability distribution as well as the standard error of genetic parameters. Likewise, Reis et al. (2009) studied additive genetic variance from selected and non-selected populations using Monte Carlo simulation.

Testing genetic covariance would be of great benefit to plant and animal breeding, since parameters directly related to the genetic covariance, such as the correlated response to selection, are crucial to the adoption of a breeding selection method. According to Guillaume e Whitlock (2007), the genetic covariance can be useful for predicting response to indirect selection of one character from another. If the genetic covariance between two characters is not null, selecting one will affect the response to selection on the other.

As in the analysis of variance when the interest is whether a genetic factor has any effect on the variability of a particular response variable, in the *analysis of genetic covariance* (ROBERTSON, 1959; KEMPTHORNE, 1969) one may attempt to identify whether there is any significant effect of this factor on the joint variation of two particular response variables. In this context, under the assumptions of the analysis of variance model, it is known that the two variances (mean squares) obtained for the genetic and residual effects both follow independent χ^2 distributions, and therefore the ratio between them produces an F variable, which can be used to assess the magnitude of the numerator relative to its denominator. In the case of the analysis of genetic covariance or, more broadly, in a multivariate analysis of variance, covariances (*mean cross-products*) are obtained, also containing genetic and residual effects. The latter is sometimes called environmental covariance. However, unlike a mean squares ratio, the distribution of a mean cross-products ratio from normally distributed variables is still unknown.

In the literature only an approximate test of the genetic correlation coefficient is available. The result of this test is commonly used to make indirect inferences about the genetic covariance, resulting in higher approximation errors. Furthermore, Falconer e Mackay (1996) stated that estimates of genetic correlations are usually subjected to sampling errors, and therefore these estimates are seldom accurate. The authors also pointed out that the sampling variance of a genetic correlation is quite complex. Some authors (LEMOS et al., 1992; MALIK et al., 2005; BARROS et al., 2010; AJAYI et al., 2014) have also tested genetic correlation through the Student's t-test. Nevertheless,

there is some misunderstanding about the degrees of freedom associated with the genetic correlation, as verified by (FERREIRA et al., 2008). The authors stated that bootstrap is a reliable way for testing genetic and environmental correlation. It is a fact that in these hypothesis testing approaches only the genetic information is used. Hence, two questions arise: 1) can the genetic covariance contribute more to the phenotypic variation than the residual covariance, even when the genetic correlation is low? 2) is it possible to test genetic covariance by considering its contribution to the phenotypic variation? The answer to the first question is intuitive and straightforward. The answer to the second provides the background of this study.

In this study it is presented tests for genetic covariance upon a twofold way: tests that take into account the genetic and environmental effects and tests that only consider the genetic information. The first way refers to tests based on mean cross-products ratio via nonparametric bootstrap resampling and Monte Carlo simulation of Wishart matrices. The second way of testing genetic covariance refers to tests based on an adaptation of Wilks' and Pillai's statistics for evaluating independence of two sets of variables.

Because the first type consists of building empirical distributions under the null hypothesis, i.e., null genetic covariance, these distributions were graphically analyzed. Nonetheless, this work aims to study analytically the mean cross-products ratio obtained from variables normally distributed with zero mean and finite variance.

Writing computational algorithms in R language to perform the proposed tests is also an objective of this study.

2 LITERATURE REVIEW

2.1 Genetic covariance and correlation

2.1.1 Importance and use

According to Resende (2002) in the context of plant breeding, understanding the genetic correlation is particularly useful, especially for the implementation of indirect selection on the characters that present difficulties of measurement, identification and/or low heritability. Selecting another character more easily assessed and with high heritability provides greater genetic progress, saving time, labor and resources.

Evolution by natural selection requires heritable variation and the most common way of representing the pattern and magnitude of the genetic basis of a number of characteristics is through the matrix **G** of genetic variances and covariances. As **G** contains additive genetic covariances information, it may be useful to predict the indirect response to selection of a character from another. If the covariance between two characters is different from zero, select one will affect the response to selection on the other (GUILLAUME; WHITLOCK, 2007).

Robertson (1959) reports that for continuous genetic variables, the correlation coefficient between two characters is an integral part of the discussion of correlated responses to selection and the combination of different measures ensures maximum improvement, then called selection index.

According to Falconer e Mackay (1996), correlated traits are of interest for three main reasons. First, in connection with genetic causes of correlation via pleiotropy[1], leading cause of genetic correlation and property common to the main genes. Second, in connection with the changes caused by selection, since it is important to know how improving a character will cause simultaneous changes in another character. And third, in connection with natural selection, since the relationship between a metric character and adaptation are primary agents determining the properties of this character in a natural population.

The association between two characters can be directly observed by the correlation among phenotypic individual values, that is, the *phenotypic correlation*. Knowing the genotypic values and environmental deviations for both characters, one can access independently the genetic and environmental causes of the correlation. If, in addition, genetic values of individuals are known, then it is possible to separate the variance of a character into two parts, additive genetic versus the rest (environmental), i.e., environmental and non-additive genetic deviations. Thus, the phenotypic covariance (σ_P) can be taken as the sum of the genetic (σ_G) and environmental (σ_E) covariances, i.e.,

[1]Property of a gene through which affects two or more characters.

$$\sigma_P = \sigma_G + \sigma_E \tag{1}$$

The genetic correlation explains the additive components (heritable part of the association) and the environmental correlation, the non-additive components. They determine the phenotypic correlation calculated from measurements of the characteristics in the population (KOMINAKIS, 2003).

The genetic covariance between characters from a population of random crosses can be maintained in two ways: through gene with pleiotropic (multiple) effects and via linkage disequilibrium (statistical dependence) among alleles at different loci affecting different characters (LANDE, 1980; MEREDITH, 1984).

For genetic breeding, specifically for quantitative genetics, the knowledge of genetic covariance (or correlation) is of great importance in the following contexts:

- Indirect selection or prediction of gain by correlated response to selection;
- Development of selection indices to select multiple characters simultaneously;
- Using breeding strategies in accordance with the determination of the extension of the genotype-environment interactions;
- Understanding the evolutionary process of characters.

According to Wootton e Smith (2014), genetic correlations combined with selection provide a measure of response of direct causal relationship among life-history traits, and thereby a qualitative measure of response to selection. In this context, they are superior to purely phenotypic correlations. The limitations of genetic correlation studies are that the outcomes may be specific to the environment in which they are conducted.

Vencovsky e Barriga (1992) emphasize that one of the most important implications of the correlations relates to selection. So it is worth asking, for example, what will be the change in the average of the character Y when selecting for the character X? In recurrent selection involving the additive effects of alleles, this indirect change is given by

$$RC_{Y,X} = ds_X \times \left(\frac{\sigma_{G(X,Y)}}{\sigma^2_{P(X)}} \right) \times a, \tag{2}$$

where $RC_{Y,X}$ is the correlated response to selection in Y when selecting the character X; ds_X is the selection differential related to X; $\sigma_{G(X,Y)}$ is the genetic covariance between X and Y; $\sigma^2_{P(X)}$ is the phenotypic covariance of X; the coefficient a depends on the scheme of selection adopted.

The authors also point out that a genetic breeding program on Y can be obtained more efficiently by selecting only X. That is the situation where $RC_{Y,X} > G_{sY}$, being G_{sY} the expected progress value on Y, which is not common, but could happen in recurrent

selection when $\hat{h}_X r_G > \hat{h}_Y$ (FALCONER; MACKAY, 1996), where $\hat{h}_{\cdot}$ represents the square root of the narrow-sense heritability coefficient (only the additive part), i.e., when the selection does not explore all the genetic variance.

2.1.2 Analysis of genetic (co)variance

The analysis of genetic covariance is a procedure analogous to the analysis of variance when the goal is to check whether the genetic variation in the population of genotypes being assessed is not null, with the difference that the first considers two or more variables simultaneously.

Data from experiments of evaluations of genotypes, as in many agronomic experiments, usually consist of observations of several variables, which are seldom considered independent from each other. In cases where there is interdependence among variables, analyses of variance and genetic covariance would be more informative in assessing genetic and other experimental effects of interest, because the information of covariances (or mean cross-products) associated with each source of variation is taken into account.

According to Dutilleul e Carriere (1998), the two-way analysis of variance with genotype and environment as crossed factors is the usual basis for estimating genetic correlation (and covariance), as presented by Robertson (1959) and Kempthorne (1969), as follows.

Consider performing the analysis of genetic covariance, say for two variables X and Y. It is perfectly possible to consider the composed observation $x + y$ for each observational unit and do the usual analysis of variance for both variables, and also for the *dummy* variable $X+Y$. The formulas used to obtain the mean squares can be applied to obtain variance components associated with $X + Y$. Thus, it follows that

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y), \tag{3}$$

then,

$$\sigma^2_{h(X+Y)} = \sigma^2_{h(X)} + \sigma^2_{h(Y)} + 2\sigma_{h(XY)}, \tag{4}$$

where $\sigma^2_{h(X+Y)}$, $\sigma^2_{h(X)}$ e $\sigma^2_{h(Y)}$ are the variance components of $X+Y$, X and Y, respectively, assigned to the source h; $\sigma_{h(XY)}$ is the covariance component of X and Y assigned to the source h.

Thus, the estimate of the covariance component σ_{XY} can be obtained by

$$\hat{\sigma}_{h(XY)} = \frac{\hat{\sigma}^2_{h(X+Y)} - \hat{\sigma}^2_{h(X)} - \hat{\sigma}^2_{h(X)}}{2}, \tag{5}$$

at which the variance components are determined using the *Method of Moments*[2].

[2]Method of estimation that consists of matching r sample moments to their respectives populational,

Kempthorne (1969) highlights that the procedure serves completely to define the estimation of covariance components. For example, consider data of two variables, say X e Y, from an experiment carried out under a randomized block design where g genotypes were evaluated on r blocks. For Y (also valid for X), the *random model* of analysis of variance would be

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{6}$$

where,
y_{ij} is the observation of Y for the i-th genotype at the j-th block;
μ is the population mean of Y;
α_i is the random effect associated with the i-th genotype ($i = 1, 2, ..., g$);
β_j is the random effect associated with the j-th block ($j = 1, 2, ..., r$);
ϵ_{ij} is the effect of the random error related to y_{ij}.

The scheme of a analysis of genetic covariance for X and Y could be summarized in Table 1.

Table 1 – Analysis of genetic covariance for X and Y in a randomized block design

Source	Degrees of freedom	Mean squares			Mean cross-products
		X	Y	Z^*	
Blocks	$r-1$	QB_X	QB_Y	QB_Z	$PB_{XY} = \frac{1}{2}(QB_Z - QB_X - QB_Y)$
Genotypes	$g-1$	QG_X	QG_Y	QG_Z	$PG_{XY} = \frac{1}{2}(QG_Z - QG_X - QG_Y)$
Residuals	$(g-1)(r-1)$	QR_X	QR_Y	QR_Z	$PR_{XY} = \frac{1}{2}(QR_Z - QR_X - QR_Y)$

$^*Z = X + Y$

The expected values of the mean squares for X are presented in eq. 7. The same is valid for Y and $Z = X + Y$.

$$\begin{aligned} E(QB_X) &= \sigma_{XX} + g\sigma_{B(XX)} \\ E(QG_X) &= \sigma_{XX} + r\sigma_{G(XX)} \\ E(QR_X) &= \sigma_{XX} \end{aligned} \tag{7}$$

where σ_{XX}, $\sigma_{B(XX)}$ and $\sigma_{G(XX)}$ are the residual variance, the variance due to block effects and the genetic variance of X, respectively.

The expected value of the mean cross-products are:

forming a system of r equations. In the analysis of variance it is used to estimate variance components, matching mean squares to their respectives expected values.

$$\begin{aligned} E(PB_{XY}) &= \sigma_{XY} + g\sigma_{B(XY)} \\ E(PG_{XY}) &= \sigma_{XY} + r\sigma_{G(XY)} \\ E(PR_{XY}) &= \sigma_{XY} \end{aligned} \tag{8}$$

where σ_{XY}, $\sigma_{B(XY)}$ and $\sigma_{G(XY)}$ are the residual covariance, the covariance due to block effects and the genetic covariance between X and Y, respectively.

As early described, obtaining the expected values allows one to estimate (co)variance components related to each source. Then,

$$\begin{aligned} \hat{\sigma}_{XX} &= QR_X \\ \hat{\sigma}_{XY} &= PR_{XY} \\ \hat{\sigma}_{B(XX)} &= \tfrac{1}{g}(QB_X - QR_X) \\ \hat{\sigma}_{B(XY)} &= \tfrac{1}{g}(PB_{XY} - PR_{XY}) \\ \hat{\sigma}_{G(XX)} &= \tfrac{1}{r}(QG_X - QR_X) \\ \hat{\sigma}_{G(XY)} &= \tfrac{1}{r}(PG_{XY} - PR_{XY}), \end{aligned} \tag{9}$$

where the latter equation expresses the moments estimator of the genetic covariance between X and Y. Therefore, the genetic correlation can be calculated by

$$\hat{\rho}_{G(XY)} = \frac{\hat{\sigma}_{G(XY)}}{\left(\hat{\sigma}_{G(XX)}\hat{\sigma}_{G(YY)}\right)^{1/2}}. \tag{10}$$

And the residual correlation is obtained by

$$\hat{\rho}_{R(XY)} = \frac{\hat{\sigma}_{XY}}{\left(\hat{\sigma}_{XX}\hat{\sigma}_{YY}\right)^{1/2}}. \tag{11}$$

The analogy with the familiar formula of correlation coefficient is clear. The genetic covariance between two phenotypes is quite distinct from the genetic correlation. It is possible for two traits to have a very high genetic correlation yet have little genetic covariance. Low genetic covariance could arise if either trait had low genetic variance (NEALE; CARDON, 1992).

Now, an important question is whether the genetic covariance is significant, i.e., whether the joint variation between X and Y due to genetic effects is greater than that due to environment (residuals) on the population, so that quantities such as the correlated response to selection (eq. 2), directly related to genetic covariance, has an effective meaning.

2.1.3 Genetic correlation test

Estimates of genetic correlation are usually subjected to very large sampling errors and therefore these estimates are rarely accurate, as state by Falconer e Mackay (1996). The authors also point out that the sample variance of a genetic correlation is complicated and, considering two characters, X and Y for example, it has the following approximate formula (derived from Reeve (1955) and Robertson (1959)) for the standard error of the additive genetic correlation (r_G):

$$\sigma_{(r_G)} = \frac{1 - r_G^2}{\sqrt{2}} \sqrt{\frac{\sigma_{(h_X^2)} \sigma_{(h_X^2)}}{h_X^2 h_Y^2}}, \tag{12}$$

where σ denotes standard error; h_X^2 e h_Y^2 are the narrow-sense heritability for the variables X and Y, respectively.

Waitt e Levin (1998) state that genetic correlations are often difficult (require large sample sizes of individuals of known relatedness) or impossible (rare, endangered or extinct species) to obtain (CHEVERUD, 1988). In comparison, phenotypic correlations are easily and accurately estimated requiring only moderate sample sizes and no knowledge of relatedness among individuals. Roff (1995) states that sample sizes required to achieve a small standard error for genetic correlations are typically enormous. Nonetheles, according to Lynch e Walsh (1998), the *Delta Method*[3] is considered better for estimating the variance of a ratio of unknown distributions.

Zeng et al. (2007) estimated the genetic correlation between yield and fiber quality of *Gossypium* species using the methodology described by Kempthorne (1969). However, the significance was not determined, because there was no adequate probability table available for a statistical test of genetic correlation. Lemos et al. (1992) used the same estimation procedure to estimate genotypic, phenotypic and environmental correlations among maize variables. The authors performed the Student's t-test to access significance. Likewise, and Barros et al. (2010) also tested genetic correlations among maize variables by using the Student's t-test. Ajayi et al. (2014) tested genotypic correlations among cowpea variables using the t-test.

According to Ferreira et al. (2008), there is no accurate test to assess the significance of genetic correlations, and the hypothesis that this coefficient is zero can not be evaluated with the usual t test on $n - 2$ degrees of freedom. The estimate of genetic correlation is obtained for the mean square due to genotype and residual, each associated with a different number of degrees of freedom, which makes difficult to establish the degrees of freedom associated with the estimate of the genetic correlation. The authors stated that the bootstrap method has high reliability and lends as an appropriate and useful procedure that can be adopted for breading purposes to test the significance of genotypic

[3]For more details, see Appendix 1 of Lynch e Walsh (1998) or see Casella e Berger (2002).

and environmental correlations.

2.2 Simulation-based tests

According to Waller et al. (2003), the basic idea of empirical tests is a very simple one and essentially operationalizes frequency-based statistical inference. Suppose one wishes to test a certain null hypothesis. One selects some test statistic denoted S and calculates its value for the observed data, say s. Under the null hypothesis, S will follow a probability distribution based on the randomness generated within the model. One determines the weight of statistical evidence against H_0 (the statistical significance of the observed value) by assessing how consistent the observed value appears to be with the distribution of the test statistic S given that the null hypothesis is true. Thus, the p-value represents the probability under the null hypothesis that the test statistic S (a random variable) exceeds the observed value s. A Monte Carlo test is simply a computational implementation of this concept. One generates a large number, say $nsim$, of independent realizations from the model and calculates the observed value of S for each realization, denoted $s_i, i = 1, ..., nsim$. A histogram of the values associated with the simulated data sets $(s_1, ..., s_{nsim})$ provides an estimate of the probability density of the test statistic under the null hypothesis.

Waller et al. (2003) used Monte Carlo simulations for accessing the goodness-of-fit for ecologic simulation models. The authors concluded that Monte Carlo methods offer an approach to draw statistical inference beyond just mean (average) behavior from ecological simulation models, particularly when realizations of such models violate many traditional statistical assumptions (e.g. independence). In order to evaluate equality of covariance matrices, Goodnight e Schwartz (1997) used bootstrap methods. Viana et al. (2000) tested the association of *Trichocereus pasacana (Cactaceae)* with potential nurse plants via data randomization procedures.

Monte Carlo testing is similar 'in spirit' to permutation tests (FISHER, 1935) and nonparametric bootstrap hypothesis tests (EFRON; TIBISHIRANI, 1993). However, permutation tests, their randomized counterparts, and bootstrap tests typically involve resampling the observed data in some manner, while Monte Carlo tests involve the generation of *new* data under the null hypothesis. Parametric bootstrap methods are based on the same concept as Monte Carlo tests. According to Efron e Tibishirani (1993), the difference between Monte Carlo and bootstrapped p-values will often be small.

An example of a *simulation-based* test is the Fisher-Pitman permutation test (PITMAN, 1937), which can be an alternative to the F-test when analyzing differences among independent samples with unequal variances (BOIK, 1987; BERRY et al., 2002). Another example is Mantel (1967) test, which consists of a permutation test used to verify a linear relation between distance matrices. Manly et al. (1986) proposed a randomization procedure for comparing the means of two or more groups with respect to several

variables.

According to Good (2006), the power of permutation tests is quite high and for small samples, a permutation test may be the most powerful test available. On the other hand, for very large samples, a permutation test will be as powerful as the most powerful parametric test. Manly (2007) states that when data are obtained from non-standard distributions, there is some evidence to suggest that randomization tests have more power than classical (parametric) tests.

2.3 Bootstrap

The bootstrap technique was first systematically considered by Efron (1979) as a basic computational method to estimate the standard error of an estimator $\hat{\theta}$ based on random data sample, whose probability distribution is known (EFRON; TIBISHIRANI, 1993), although the general idea of the method has been used before. The authors comment that the bootstrap estimate for the standard error does not require any theory calculation, however complicated can be $\hat{\theta}$ from a mathematical point of view. According to Goodnight e Schwartz (1997), the power of the bootstrap approach lies in its adaptability to complex designs, in its freedom from assumptions (usually balance and normality).

The essence of bootstrap resampling is the idea that, in the absence of any knowledge about the population, the distribution of values in a random sample of size n from this population is the best 'guide' for the distribution of the population values. In other words, the n observed sample values, each with probability $1/n$, are used for modelling the population. The sampling is done with replacement, and it is the only practical difference between the bootstrap resampling and randomization in many applications (MANLY, 2007).

2.3.1 Data generation process

The procedure for generating the bootstrap samples is called *bootstrap data generating process*, or simply *bootstrap DGP*. Some bootstrap DGPs may be fully parametric, others may be fully nonparametric, and still others may be partly parametric. What choices are available depend on the model being estimated and on the assumptions that the investigator is willing to make.

To describe the nonparametric bootstrap DGP, considere $\hat{F}$ the empirical distribution function of the observed data vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, with probability $1/n$ for each x_i. Now consider the statistic $\hat{\theta} = f(\mathbf{x})$ of interest. A *bootstrap sample* is defined as a random sample of size n obtained from $\hat{F}$, denote $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$, i.e., a *randomized with replacement* version of $\mathbf{x}$. Corresponding to this bootstrap sample, there is an estimator $\hat{\theta}^* = f(\mathbf{x}^*)$.

According to Efron e Tibishirani (1993), the algorithm used to obtain the bootstrap

standard error consists of:

1. Obtaining B independent bootstrap samples, $\mathbf{x}^{*1}, \mathbf{x}^{*2}, ..., \mathbf{x}^{*B}$, each one of size n.

2. Evaluating the statistic of interest for each bootstrap sample

$$\hat{\theta}^*(b) = f(\mathbf{x}^{*b}), \qquad b = 1, 2, ..., B. \tag{13}$$

3. Estimating the standard error $se_F(\hat{\theta})$ as the standard deviation from the B bootstrap estimates

$$\hat{se}_{boot}(\hat{\theta}) = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \bar{\hat{\theta}}^* \right]^2 \right)^{\frac{1}{2}} \tag{14}$$

where $\bar{\hat{\theta}}^* = \dfrac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$.

The reason why it is called *nonparametric bootstrap* is due to the fact that estimates are based on $\hat{F}$, an empirical distribution function, i.e., not derived from a parametric model for the data.

The number B of resampling usually vary from 25 to 200, according to Efron e Tibishirani (1993). On the other hand, Krzanowski (2000) stated that common values for B used in practice range between 100 and 1,000, depending on the complexity of the estimator being studied. Nonetheless, Manly (2007) presents some calculation methods and discussions on how many randomizations are needed considering the significance level of the test.

Regarding to the quality of the bootstrap estimator, Casella e Berger (2002) comment that in many cases the technique offers a reasonable estimator, which is consistent. Precisely, they establish

$$\hat{se}_{boot}(\hat{\theta}) \overset{B \to \infty}{\to} se_F(\hat{\theta}). \tag{15}$$

2.3.2 Bootstrap inference

The bootstrap resampling method in its simplest form has been used to compute confidence intervals for population quantities or to build significance tests. Two methods of obtaining bootstrap confidence intervals are widely used. The first is based on percentiles, where the limits are calculated directly from the empirical distribution. It turns out that the distribution of these estimates is often asymmetric, as it is based on a single sample, making the confidence interval biased. There is, however, an alternative method, *the bias-corrected and accelerated*.

Another method to build bootstrap confidence intervals consists of using quantiles from the standard normal distribution to modify the lower and upper limits of the interval

(QUINN; KEOUGH, 2010). Under most circumstances, it is seen that for a large n, the distribution of $\hat{\theta}$ becomes more and more normal, with mean θ and standard error $\hat{se}_{boot}$ (MANLY, 2007). Thus, the following probability statement is valid:

$$P_\theta(\hat{\theta} - z_{\alpha/2}\hat{se}_{boot} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\hat{se}_{boot}) = 1 - \alpha,$$

where $z_{\alpha/2}$ is the percentile $100(1-\alpha/2)\%$ from the standard normal distribution. Hence, a $100(1-\alpha)\%$ boostrap confidence interval for θ can be obtained by

$$\hat{\theta} \pm z_{\alpha/2}\hat{se}_{boot}. \tag{16}$$

This result can also be used to build significance tests for θ, since the following relationship is verified

$$\frac{\hat{\theta} - \theta}{\hat{se}_{boot}} \xrightarrow{d} N(0,1). \tag{17}$$

2.4 Monte Carlo simulation

According to Manly (2007), a Monte Carlo significance test for an observed test statistic is accessed by comparing it to a vector of sample test statistics obtained by generating random samples from some assumed model. If the assumed model implies that all sorted data are equally likely, then it works like a randomization test with random sampling from the randomization distribution (empirical distribution). Thus, the bootstrap resampling technique can be thought of a Monte Carlo method applied in a particular way.

Henderson (1985) stated that for breeding purposes, some genetic evaluation methods are less vulnerable to selection effects than others, what makes necessary to make a comparison between them. The author suggests data simulation as an alternative, specially when the involving field data require difficult mathematical manipulations.

2.4.1 Generating random numbers

Monte Carlo methods are heavily dependent on how quickly and efficiently is the production of random numbers (LANDAU; BINDER, 2009). According to Caflisch (1998) and Gamerman e Lopes (2006), most of the methods used to generate *pseudo*-random numbers are linear congruential methods.

Random variables used in the Monte Carlo methods are generated by a *pseudo*-random number generator. This is because a computer program itself is not random. A pseudo random number generator simulates randomness without actually being random. However, the pseudo random numbers are generated so that they have many of the random number sequence properties (NIEDERREITER, 1992). Gamerman e Lopes (2006)

highlights these important basic properties: uniformity and independence.

The pseudo-random number generators produce uniformly distributed variables. Not uniform variables can be generated by transforming uniform variables using, for example, the *Probability Integral Transformation* method[4]. For a summary of this method, consider generating values of X, a continuous random variable with F_X distribution function and one defines the random variable $U = F_X(X)$. Hence, U is uniformly distributed in the interval [0, 1], that is $P(U \leq u) = u$. If F_X is strictly increasing, then its inverse, F_X^{-1}, is defined by

$$F_X^{-1}(u) = x \quad \Leftrightarrow \quad F_X(x) = u. \tag{18}$$

This way, generating $U_1, U_2, ..., U_n$ independent random variables from the standard uniform distribution, we get $X_i = F_X^{-1}(U_i)$, for $i = 1, 2, ..., n$ independent and identically distributed (i.i.d.) random variables with density f_X.

This is a very convenient method, but not necessarily easy to implement because of the difficulties that can arise when computing the inverse F_X^{-1}. This would be the case, for example, when generating χ_1^2 variables. Some difficulty is also found in obtaining the inverse of $N(0, 1)$ distribution, being necessary to compute the *error function*[5]. For the latter distribution as well as for some other probability distributions, special transformations are an useful alternative to the probability integral transformation. The simplest among these methods is the *Box-Muller algorithm* (BOX; MULLER, 1958).

2.4.2 Simulating multivariate normal

To generate a p-dimensional normal variable, we start with the generation of p independent unidimensional variables. For this, one can use the *Box-Muller algorithm*, which in one process generates two independent standard normal variables, say Z_1 and Z_2, from two independent uniform variables, say U_1 and U_2. The algorithm formulas are:

$$\begin{aligned} R &= \sqrt{-2\ln(U_1)} \\ \Theta &= 2\pi U_2 \\ Z_1 &= R\cos(\Theta) \\ Z_2 &= R\sin(\Theta) \end{aligned} \tag{19}$$

Thus, it is possible to generate several pairs of independent standard normal variables by generating the same number of pairs of independent standard uniform variables.

[4]The proof of the *Probability Integral Transformation Theorem* can be found at Casella e Berger (2002).

[5]For further details, see Caflisch (1998).

The generation of p-dimensional normal with any *positive definite*[6] scale matrix can be done using the *Cholesky decomposition*[7], as follows: consider simulating $\mathbf{X} \in \Re^p$ a random variable with $N_p(\mathbf{0}, \mathbf{\Sigma})$ distribution. Via Cholesky decomposition we get $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is lower triangular. Note that $\mathbf{L}$ does exist because $\mathbf{\Sigma}$ is symmetric and positive definite. Now, consider $\mathbf{Z} \in \Re^p$ a random vector consisting of p independent standard normal variables. Then, the covariance matrix of $\mathbf{Z}$ is $\mathbf{I}$ (the identity). Therefore,

$$\mathbf{X} = \mathbf{L}\mathbf{Z}. \tag{20}$$

Thus, $\mathbf{X}$ is a p-dimensional normal vector with covariance matrix

$$Cov(\mathbf{X}) = \mathbf{L}\mathbf{I}\mathbf{L}' = \mathbf{\Sigma}.$$

To get a normally distributed random vector with mean $\boldsymbol{\mu}$, simply take $\mathbf{X} = \mathbf{L}\mathbf{Z} + \boldsymbol{\mu}$.

The simulation of multidimensional normal variables with a particular structure $\mathbf{\Sigma}$ of variances and covariances is of special interest for generating *Wishart* matrices, especially when studying the distribution of variances-covariance sample matrices.

2.5 Wishart distribution

Wishart matrices occupy a central place in the development of multivariate theory and applications (GHOSH; SINHA, 2002). According to Gauthier e Possamai (2009), in financial mathematics Wishart processes have emerged as an efficient tool to model stochastic covariance structures. Gelman et al. (2004) stated that an inverse-Wishart distribution is used often in Bayesian modeling because it is a proper conjugate prior for an unknown covariance matrix in a multivariate normal model. Vester e Waal (2015) used Wishart distribution for modeling the volatility of a rainfall indicator.

2.5.1 Derivation

In the univariate theory, if $x_1, x_2, ..., x_n$ are a sequence of independent and identically distributed (i.i.d.) random variables, each having $N(0, \sigma^2)$ distribution, then $\frac{1}{\sigma^2}\sum x_i^2$ has a chi-squared distribution on n degrees of freedom, denoted χ_n^2. The multivariate case occurs when $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ form a sequence of independent p-variate random vectors, each with distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$. From these vectors a symmetric matrix $\mathbf{C}$ of dimension p can be defined by $\mathbf{C} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}'_i$, whose main diagonal contains sum of squares, whereas the off-diagonal elements correspond to sum of cross-products. The joint distribution of

[6] A symmetric matrix, say $\mathbf{A}$, is called positive definite if all its eigenvalues are positive. Denote $\mathbf{A} > 0$.

[7] Cholesky decomposition of factoration receives the name of its creator, André-Louis Cholesky, and is often used to solve the normal equations in problems of least square or in Monte Carlo methods to simulate systems with multiple correlated variables. A variation of Cholesky decomposition is the form $\mathbf{A} = \mathbf{R}'\mathbf{R}$, where $\mathbf{R}$ is upper triangular (the *Cholesky triangle*). For more details, see Golub e Loan (1996).

all elements of $\mathbf{C}$ (or simply the distribution of $\mathbf{C}$) was obtained by Wishart (1928) and is called p-variate *Wishart distribution* with n degrees of freedom and scale parameter $\mathbf{\Sigma}$, denoted $W_p(n, \mathbf{\Sigma})$ (KRZANOWSKI, 2000). If $p > n$ then $W_p(n, \mathbf{\Sigma})$ is called *singular Wishart distribution* (COOK, 2011).

Being $\mathbf{S}$ a sampling variance-covariance matrix (supposed to be positive definite), obtained from n independent random vectors with distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$, then $(n-1)\mathbf{S} \sim W_p(n-1, \mathbf{\Sigma})$. The probability mass function of $\mathbf{S}$ can be writen as follows (adapted from Anderson (2003)):

$$f_{\mathbf{S}}(\mathbf{S}|\mathbf{\Sigma}, n-1) = \frac{|\mathbf{S}|^{\frac{n-p-2}{2}} e^{-tr(\mathbf{\Sigma}^{-1}\mathbf{S})/2}}{2^{\frac{p(n-1)}{2}} \pi^{\frac{p(p-1)}{4}} |\mathbf{\Sigma}|^{\frac{n-1}{2}} \prod_{j=1}^{p} \Gamma\left(\frac{n-j}{2}\right)}, \quad \mathbf{S} > 0, \tag{21}$$

where $\Gamma(.)$ is the gamma function and $tr(.)$ the trace function.

2.5.2 Properties

Wishart distribution can be understood as a multivariate generalization of a chi-squared distribution. Indeed, many of its properties are either bound up with χ^2 or mirror those of the χ^2 distribution. Some of them, more interesting to this study, are listed here.

1. If $p = 1$, then $\mathbf{C} = \sum_{i=1}^{n} x_i^2$, where x_i are i.i.d. $N(0, \sigma^2)$. Hence $W_1(n, \sigma^2) \sim \sigma^2 \chi_n^2$.

2. If $\mathbf{C}_1$ and $\mathbf{C}_2$ are independent with $\mathbf{C}_1 \sim W_p(n_1, \mathbf{\Sigma})$ and $\mathbf{C}_2 \sim W_p(n_2, \mathbf{\Sigma})$, then $\mathbf{C}_1 + \mathbf{C}_2 \sim W_p(n_1 + n_2, \mathbf{\Sigma})$.

3. If $\mathbf{B}$ is any $(q \times p)$ matrix of constants and $\mathbf{C} \sim W_p(n, \mathbf{\Sigma})$, then $\mathbf{BCB'} \sim W_q(n, \mathbf{B\Sigma B'})$.

4. If $\mathbf{b}$ is any p-variate vector of constants and $\mathbf{C} \sim W_p(n, \mathbf{\Sigma})$, then $\mathbf{b'Cb} \sim \sigma^2 \chi_n^p$, where $\sigma^2 = \mathbf{b'\Sigma b}$.

5. If $\mathbf{C} \sim W_p(n, \mathbf{\Sigma})$ then $\mathbf{C}^{-1}$ is said to have *inverted Wishart* distribution with the same parameters, denoted by $IW_p(n, \mathbf{\Sigma})$.

   For the following properties, consider c_{ij} as the element at the i-th row and j-th column of the matrix $\mathbf{C}$ and σ_{ij} as the element at the i-th row and j-th column of the matrix $\mathbf{\Sigma}$, admitted $\mathbf{C} \sim W_p(n, \mathbf{\Sigma})$.

6. The correlation coefficient can be obtained by $r_{ij} = c_{ij}/\sqrt{c_{ii}c_{jj}}$.

7. $E(\mathbf{C}) = n\mathbf{\Sigma}$ and $Var(c_{ij}) = n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$.

8. If $\sigma_{ij} = 0$, then $r_{ij}\sqrt{\frac{n-1}{1-r_{ij}^2}}$ is distributed as t-Student on $n-1$ degrees of freedom.

9. $\frac{1}{2}\ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right)$ is asymptotically normally distributed with mean $\frac{1}{2}\ln\left(\frac{1+\rho_{ij}}{1-\rho_{ij}}\right)$ and variance $\frac{1}{n-2}$, where $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$.

For further details about these properties, see Johnson e Kotz (1972), Mardia et al. (1979), Krzanowski (2000).

2.6 Wilks' Lambda distribution

Wilks (1932) Lambda statistic is widely used for various statistical tests in multivariate analysis since it, supposedly, plays the same role as the Fisher–Snedecor F univariate statistics (PHAM-GIA, 2008).

Consider $\mathbf{C}_0 \sim W_p(n_0, \mathbf{\Sigma})$ and $\mathbf{C}_1 \sim W_p(n_1, \mathbf{\Sigma})$, two independent Wishart matrices of order p, where $n_0 \geq p$ and $n_1 \geq p$. The generalized likelihood ratio

$$\Lambda = \frac{|\mathbf{C}_0|}{|\mathbf{C}_0 + \mathbf{C}_1|} \qquad 0 \leq \Lambda \leq 1 \tag{22}$$

is called *Wilks' lambda*. Its distribution depends only on three parameters: p is the order of the matrices, n_0 is typically the residual degrees of freedom and n_1 is the hypothesis degrees of freedom. We denote the *pdf* of the Wilks' lambda distribution by $\Lambda(p, n_0, n_1)$.

The exact distribution of Wilks's statistic is difficult to track, because, so far, its density lacks a closed form expression, except for some simple values of its parameters (SCHATZOFF, 1966). According to Pham-Gia (2008) and Timm (2002), it was established that Wilks' Lambda has the same density as a product of independent univariate beta variables on $(n_0 - i + 1)/2$ and $n_1/2$ degrees of freedom for $i = 1, 2, ..., p$. For certain special cases, the Wilks' Lambda distribution reduces to an F distribution, but in general its use is quite complicated, as it is a function of three parameters (p, n_0 and n_1). Nonetheless, various approximations has been obtained and are useful in practice (KRZANOWSKI, 2000). Bartlett (1947) showed that

$$-\left(n_0 + n_1 - \frac{p + n_1 + 1}{2}\right)\log_e \Lambda(p, n_0, n_1),$$

is asymptotically a $\chi^2_{n_1 p}$. An F-approximation is given by Rao (1952).

2.7 Using Wilks' Lambda for testing independence of sets of variables

Consider a random vector $\mathbf{x}$ with k-variate normal distribution, $N_k(\boldsymbol{\mu}, \mathbf{\Sigma})$. Suppose partitioning $\mathbf{x}$ into two subvectors of interest, say $\mathbf{x}_1$ and $\mathbf{x}_2$, with resulting dimensions p and q, respectively, where $p+q = k$. The corresponding partitioning population covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

with analogous partitioning of the estimate $\mathbf{S}$:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

The hypothesis of independence of $\mathbf{x}_1$ and $\mathbf{x}_2$ can be expressed by

$$H_0 : \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

It means that every variable in $\mathbf{x}_1$ is independent of every variable in $\mathbf{x}_2$. Note that there is no restriction on $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$. The likelihood ratio test statistic (WILKS, 1935) for H_0 is given by

$$\Lambda = \frac{|\mathbf{S}|}{\left|\mathbf{S}_{(H_0)}\right|} = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}|\,|\mathbf{S}_{22}|} \qquad 0 \leq \Lambda \leq 1 \tag{23}$$

which is distributed as $\Lambda(p, q, n-1-q)$ or, equivalently, $\Lambda(q, p, n-1-p)$. Thus, Wilks' Λ compares an estimate of $\mathbf{\Sigma}$ without restriction to an estimate of $\mathbf{\Sigma}$ under $H_0 : \mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}^T = \mathbf{0}$. Then, a value of Λ near zero indicates high correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$, whereas a value near one indicates low correlation.

There are pq constraints in the specification of H_0, so if the null hypothesis is true, then an approximation to a chi-square distribution is given by

$$-n \log_e \Lambda \xrightarrow{d} \chi^2_{pq} \tag{24}$$

According to Krzanowski (2000), the χ^2 approximation can be improved on replacing n by $n' = n - \frac{1}{2}(p + q + 3)$.

An asymptotically equivalent test statistic is presented by the Pillai (1955):

$$T_n = tr\left(\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\right) \tag{25}$$

Under H_0, $nT_n \xrightarrow{d} \chi^2_{pq}$.

2.8 Distribution analyses: density estimation

In many situations, the detailed shape of the underlying density function, say f, is of primary interest. In genetics for example, empirical distributions are frequently evaluated. Brown (1969) studied the empirical distribution of the sample genetic correlation coefficient using histograms. Through kernel densities, Schork (2002) estimated the distribution of marker-allele frequencies. Segal e Wiemels (2002) used Gaussian kernel estimation for studying Translocation, a physical movement of genetic material from one

chromosome to another. In order to analyze empirical distributions of genetic distances obtained from simulation, Worby et al. (2014) used Gaussian kernel densities. Through histograms, Blows e Mcguigan (2014) analysed the empirical distribution of eigenvalues of the **G** matrix obtained from simulated data.

Deng e Wickham (2011) state that density estimation builds an estimate of f using an observed data sample, supposed to be independent and identically distributed. Density estimation can either be parametric, where the data are obtained from a known probability family, or nonparametric, which attempts to flexibly estimate an unknown distribution.

The histogram is, of course, a widely used tool for displaying the distribution shape of a set of data, since it indicates the shape of f. However, viewed as a density estimate, the histogram may be criticized in three ways (BOWMAN; AZZALINI, 1997):

1. Information has been thrown away in replacing the observations by the central points of the interval in which it falls.

2. In most circumstances, f is assumed to be smooth, but its estimator is not, due to the sharp edges of the boxes from which it is built.

3. The behaviour of the estimator is dependent on the choice of width of the intervals (or equivalently boxes) used, and also, to some extent on the starting position of the grid of intervals.

An approach that removes the first two of these problems is *kernel density estimation*, at which a smooth function is used as the basic building block and also, these smooth functions are centered directly over each observation.

Let $x_1, x_2, ..., x_n$ denote a sample of size n from a random variable X with density f. The basic kernel estimator of f at the point x is given by:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x - x_i|h), \tag{26}$$

where k is a function that satisfies the conditions

$$\int k(x)dx = 1,$$

$$\int xk(x)dx = 0,$$

$$\int x^2 k(x)dx = \mu_2'(k) > 0.$$

In this context, k is called *kernel* function, whose variance is controlled by the parameter h. Hence, the kernel is itself a probability density. The most common kernel functions are described in Table 2. And Figure 1 illustrates a Gaussian kernel function.

Because of its role in determining the manner in which the probability associated with each observation spread over the surrounding sample space, h is called the *smoothing*

Table 2 – Six common kernel functions

Kernel	Formula	Support interval	$\mu_2'(k)$
Gaussian	$k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$	$x \in \Re$	1
Rectangular	$k(x) = \frac{1}{2}$	$\lvert x\rvert < 1$	1/3
Triangular	$k(x) = 1 - \lvert x\rvert$	$\lvert x\rvert < 1$	1/6
Epanechnikov	$k(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}x^2\right)$	$\lvert x\rvert < \sqrt{5}$	1
Biweight	$k(x) = \frac{15}{16}\left(1 - x^2\right)^2$	$\lvert x\rvert < 1$	1/7
Cosine	$k(x) = 1 + \cos(2\pi x)$	$\lvert x\rvert < 1/2$	≈ 0.0326

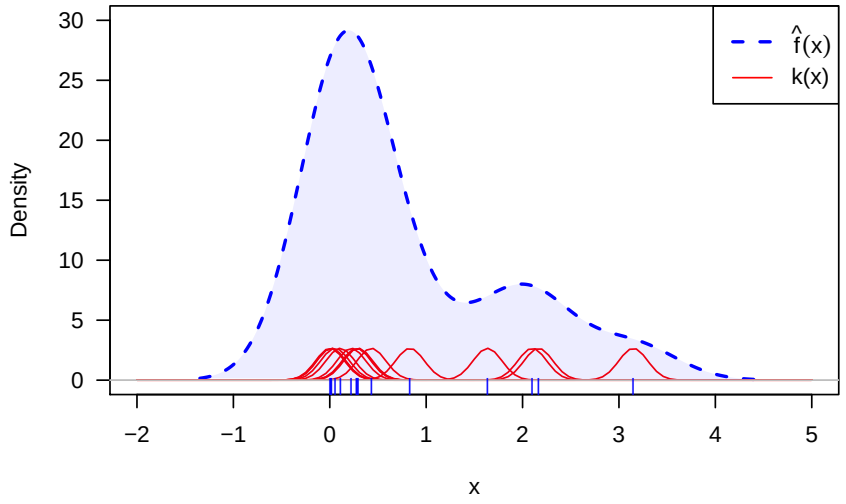

Figure 1 – Illustrating a Gaussian kernel function

parameter or *bandwidth* (BOWMAN; AZZALINI, 1997). Kernels can be scaled so that bandwidth is the standard deviation of the kernel. The choice of h affects directly the behaviour of the density estimate (BOWMAN; AZZALINI, 1997; SHEATHER, 2004). The smaller is h the more winding is the density estimates. On the other hand, large bandwidths produce very smooth estimates. In addition, most researchers agree that the choice of kernel is not as important as the choice of h. Taylor (1989) presents a bootstrap choice of h.

Some other methods of density estimation are: *variable bandwidths*, the *nearest neighbour methods*, the *ortogonal series methods*, *local likelihood and semiparametric density*. Further details on density estimation can be found in Wand e Jones (1995), Simonoff (1996), Bowman e Azzalini (1997).

3 MATERIALS AND METHODS

3.1 Experimental data

3.1.1 Genetic material

The genetic material was obtained from the crossing of the maize lines L20-01F and L02-03D, arising from populations developed by the maize breeding program of the Genetics Department of ESALQ/USP. The L20-01F is derived from the IG-1 population, which has orange hard grains, whereas L02-03D is derived from the IG-2 population, which has yellow toothed grains. These populations have tropical origin, early cycle, short stature and belong to different heterotic groups, contrasting on several agronomic traits (MANGOLIN et al., 2004).

From the crossing of L20-01F and L02-03D was obtained the F1 generation, then being selfed three plants of this generation, resulting in the F2 population. Approximately 500 F2 seeds were sown and the plants subsequently selfed to yield 256 F2:3 progenies with well formed ears. For obtaining the number of seeds required for the evaluations, the F2:3 progenies were grown in rows with 60 plants and crossed among plants were performed on each progeny. Each plant was used only once as male or female.

3.1.2 Description of the experiments

Experiments were installed under a incomplete block design, more precisely square lattice of dimension 16. Plots were 4.0 m long and spaced 0.8 m between rows, and 0.20 m between plants within rows; and they were overplanted and thinned to 20 plants per plot (62,500 plants ha^{-1}). The experiments were carried out at 12 environments, each one corresponding to the combination $local \times year$, namely:

1. Estação Experimental Departamento de Genética da ESALQ/USP (E. E. Depto. Genética), years 2002/2003.
2. Estação Experimental Fazenda Caterpillar (E. E. Caterpillar), years 2002/2003.
3. Estação Experimental Fazenda Areão (E. E. Areão), years 2002/2003.
4. Estação Experimental Fazenda Anhembi (E. E. Anhembi), years 2002/2003.
5. Estação Experimental Departamento de Genética da ESALQ/USP (E. E. Depto. Genética), years 2003/2004.
6. Estação Experimental Fazenda Caterpillar (E. E. Caterpillar), years 2003/2004.
7. Estação Experimental Fazenda Anhembi (E. E. Anhembi), years 2003/2004.
8. Estações Experimentais ESALQ/Anhembi, years 2004/2005.

9. Estação Experimental Departamento de Genética da ESALQ/USP (E. E. Depto. Genética), years 2005/2006.
10. Estação Experimental Fazenda Anhembi (E. E. Anhembi), years 2005/2006.
11. Estação Experimental Departamento de Genética da ESALQ/USP (E. E. Depto. Genética), years 2006/2007.
12. Estação Experimental Fazenda Anhembi (E. E. Anhembi), years 2006/2007.

3.1.3 Response variables

The following variables were measured:

1. (STD) Stand, corresponding to the number of plants per plot;
2. (MO) Grains moisture content, in %, obtained from a sample of grains from each plot, using an electronic determinator Dickey-John;
3. (NE) Number of ears per plot;
4. (W500) Weight of 500 grains, in g, measured after threshing on each plot;
5. (LE) Average length of ears, in cm;
6. (DE) Average diameter of ears, in cm;
7. (DC) Average diameter of cobs, in cm;
8. (NROWS) Average number of rows per ear;
9. (NGROW) Average number of grains per row;
10. (YP) Grains yield per plot, in g;
11. (PH) Average height of plants per plot, in m
12. (EH) Average height of ears, in cm;
13. (MF) Number of days from the sowing until 50% of plants in a plot to present anthesis;
14. (FF) Number of days from sowing until 50% of plants to present visible style-stigma.

For statistical analysis, the variable YP was corrected to the standard moisture of 15.5% and converted to g plant^{-1} by dividing the corresponding value in g plot^{-1} by the plot stand. Thus, STD and MO will be used only to correct YP. For the analysis of YP, all plants in a plot were considered, while LE, DE, DC, NROWS and NGROW were obtained

from a sample of five ears from each plot, being the sampling arithmetic mean used in the analyses. Samples were taken from the five best formed ears of each plot. Hence, there are actually twelve response variables, but in this study, only for simplicity, the following variables were considered: DE, DC, PH, and NGROW. The analysis involving all variables can be found in the Appendix.

3.1.4 Univariate analysis of variance

For each variable, analyses of variance (ANOVA) were performed per environment, according to the following random linear model

$$y_{ijk} = \mu + \alpha_i + \xi_j + \beta_{k(j)} + \epsilon_{ijk} \tag{27}$$

where
y_{ijk} is the observation taken from the i-th progeny on the k-th block in the j-th replication;
μ is the population mean of the variable Y;
α_i is the random effect of the i-th progeny ($i = 1, 2, ..., 256$);
ξ_j is the random effect of the j-th replication ($j = 1, 2$);
$\beta_{k(j)}$ is the random effect of the k-th block ($k = 1, 2, ..., 16$) in the j-th replication;
ϵ_{ijk} is the random error associated with y_{ijk}.

These analyses were used only to calculate adjusted (*Least Square*) means of progenies for the effect of blocks, as presented by Cochran e Cox (1966).

3.1.5 Multivariate analysis of variance

Joint (for environments) analysis of variance was performed through multivariate analysis of variance (MANOVA), using the adjusted means of progenies for the effect of blocks, as a randomized block design model, expressed by

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\lambda}_j + (\boldsymbol{\alpha\lambda})_{ij} \tag{28}$$

where
$\mathbf{y}_{ij}$ is the p-dimensional vector of adjusted means of the i-th progeny on the j-th environment;
$\boldsymbol{\mu}$ is the p-dimensional vector of population means of the response variables;
$\boldsymbol{\alpha}_i$ is the p-dimensional vector of random effects of the i-th progeny ($i = 1, 2, ..., 256$);
$\boldsymbol{\lambda}_j$ is the p-dimensional vector of random effects of the j-th environment ($j = 1, 2, ..., 12$);
$(\boldsymbol{\alpha\lambda})_{ij}$ is the p-dimensional vector of random effects of the interaction between the i-th progeny and the j-th environment.

By means of MANOVA the matrices of *mean squares and cross-products* of environment, $\mathbf{M}_E$, progenies, $\mathbf{M}_G$, and interaction progenies $\times$ environments, $\mathbf{M}_{GE}$ were

obtained. These two latter were used to calculate the statistic of interest

$$\frac{\mathbf{M}_{G(ij)}}{\mathbf{M}_{GE(ij)}}, \quad \forall i \neq j, \quad i, j = 1, 2, ..., p. \tag{29}$$

where p is the number of response variables, $\mathbf{M}_{G(ij)}$ and $\mathbf{M}_{GE(ij)}$ are the mean cross-products of progenies and interaction progenies $\times$ environments, respectively, for the variables located at the i-th row and j-th column.

It is stressed that the mean cross-products of interaction progenies $\times$ environments were used as denominator for the test statistic, by analogy of the corresponding F-test, the ratio of mean squares, since it is found that the expected value of mean squares and cross-products are as shown in Table 3.

Table 3 – Scheme of MANOVA for joint analysis

Source	D.f.	$\mathbf{M}$	E($\mathbf{M}$)
Environments (E)	$e-1$	$\mathbf{M}_E$	
Progenies (G)	$g-1$	$\mathbf{M}_G$	$\mathbf{\Sigma}_{GE} + e\mathbf{\Sigma}_G$
G $\times$ E	$(g-1)(e-1)$	$\mathbf{M}_{GE}$	$\mathbf{\Sigma}_{GE}$

The estimation of the genetic covariance matrix, $\mathbf{\Sigma}_G$, was performed using the method of moments. Then,

$$\mathbf{G} = \hat{\mathbf{\Sigma}}_G = \frac{\mathbf{M}_G - \mathbf{M}_{GE}}{e}. \tag{30}$$

Note that the estimation procedure used here is the standard one, i.e., based on the assumption that the genetic variance (computed from the genotypic means) is constant over environments. For more details about the estimation procedures, see Yamada (1962) and Dutilleul e Carriere (1998).

3.2 Bootstrapping

The vectors $\mathbf{y}_{i,j}$ of adjusted means were arranged in *arrays*, as shown in Table 4.

The bootstrap process consisted of resampling rows of Table 4, with replacement, obtaining a new randomized array with the same dimensions ($e \times p \times g$). For illustrating the scheme of bootstrap resampling, consider a hypothetical array of $e = 4$ environments (rows), $p = 2$ variables (columns) and $g = 3$ progenies (matrices), as shown in Figure 2.

The third dimension of the array is being bootstrapped. In Figure 2 only three bootstrap samples were generated. For the experimental data, a $12 \times 4 \times 256$ array, the number B of bootstrap samples was equal to 400.

After computing the bootstrapped arrays, the matrices $\mathbf{M}_G^*$ and $\mathbf{M}_{GE}^*$ (asterisc indicates bootstrapped matrices) were determined via MANOVA, as described early. Then,

Table 4 – Three dimensional array for bootstrap resampling

Progeny	Environment			
	1	2	$\cdots$	e
1	$\mathbf{y}_{1,1}$	$\mathbf{y}_{1,2}$	$\cdots$	$\mathbf{y}_{1,e}$
2	$\mathbf{y}_{2,1}$	$\mathbf{y}_{2,2}$	$\cdots$	$\mathbf{y}_{2,e}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
g	$\mathbf{y}_{g,1}$	$\mathbf{y}_{g,2}$	$\cdots$	$\mathbf{y}_{g,e}$

```
Original......   Bootstrapped..............................
, , 1            , , 1           , , 2           , , 3

     [,1] [,2]        [,1] [,2]       [,1] [,2]       [,1] [,2]
[1,] y111 y121   [1,] y111 y121  [1,] y112 y122  [1,] y113 y123
[2,] y211 y221   [2,] y211 y221  [2,] y212 y222  [2,] y213 y223
[3,] y311 y321   [3,] y311 y321  [3,] y312 y322  [3,] y313 y323
[4,] y411 y421   [4,] y411 y421  [4,] y412 y422  [4,] y413 y423

, , 2            , , 1           , , 1           , , 2

     [,1] [,2]        [,1] [,2]       [,1] [,2]       [,1] [,2]
[1,] y112 y122   [1,] y111 y121  [1,] y111 y121  [1,] y112 y122
[2,] y212 y222   [2,] y211 y221  [2,] y211 y221  [2,] y212 y222
[3,] y312 y322   [3,] y311 y321  [3,] y311 y321  [3,] y312 y322
[4,] y412 y422   [4,] y411 y421  [4,] y411 y421  [4,] y412 y422

, , 3            , , 3           , , 2           , , 1

     [,1] [,2]        [,1] [,2]       [,1] [,2]       [,1] [,2]
[1,] y113 y123   [1,] y113 y123  [1,] y112 y122  [1,] y111 y121
[2,] y213 y223   [2,] y213 y223  [2,] y212 y222  [2,] y211 y221
[3,] y313 y323   [3,] y313 y323  [3,] y312 y322  [3,] y311 y321
[4,] y413 y423   [4,] y413 y423  [4,] y412 y422  [4,] y411 y421
```

Figure 2 – Illustrating the scheme of bootstrap resampling on the third dimension of an $4 \times 2 \times 3$ array

ratios of mean cross-products $\mathbf{M}^*_{G(ij)}/\mathbf{M}^*_{GE(ij)}$, $\forall i \neq j$ $(i, j = 1, 2, ..., p.)$, were calculated, obtaining 400 estimates of each test statistics of interest (eq. 29), whose empirical distribution was built.

3.3 Simulating Wishart matrices

Under the assumptions of the MANOVA model, eq. 28, for the random effects it was considered that

$$\begin{aligned} \boldsymbol{\alpha}_i &\sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_G), \\ \boldsymbol{\alpha\lambda}_{ij} &\sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_{GE}). \end{aligned}$$

where $i = 1, 2, ..., 256$ and $j = 1, 2, ..., 12$. Thus, matrices of mean squares and cross-products, $\mathbf{M}_G$ and $\mathbf{M}_{GE}$, obtained via MANOVA for the p-dimension vectors of adjusted means, as presented in Table 3, were considered to be positive definite having a p-dimensional Wishart distribution, i.e.,

$$\begin{aligned} (g-1)\mathbf{M}_G &\sim W_p(g-1, \boldsymbol{\Sigma}_{GE} + a\boldsymbol{\Sigma}_G), \\ (g-1)(e-1)\mathbf{M}_{GE} &\sim W_p((g-1)(e-1), \boldsymbol{\Sigma}_{GE}). \end{aligned} \tag{31}$$

Now, taking into account the null hypothesis (H_0):

$$H_0 : \boldsymbol{\Sigma}_G = \mathbf{0}, \tag{32}$$

then

$$(g-1)\mathbf{M}_G^{(H_0)} \sim W_p(g-1, \boldsymbol{\Sigma}_{GE}). \tag{33}$$

Under these conditions, 10,000 Monte Carlo simulations of $\mathbf{M}_G^{(H_0)}$ and $\mathbf{M}_{GE}$ were performed. Considering $(g-1)\mathbf{M}_{G(ij)}^{(k,H_0)}$ the k-th Monte Carlo result for $\mathbf{M}_{G(ij)}$ under H_0, since $E[W_p(n, \boldsymbol{\Sigma})] = n\boldsymbol{\Sigma}$ (property 7, subsection 2.5.2), the following test statistic was computed

$$\frac{(g-1)\mathbf{M}_{G(ij)}^{(k,H_0)}}{(g-1)(e-1)\mathbf{M}_{GE(ij)}^{(k)}} \times \frac{(g-1)(e-1)}{g-1}, \quad \forall i \neq j, \tag{34}$$

where $i, j = 1, 2, ..., p$.

Thus, $k = 1, 2, ..., 10,000$ estimates of each test statistics of interest (eq. 29) were obtained, whose empirical distribution was built.

3.4 Empirical distribution

The empirical distribution of the statistic defined in eq. 29 was evaluated by computing the empirical cumulative distribution function and the kernel density.

The empirical cumulative distribution function (ECDF), denoted by $\hat{F}$, consists of a step function with jumps a/n at observed values, where a is the number of tied observations. Considering the vector of observations $\mathbf{x} = (x_1, x_2, ..., x_n)$ from the random variable X, $\hat{F}(x)$ is the proportion of observations less than or equal to a given value x (GOOD, 2006), i.e.,

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x), \tag{35}$$

where $I(\cdot)$ is an indicator variable.

The kernel density was estimated according to the eq. 26, based on the Gaussian kernel (Table 2).

3.5 p-value

Let T_{ij} the random variable represented by the test statistic defined in eq. 29 for the variables located at the i-th row and j-th column of the mean squares and cross-products matrices. Now consider t_{ij} the value assumed by this variable for a certain data set. Through the empirical distribution obtained via Bootstrap or Monte Carlo simulation, the two-sided p-value was calculated as follows:

$$p = \hat{F}(-|t_{ij}|) + [1 - \hat{F}(|t_{ij}|)], \tag{36}$$

Then, the following hypotheses were evaluated:

$$H_0 : \sigma_{G(ij)} = 0 \quad vs. \quad H_1 : \sigma_{G(ij)} \neq 0 \tag{37}$$

where $\sigma_{G(ij)}$ is the genetic covariance between the i-th and j-th variables.

And, of course, $\forall\ i = j$,

$$p = 1 - \hat{F}(t_{ij}), \tag{38}$$

which corresponds to the (simulated) p-value of the F-test for testing $H_0 : \sigma_{G(ij)} = 0$.

Because the empirical distribution may not be symmetric around the origin, before calculating two-sided p-values the median of the distribution was subtracted from the the simulated and observed values t_{ij}. That is equivalent to compute right tail p-values from the distribution of the absolute values.

3.6 Partial Wilks' Lambda

Consider the estimate of the genetic covariance matrix $\mathbf{\Sigma}_G$, $\mathbf{G}$ (eq. 30), with dimension $p \times p$.

$$\mathbf{G} = \begin{pmatrix} \hat{\sigma}_{G(11)} & \hat{\sigma}_{G(12)} & \cdots & \hat{\sigma}_{G(1p)} \\ \hat{\sigma}_{G(21)} & \hat{\sigma}_{G(22)} & \cdots & \hat{\sigma}_{G(2p)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{G(p1)} & \hat{\sigma}_{G(p2)} & \cdots & \hat{\sigma}_{G(pp)} \end{pmatrix}$$

Now consider a submatrix of dimension 2×2 from $\mathbf{G}$, for any two variables, say i-th and j-th $(i, j = 1, 2, ..., p)$.

$$\mathbf{G}_{ij} = \begin{pmatrix} \hat{\sigma}_{G(ii)} & \hat{\sigma}_{G(ij)} \\ \\ \hat{\sigma}_{G(ij)} & \hat{\sigma}_{G(jj)} \end{pmatrix}$$

The null hypothesis $H_0 : \sigma_{G(ij)} = 0$ related to the genetic covariance of the i-th and j-th variable can be evaluated by applying Wilks' Lambda (eq. 23) approach for testing independence of two sets of variables. In this case, $p = q = 1$ and because there are $g - 1$ independent observations, $n = g - 1$. Then, under H_0

$$\mathbf{\Sigma}_{G(ij)} = \begin{pmatrix} \sigma_{G(ii)} & 0 \\ \\ 0 & \sigma_{G(jj)} \end{pmatrix}.$$

Thus, the likelihood ratio test statistic is given by

$$\Lambda = \frac{\hat{\sigma}_{G(ii)}\hat{\sigma}_{G(jj)} - \hat{\sigma}^2_{G(ij)}}{\hat{\sigma}_{G(ii)}\hat{\sigma}_{G(jj)}} \tag{39}$$

which is distributed as $\Lambda(1, 1, g - 3)$, being g the number of progenies. There is only one constraint in the specification of H_0, $\sigma_{G(ij)} = 0$, so if the null hypothesis is true, then the approximation to a chi-square distribution is given by

$$-(g - 1) \log_e \Lambda \xrightarrow{d} \chi^2_1 \tag{40}$$

Note that, since it is considered only part of the original matrix $\mathbf{G}$, this procedure is called here as *Partial* or *Pairwise* Wilks' Lambda. The equivalent Pillai test statistic (as defined in eq. 25) is:

$$T_n = \frac{\hat{\sigma}^2_{G(ij)}}{\hat{\sigma}_{G(ii)}\hat{\sigma}_{G(jj)}} \tag{41}$$

whose chi-square approximation is also given by $(g - 1)T_n \xrightarrow{d} \chi^2_1$.

It is noteworthy that $\Lambda = 1 - \hat{\rho}^2_{G(ij)}$ and $T_n = \hat{\rho}^2_{G(ij)}$. Then, because these procedures proposed here are invariant to scaling, both allow one to test either the genetic covariance or correlation. Hence, comparisons with the exact p-values were done in order to evaluate the consistency of the inferences made with the approximate tests. The exact p-value was calculated from the exact correlation probability density function (PDF) under the null hypothesis $H_0 : \rho_{G(ij)} = 0, \ \forall i \neq j$. According to Weisstein (2015), the PDF for the correlation coefficient (r) of two Gaussian variables is given by:

$$f(r|\nu) = \frac{(1 - r^2)^{\frac{\nu-2}{2}}}{\mathbf{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} I_{[-1,1]}(r) \tag{42}$$

where ν is the degrees of freedom (of progenies in this case) and $\mathbf{B}(.,.)$ is the beta function:

$$\mathbf{B}(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx \tag{43}$$

Thus, since $f(r|\nu)$ is symmetric around zero, the exact two-sided p-value related to a certain genetic correlation, say $r_{G(ij)}$, is calculated by

$$p = 2\int_{|r_{G(ij)}|}^1 f(r|\nu)dr \tag{44}$$

3.7 Evaluating the sample size effect

For evaluating the sample size effect on the empirical distributions, part of the original data (the first 5 progenies and last 4 environments) was selected. Afterwards, the same methods described earlier were applied and the distributions were once again studied.

3.8 Evaluating collinearity effects

Because the empirical test based on bootstrap resampling and Wishart simulation are built element-wise, there is no reason to be concerned about collinearity effects on the matrices $\mathbf{M}_G$ and $\mathbf{M}_{GE}$. However, for evaluating Wilks' Lambda statistic, the determinant of the matrix $\mathbf{G}$ is computed. Hence, a multicollinearity diagnosis has to be done. It was based on the condition number (CN), that consists of the ratio between the largest and smallest eigenvalue of the underlying matrix. According to Montgomery e Peck (1982), the following classification can be applied:

- CN $<$ 100: weak multicollinearity
- 100 $\leq$ CN $\leq$ 1,000: moderate to severe multicollinearity
- CN $>$ 1,000: severe multicollinearity

Since the calculation of Pillai's test statistic, as defined in eq. 41, is not affected by collinearity effects, it may constitutes an alternative for overcoming collinearity problems.

3.9 Computing

All the analyses were performed using the software R version 3.1.2 (R CORE TEAM, 2014), as described in Table 5.

Those procedures not mentioned here were implemented in R language. Three functions were developed to perform the empirical test based on the approaches of Wilks' Lambda, Bootstrap and Wishart simulation. All of them were designed to receive as input an object of class `'manova'`, which contains the model matrix, the degrees of freedom and

Table 5 – Statistical procedures in R

Procedure	Function	Package	Reference
ANOVA	`aov()`	*stats*	R Core Team (2014)
LS means	`popMeans()`	*doBy*	Hojsgaard et al. (2012)
MANOVA	`manova()`	*stats*	R Core Team (2014)
Multivariate normality test	`mvShapiro.Test()`	*mvShapiroTest*	Gonzalez-Estrada (2013)
Simulation of Wishart matrices	`rWishart()`	*stats*	R Core Team (2014)
ECDF	`ecdf()`	*stats*	R Core Team (2014)
Kernel density	`density()`	*stats*	R Core Team (2014)

the matrices of sum of squares and cross-products of each source. All the codes used to execute the empirical tests and related functions are available in the appendices of this thesis.

4 RESULTS AND DISCUSSION

4.1 On the exact distribution of the mean cross-products ratio

Consider a $g \times 2$ matrix $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_X \;\; \boldsymbol{\alpha}_Y]^T$ containing g genetic effects of the variables X and Y, as defined in eq. 28. Also, consider a $ge \times 2$ matrix $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_X \;\; \boldsymbol{\epsilon}_Y]^T$ containing the random error term of the MANOVA model. Assuming

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix} \right),$$

the matrix of mean squares and cross-products $\mathbf{M}_G$ (as shown in Table 3) can be obtained by

$$\mathbf{M}_G = \frac{e}{g-1}(\hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}}). \tag{45}$$

And the residual matrix could be obtained by

$$\mathbf{M}_{GE} = \frac{1}{(g-1)(e-1)}(\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}). \tag{46}$$

Thus, the ratio of the off-diagonal elements from $\mathbf{M}_G$ and $\mathbf{M}_{GE}$ corresponds to the statistic of interest for measuring the significance of the genetic covariance component of $\mathbf{M}_G$, say $\sigma_{G(XY)}$. But what distribution does this ratio have? Note that the ratios involving diagonal elements correspond to F variables. An answer to the question should be obtained from first studying the distribution of the cross-products term resulting from $\boldsymbol{\alpha}^T\boldsymbol{\alpha}$. For the simplest case take $g = 1$, then if $\alpha_{1X} \sim N(0, \sigma_{G(XX)})$ and $\alpha_{1Y} \sim N(0, \sigma_{G(YY)})$ are independent (under $H_0 : \sigma_{G(XY)} = 0$), the distribution of $\alpha_{1X} \times \alpha_{1Y}$ can be given in terms of Meijer G-function (SPRINGER; THOMPSON, 1970). The authors prove the following theorem

Theorem 4.1. *The probability density function of the product $z = \prod_{i=1}^{n} x_i$ of n independent Gaussian random variables $N(0, \sigma_i)$, $i = 1, 2, ..., n$, is a Meijer G-function multiplied by a normalizing constant H, i.e.,*

$$g(z) = HG_{N0}^{N0}\left(z^2 \prod_{i=1}^{n} \frac{1}{2\sigma_i} | 0 \right), \tag{47}$$

where

$$H = \left[(2\pi)^{n/2} \prod_{i=1}^{n} \sigma_i \right]^{-1}. \tag{48}$$

Ware e Frank (2003) present the following procedure to compute the *pdf* of a product of two independent Gaussian variables: Take $X \sim N(\mu_x, \sigma_{xx})$ and $Y \sim N(\mu_y, \sigma_{yy})$, inde-

pendent variables and $Z = XY$. Now, define the conditional distribution of $Z|(Y = y)$, which is $Z|y \sim N(y\mu_x, y^2\sigma_{xx})$. Then, find the joint distribution of Z and Y:

$$f_{ZY}(z, y) = f_{Z|Y}(z|y)f_Y(y) \tag{49}$$

Finally, the marginal density $f_Z(z)$ is obtained by integrating $f_{Z|Y}(z|y)f_Y(y)$ with respect to y, i.e.,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z|Y}(z|y)f_Y(y)dy, \tag{50}$$

which can be solved using a numerical procedure.

The *moment generating function* of $Z = XY$ was derived by Craig (1936):

$$\begin{aligned} M_{XY}(t) &= E(e^{txy}) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{txy} f_X(x)f_Y(y)dxdy \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_{xx}} + \frac{(y-\mu_y)^2}{\sigma_{yy}}\right] + txy\right)dxdy \\ &= \frac{\exp([(\sigma_{xx}\mu_y^2 + \sigma_{yy}\mu_x^2)t^2 + 2t\mu_x\mu_y]/2(1 - t^2\sigma_{xx}\sigma_{yy}))}{(1 - t^2\sigma_{xx}\sigma_{yy})^{1/2}} \end{aligned} \tag{51}$$

Then, the mean and variance are

$$\begin{aligned} E(XY) &= M'_{XY}(0) \\ &= \mu_x\mu_y \\ Var(XY) &= M''_{XY}(0) - [M'_{XY}(0)]^2 \\ &= \mu_y^2\sigma_{xx} + \mu_x^2\sigma_{yy} + \sigma_{xx}\sigma_{yy} \end{aligned} \tag{52}$$

Here it is presented a result based on 100,000 simulations of two independent Gaussian variables, $X \sim N(0, 4)$ and $Y \sim N(0, 1)$. The kernel density obtained is shown in Figure 3.

Note that the product xy is not $N(0, 4)$. Indeed, the empirical distribution does not seem to be normally distributed. According to Aroian (1947), only under certain conditions the probability density function of xy approximates to a normal curve. Specifically when the inverses of the coefficients of variation are large, i.e., when $\mu_x >> \sigma_x$ and $\mu_y >> \sigma_y$. This can be seen in Figure 4, where 100,000 random values are simulated from $X \sim N(20, 4)$ and $Y \sim N(10, 1)$ in order to compute the empirical distribution of the product XY.

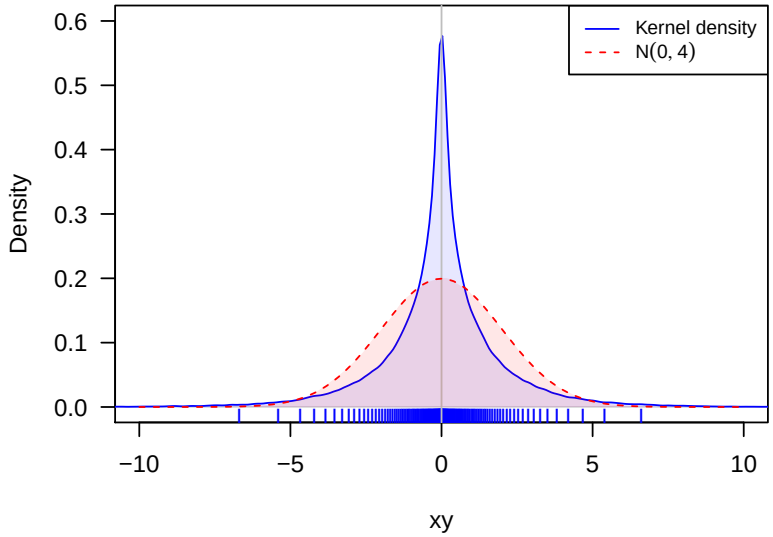

Figure 3 – Distribution of the product of two independent Gaussian variables, $X \sim N(0, 4)$ and $Y \sim N(0, 1)$

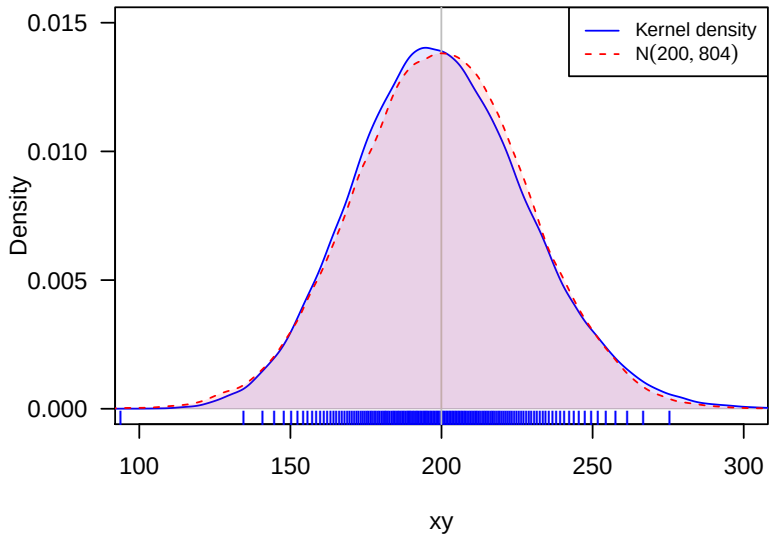

Figure 4 – Distribution of the product of two independent Gaussian variables, $X \sim N(20, 4)$ and $Y \sim N(10, 1)$

According to Aroian (1947) and Aroian et al. (1978), when $r_{XY} \to 0$, the Pearson's Type III function and the Gram-Charlier Type A series are excellent approximations to the probability density function (PDF) of xy. Regardless, although the product of two Gaussian variables does not produce a Gaussian variables, Bromiley (2003) states that the product of two Gaussian PDFs is proportional to a Gaussian PDF.

Hereupon, because the genetic effects α usually have mean zero, it seems to be tricky getting a large value for the inverse of its coefficient of variation. Hence, studying the distribution of a mean cross-products ratio as a quotient of two Gaussian variables may not be suitable.

In the following sections, it is presented three alternatives for evaluating the null hypothesis $H_0 : \sigma_{G(XY)} = 0$. The first is based on the χ^2 approximation from Wilks' Λ statistic for testing independence of two random vectors normally distributed. The second is the bootstrap approach. Finally, the third procedure is simulation-based.

4.2 Analysis of variance and components of covariance

4.2.1 Analysing part of the data

A significant ($p < 0.0001$) effect of progenies on the multivariate response was obtained from the F approximation for the Wilks' Lambda criterion (Table 6). Thus, at least one *non*-null (co)-variance component is expected.

Table 6 – Multivariate analysis of variance of the maize variables: NGROW, DE, DC, PH. Results obtained with part of the data

Source	D.f.	Wilks	Approx. F	Num. D.f.	Den. D.f.	p-value
Environments (E)	3	0.0604	3.79	12	24.10	0.0026
Progenies (G)	4	0.0063	7.45	16	28.13	< 0.0001
Residuals (G $\times$ E)	12					

The Shapiro-Wilk normality test was applied over the residual matrix. The p-value was 0.374, which indicates that multivariate normality is a reasonable assumption.

The matrices of *mean squares and cross-products*, $\mathbf{M}$, obtained via MANOVA are:

$$\mathbf{M}_G = \begin{bmatrix} 22.6317 & 2.2336 & 0.7046 & 0.2125 \\ 2.2336 & 0.4220 & 0.1455 & 0.0348 \\ 0.7046 & 0.1455 & 0.0531 & -0.0026 \\ 0.2125 & 0.0348 & -0.0026 & 0.0857 \end{bmatrix},$$

$$\mathbf{M}_{GE} = \hat{\mathbf{\Sigma}}_{GE} = \begin{bmatrix} 4.5895 & 0.1745 & 0.1103 & 0.0644 \\ 0.1745 & 0.0195 & 0.0129 & 0.0042 \\ 0.1103 & 0.0129 & 0.0110 & 0.0037 \\ 0.0644 & 0.0042 & 0.0037 & 0.0101 \end{bmatrix}$$

Then, the matrix $\mathbf{G}$, containing the (co)-variance components is

$$\mathbf{G} = \hat{\mathbf{\Sigma}}_G = \begin{bmatrix} 4.5105 & 0.76 & 0.68 & 0.12 \\ 0.5148 & 0.1006 & 1.01 & 0.17 \\ 0.1486 & 0.0332 & 0.0105 & -0.11 \\ 0.0370 & 0.0077 & -0.0016 & 0.0189 \end{bmatrix},$$

where the upper triangular elements correspond to the genetic correlations.

Note, for example, that $\hat{\sigma}_{G(1,2)} = 0.5148$. Is this value significant ($p \leq 0.05$)? Or, more precisely, is this value significant in comparison to $\hat{\sigma}_{GE(1,2)} = 0.1745$?

All the genetic correlation estimates involving PH are less than 0.2 in absolute value. Two estimates of correlation involving NGROW are around 0.7. Studying exotic maize hybrids, Munawar et al. (2013) also found a genetic correlation around 0.2 between NGROW and PH. For NGROW and cob girth, these authors found $r_G = 0.67$, which is in agreement with our findings ($r_G = 0.68$ for NGROW and DC). Tengan et al. (2012) found low phenotypic correlations between DC and PH (around 0.00 and 0.47) in a backcross breeding program involving normal and opaque-2 maize.

There is one estimate of genetic correlation that is even greater than 1.0. And that is probably because the condition number of **G** was 3549.61, indicating severe collinearity, which is probably a consequence of estimating with the method of moments. According to Hill e Thompson (1978), if **G** is non-positive definite, heritabilities and ordinary or partial genetic correlations can fall outside their valid limits. A comprehensive study of the sampling distribution of genetic correlation estimated from the MANOVA was carried out by Liu et al. (1997).

4.2.2 Analysing all the data

When considering all the 12 environments and 256 progenies, the following result (Table 7) was found with MANOVA.

Table 7 – Multivariate analysis of variance of the maize variables: NGROW, DE, DC, PH

Source	D.f.	Wilks	Approx. F	Num. D.f.	Den. D.f.	p-value
Environments (E)	11	0.0538	279.28	44	10722	< 0.0001
Progenies (G)	255	0.0362	14.19	1020	11210	< 0.0001
Residuals (G $\times$ E)	2805					

This time, residuals did not present multivariate normality ($p < 0.0001$).

The matrices of *mean squares and cross-products*, **M**, obtained via MANOVA are:

$$\mathbf{M}_G = \begin{bmatrix} 75.1352 & 3.2533 & 0.8702 & 0.9218 \\ 3.2533 & 0.5042 & 0.2591 & 0.0653 \\ 0.8702 & 0.2591 & 0.1978 & 0.0296 \\ 0.9218 & 0.0653 & 0.0296 & 0.1028 \end{bmatrix},$$

$$\mathbf{M}_{GE} = \hat{\mathbf{\Sigma}}_{GE} = \begin{bmatrix} 6.3571 & 0.2358 & 0.0849 & 0.0674 \\ 0.2358 & 0.0270 & 0.0110 & 0.0043 \\ 0.0849 & 0.0110 & 0.0093 & 0.0018 \\ 0.0674 & 0.0043 & 0.0018 & 0.0074 \end{bmatrix}$$

Then, the matrix $\mathbf{G}$, containing the (co)-variance components is

$$\mathbf{G} = \hat{\mathbf{\Sigma}}_G = \begin{bmatrix} 5.7315 & 0.53 & 0.22 & 0.33 \\ 0.2515 & 0.0398 & 0.83 & 0.29 \\ 0.0654 & 0.0207 & 0.0157 & 0.21 \\ 0.0712 & 0.0051 & 0.0023 & 0.0080 \end{bmatrix},$$

where the upper triangular elements correspond to the genetic correlations.

Observe now that the estimate $\hat{\sigma}_{G(1,2)} = 0.2515$ is closer to $\hat{\sigma}_{GE(1,2)} = 0.2358$. The correlations involving PH remained low. And, although there is no correlation greater than one, $\mathbf{G}$ also presented severe multicollinearity (NC = 2086.45).

4.3 Wilks' and Pillai's statistics

Before presenting the results obtained using the Wilks' (eq. 39) and Pillai (eq. 41) criteria, it was verified their approximations to a χ^2_1 distribution. In Figure 5 it is shown the kernel densities of the quantities $-(g-1)\log_e \Lambda$ (eq. 40) and $(g-1)T_n$ after simulating 10,000 genetic covariance matrices under the null hypothesis $H_0 : \sigma_{G(ij)} = 0,\ \forall i \neq j$, $i, j = 1, 2, 3, 4$, i.e.,

$$\mathbf{G}_{(H_0)} \sim W_4 \left(4, \begin{bmatrix} 4.5105 & 0 & 0 & 0 \\ 0 & 0.1006 & 0 & 0 \\ 0 & 0 & 0.0105 & 0 \\ 0 & 0 & 0 & 0.0189 \end{bmatrix} \right)$$

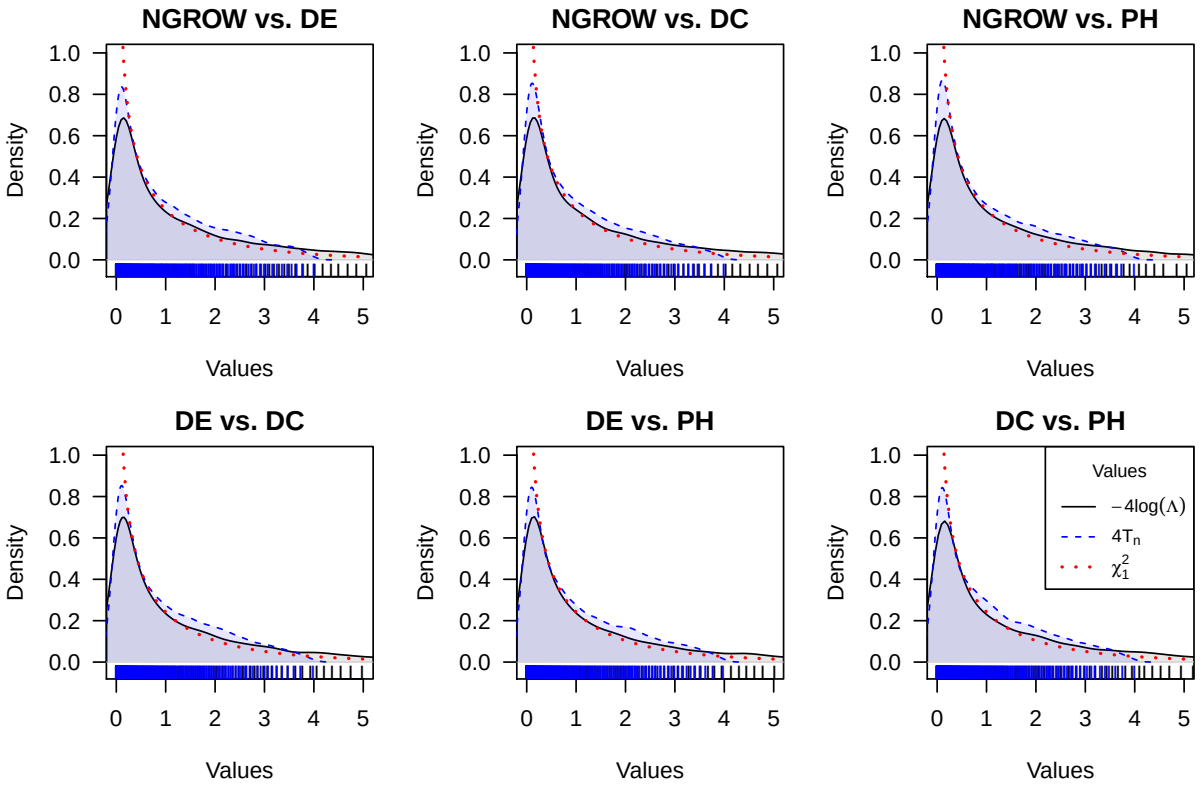

Figure 5 – Density of the pairwise Wilks' Lambda and Pillai's approximations to a χ^2_1 distribution. Results obtained with part of the data set

As we can note, the approximations for Wilks' Lambda and Pillai's T_n can be considered reliable if one intends to use them for calculating p-values, since right tails are quite similar to the corresponding χ^2_1. Therefore, some tests presented here were based only on Pillai's statistic, due to collinearity problems.

The output of the R function designed to perform the test of $H_0 : \sigma_{G(ij)} = 0, \forall i \neq j$, based on Wilks'and Pillai's approximation is given in Figure 6. It consists of three matrices. The first one is $\mathbf{G}$, where the upper triangular elements correspond to genetic correlations. The second and third matrices give values of the χ^2_1 approximation for Wilks' Lambda and Pillai's T_n, respectively, on lower triangular part, whereas the upper triangular contains the associated p-values. For example, take the variables NGROW (1) and DE (2). We know $\hat{\sigma}_{G(12)} = 0.5148$ and $\hat{\rho}_{G(12)} = 0.7641$. The second matrix shows that $-4\log(\Lambda) = 3.5070$, whose p-value is 0.0611. On the other hand, the third matrix

shows that the p-value associated with $4T_n = 2.3355$ is 0.1265. Then, we have got some evidence against $H_0 : \sigma_{G(12)} = 0$.

```
          Genetic Covariance Test

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE      DC      PH
NGROW 4.5105 0.7641  0.6817  0.1268
DE    0.5148 0.1006  1.0186  0.1755
DC    0.1486 0.0332  0.0105 -0.1109
PH    0.0370 0.0077 -0.0016  0.0189

Chi-Sq (df = 1) approx. (Wilks) and p-values (upper triangular):
       NGROW     DE     DC     PH
NGROW     NA 0.0611 0.1139 0.7989
DE    3.5070     NA    NaN 0.7235
DC    2.4997    NaN     NA 0.8240
PH    0.0649 0.1251 0.0495     NA

Chi-Sq (df = 1) approx. (Pillai) and p-values (upper triangular):
       NGROW     DE     DC     PH
NGROW     NA 0.1265 0.1728 0.7997
DE    2.3355     NA 0.0416 0.7256
DC    1.8588 4.1500     NA 0.8245
PH    0.0644 0.1232 0.0492     NA
```

Figure 6 – Output of the R function designed to test genetic covariance through Wilks' and Pillai's statistics. Results obtained with part of the data set

It is noteworthy that Wilks' approximation seems to be more sensitive than Pillai's. Observe that the difference between p-values is larger as the estimate of genetic covariance increases (in absolute value). See, for example, that for $\sigma_{G(12)}$ this difference is of around 0.06. Almost the same for $\sigma_{G(13)}$. On the other hand, for low values of genetic covariance, such as $\hat{\sigma}_{G(14)}$, $\hat{\sigma}_{G(24)}$ and $\hat{\sigma}_{G(34)}$, the differences are irrelevant. Nonetheless, due to collinearity problems between DE and DC, Wilks' statistic cannot be evaluated.

Using all the data, i.e., the 12 environments and 256 progenies, the χ^2_1 approximations were even better (Figure 7), as expected.

The output shown in Figure 8 was obtained. Now, even though some estimates are lower than before, the sensitivity of both tests (Wilks and Pillai) increased roughly, as expected. And as a consequence, all p-values are lower than 0.01.

Figure 9 shows the genetic correlations and their exact two-sided p-values, i.e., calculated from the exact PDF of the correlation coefficient under the null hypothesis ($H_0 : \rho_{G(ij)} = 0$, for $i, j = 1, 2, 3, 4$), for both partial and complete data set. We observe that p-values related to correlations greater than 0.6 (absolute value) are intermediate in comparison to those obtained with Wilks' (lowest p-values) and Pillai's approaches. On the other hand, when there is weak evidence against H_0 (NGROW vs PH, DE vs PH and DC vs PH) the exact p-values tend to be a little higher. Nevertheless, we can observe

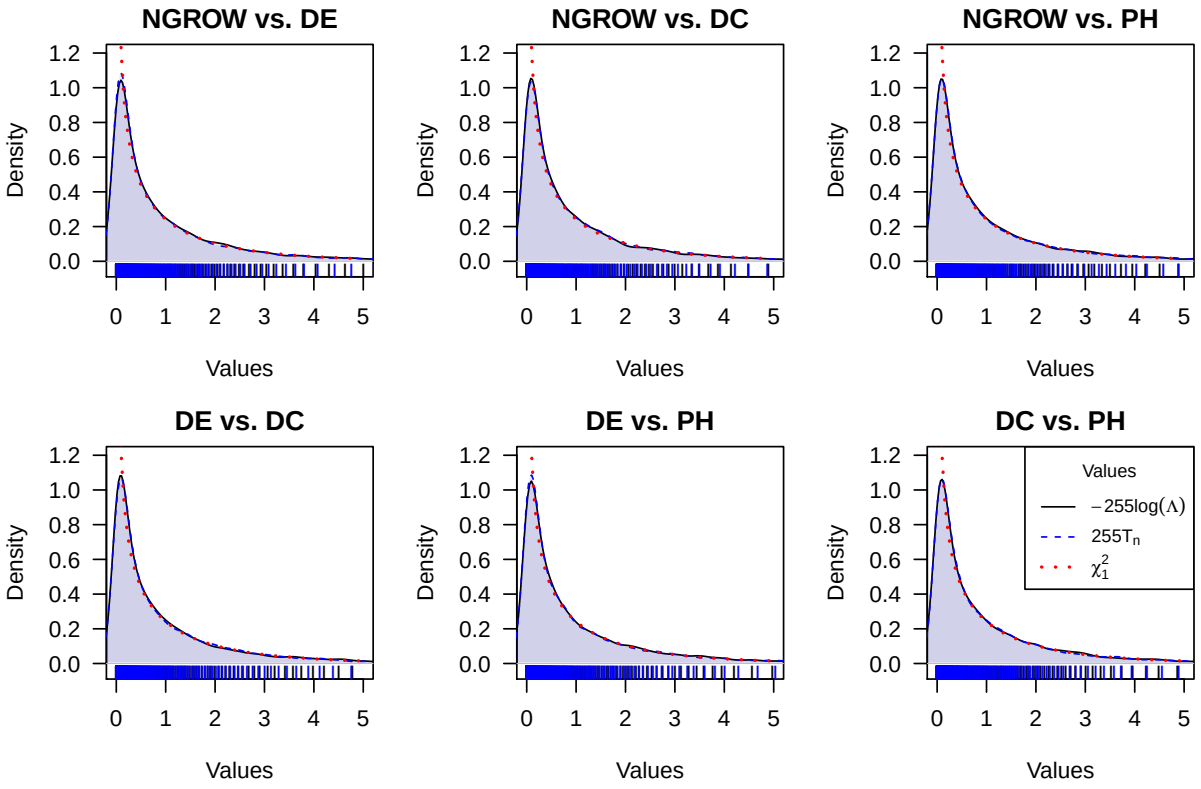

Figure 7 – Density of the pairwise Wilks' Lambda and Pillai's approximations to a χ^2_1 distribution

```
          Genetic Covariance Test

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE     DC     PH
NGROW 5.7315 0.5267 0.2181 0.3335
DE    0.2515 0.0398 0.8272 0.2860
DC    0.0654 0.0207 0.0157 0.2074
PH    0.0712 0.0051 0.0023 0.0080

Chi-Sq (df = 1) approx. (Wilks) and p-values (upper triangular):
        NGROW       DE      DC    PH
NGROW      NA   0.0000  0.0004 0e+00
DE    82.8611       NA  0.0000 0e+00
DC    12.4296 294.0248      NA 8e-04
PH    30.0702  21.7665 11.2152    NA

Chi-Sq (df = 1) approx. (Pillai) and p-values (upper triangular):
        NGROW       DE      DC    PH
NGROW      NA   0.0000  0.0005 0e+00
DE    70.7455       NA  0.0000 0e+00
DC    12.1315 174.5026      NA 9e-04
PH    28.3649  20.8634 10.9722    NA
```

Figure 8 – Output of the R function designed to test genetic covariance through Wilks' and Pillai's statistics

similarities among all respective p-values.

```
# Incomplete data (5 progenies, 4 environments)

          Exact Correlation Test

Genetic Correlations and p-values (upper triangular):
       NGROW     DE      DC     PH
NGROW 1.0000 0.0769  0.1359 0.8107
DE    0.7641 1.0000  0.0005 0.7395
DC    0.6817 1.0186  1.0000 0.8344
PH    0.1268 0.1755 -0.1109 1.0000

Alternative hypothesis: two.sided

# ------------------------------------------------------------
# Complete data (256 progenies, 12 environments)

          Exact Correlation Test

Genetic Correlations and p-values (upper triangular):
       NGROW     DE     DC    PH
NGROW 1.0000 0.0000 0.0004 0e+00
DE    0.5267 1.0000 0.0000 0e+00
DC    0.2181 0.8272 1.0000 8e-04
PH    0.3335 0.2860 0.2074 1e+00

Alternative hypothesis: two.sided
```

Figure 9 – Output of the R function designed to perform an exact test of the genetic covariance

In general and considering the sample size effect, the p-values for both approximations are in agreement (Figure 10) and consistent with respect to the genetic covariance (and correlation) estimates.

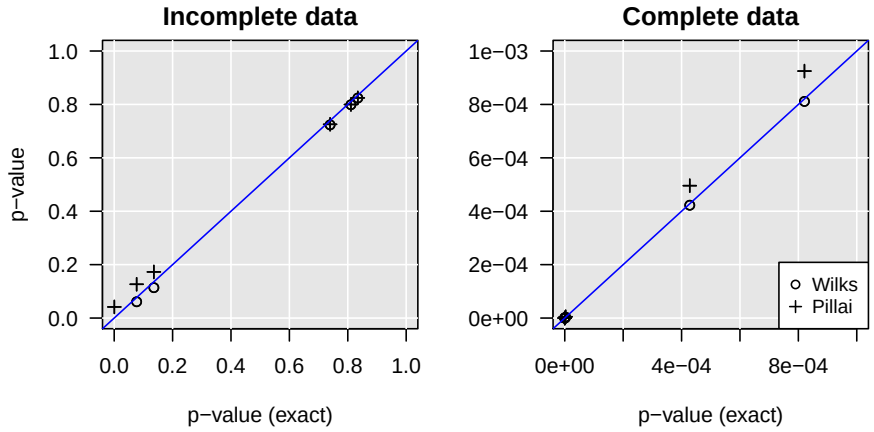

Figure 10 – Shepard's diagrams for evaluating agreement among the exact p-values and those obtained using Wilks' and Pillai's approximations

4.4 Bootstrap approach

Bootstrapping part of the data 400 times, randomized matrices $\mathbf{M}_G$ and $\mathbf{M}_{GE}$ were obtained. The empirical distributions of mean squares and cross-products ratios are shown in Figure 11. Because the bootstrap process can generate very large values of the ratio, data were winsorized in 0.5% on each tail. The first graphical line is related to mean squares ratio for the variables NGROW, DE and DC. Rods indicate the observed ratio. The density of a F distribution with 4 and 12 degrees of freedom was superimposed on the kernel densities for comparisons. Although the behaviors of the kernel densities are generally similar to the F PDF, there are some differences. Nevertheless, the locations of the empirical distributions seem to be correct. Also, the associated p-values should lead to the same conclusion about $H_0 : \sigma_{G(ij)} = 0,\ \forall i = j$.

Empirical distributions of mean cross-products ratios tended to be symmetric around some point between 0 and 1, as expected, which is not the same for every pair of variables. Approximations to a non-standard Cauchy(μ_0, γ) distribution were checked, where μ_0 is the location parameter, mode and median, and γ is the scale parameter, half the interquartile range. Both, symmetry and approximation to Cauchy distribution were characteristics related to the observed statistics. They are more evident in cases whose null hypotheses are more likely. See for example DC vs PH, whose distribution shows best symmetry and approximation. The genetic correlation is the lowest (-0.11). On the other hand, the extreme value of genetic correlation between DE and DC promotes no symmetry and Cauchy approximation. Thus, this problem is due to two probable causes: 1) the randomization process, i.e., the bootstrap resampling may not be efficient enough on generating matrices under H_0, and/or 2) the structure of the matrices $\mathbf{M}$. According to the estimates of genetic correlation, cause 2 seems more likely, since the matrix $\mathbf{G}$ presents collinearity problems.

The output of the function designed to test $H_0 : \sigma_{G(ij)} = 0$ via mean squares and cross-products ratio is shown in Figure 12. It consists of two matrices: the first contains genetic (co)variances (lower triangular) and correlations (upper triangular) and the second contains the mean squares and cross-products ratios (lower triangular) and associated empirical p-values (upper triangular, only for cross-products). Note that, because genetic covariances are effected by variable scaling, a good guide for evaluating their magnitude is the respective correlations. Then, pairs of variables presenting genetic correlation around 0.7 or above (NGROW vs DE, NGROW vs DC, DE vs DC) presented $p < 0.01$, indicating that the corresponding genetic covariance are significant. On the other hand, mean cross-products ratio of NGROW vs PH (0.12), DE vs PH (0.17) DC vs PH (-0.11) presented $p > 0.05$, indicating that respective genetic covariances can be considered null. These conclusions are accordant to those made up by using Wilks's and Pillai's approaches. However, the p-values are quite different in some cases. For example, take the correlation

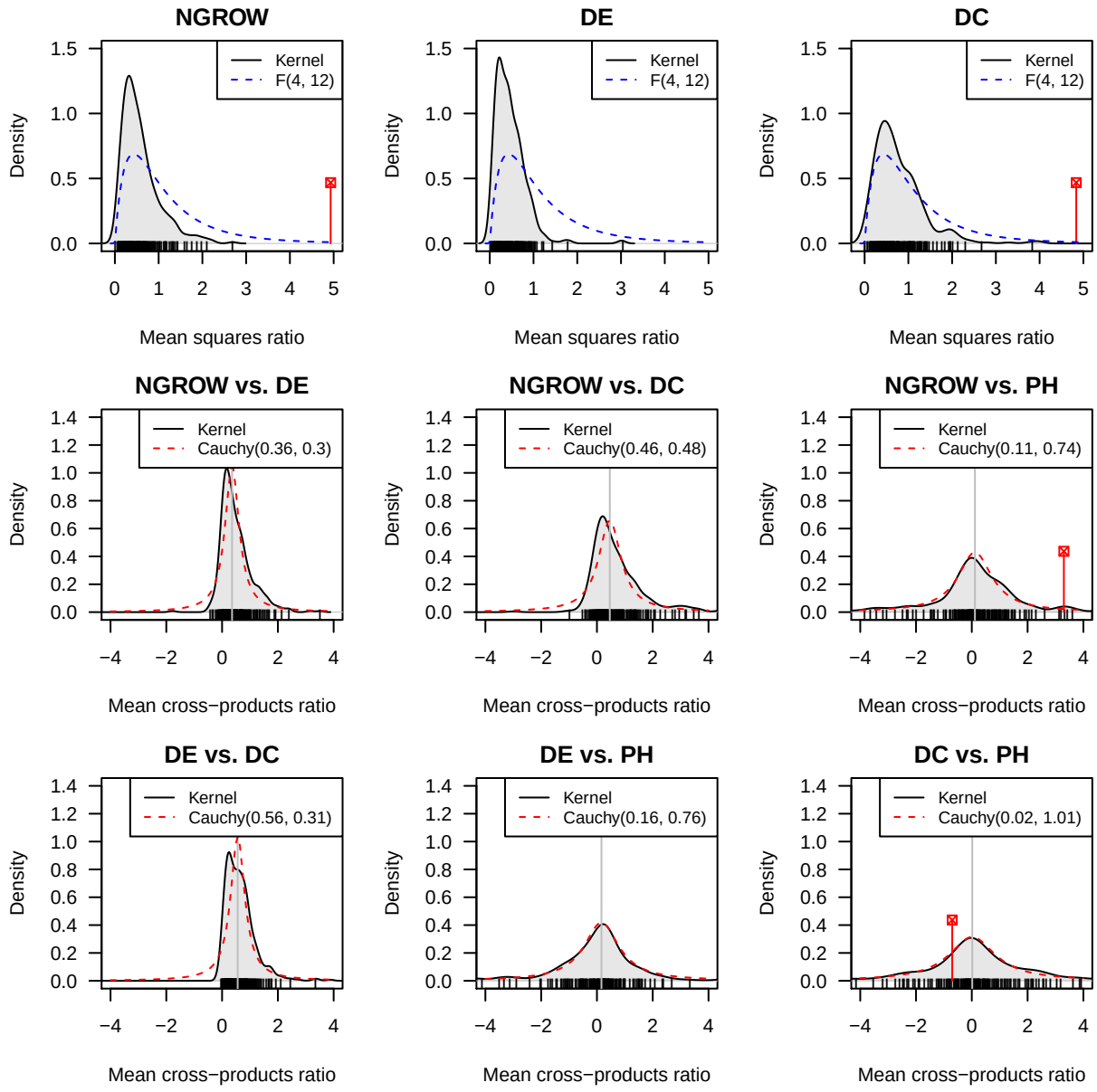

Figure 11 – Density of the bootstrapped mean squares and cross-products ratios. Results obtained with part of the data

between NGROW and PH (0.12) and between DC and PH (-0.11). Considering their absolute values, they are very similar. Based on Wilks' Lambda, p-values are 0.7989 and 0.8240, respectively. But here $p = 0.1675$ and $p = 0.5900$, respectively. Why? The answer lies on the test statistics. Here, the ratio of mean cross-products is used to measure the magnitude of the genetic covariance, i.e., the residual covariance is taken into account. Observe that the ratio related to NGROW vs PH (3.3012) is much higher than that related to DC vs PH (-0.7012), again in absolute value. It means that the genetic covariance of NGROW and PH makes a greater contribution to the corresponding phenotypic covariance than does the genetic covariance of DC and PH. Since cross-products ratios are dimensionless, comparing values is a valid task. Nonetheless, Wilks' Lambda disregards

the information of residual covariance for testing genetic covariance.

```
          Genetic Covariance Test via Bootstrap

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE      DC      PH
NGROW 4.5105 0.7641  0.6817  0.1268
DE    0.5148 0.1006  1.0186  0.1755
DC    0.1486 0.0332  0.0105 -0.1109
PH    0.0370 0.0077 -0.0016  0.0189

Mean Sq and Cross-Prods Ratios and p-values (upper triangular)
based on 400 estimates:
        NGROW      DE      DC     PH
NGROW  4.9312  0.0000  0.0075 0.1675
DE    12.8014 21.6422  0.0000 0.0875
DC     6.3903 11.2840  4.8367 0.5900
PH     3.3012  8.2161 -0.7012 8.4603

Alternative hypothesis: two.sided
```

Figure 12 – Output of the R function designed to test genetic covariance via cross-products ratio based on bootstrap resampling. Results obtained with part of the data

One could also verify that the ratio for DE vs PH is 8.2161 and for NGROW vs DC is 6.3903. In contrast, the associated p-values are 0.0875 and 0.0075, respectively. But note that the genetic correlations are 0.1755 and 0.6817, respectively. Hence, the test based on mean cross-products ratio is probably related to both the value of the genetic covariance and the magnitude of the latter with respect to the residual covariance.

Using the complete data, the approximation to a $F(255, 2805)$ distribution for mean squares ratio was not verified (Figure 13). The empirical distributions were shifted in some way. Kernel densities of mean cross-products kept their general shape, but some of them, such as that for NGROW vs PH and De vs PH, did not remain inside the expected interval [-1, 1]. Approximations to a normal distribution are verified.

Inferences were significantly affected by the sample size, as expected. As occurred for Wilks' Lambda, all p-values were less than 0.001 (Figure 14).

4.5 Simulation approach

Simulating 10,000 Wishart matrices $\mathbf{M}_G^{(H_0)}$ and $\mathbf{M}_{GE}$ the empirical distribution of mean squares and cross-products ratios were obtained (Figure 15). Data were also winsorized in 0.5% on each tail. Again, approximations to a $F(4, 12)$ and Cauchy(μ_0, γ) distributions were checked. The empirical ratio of mean squares is undoubtedly a F variable. The approximations are much more evident than those obtained by bootstrapping data (Figure 11). Kernel densities of mean cross-products showed the same general behavior as those found with the bootstrap approach. But now, location and scale parameters

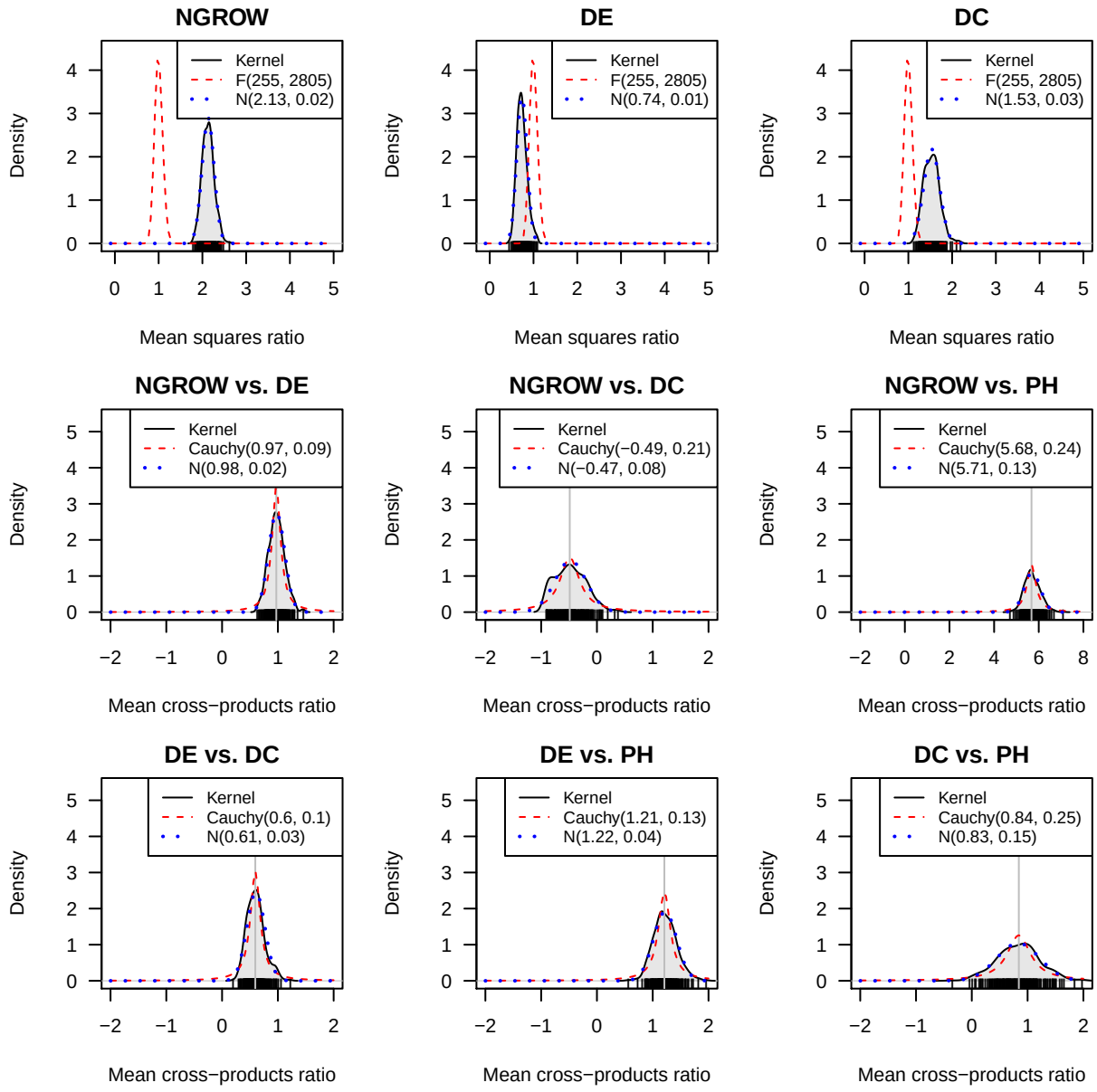

Figure 13 – Density of the bootstrapped mean squares and cross-products ratios

are greater. In fact, it was expected that the actual distribution (under H_0) locations are 1 or -1. Thus, Monte Carlo simulation showed more effectiveness. Again, the conditioning of **G** matrix seems to roughly affect the empirical distribution of mean cross-products ratios. See for example that the distribution related to DE vs DC ($r_G \approx 1.02$) is completely skewed. Once more, symmetry and approximations to a Cauchy distribution were dependent on the magnitude of genetic covariance estimate.

The output of the function designed to perform the genetic covariance test via Monte Carlo simulation of Wishart matrices is shown in Figure 16. Actually, it contains the same components as those in Figure 12, based on bootstrap. In fact, all p-values are quite similar to the latter, but slightly higher, except from that related to NGROW vs DC ($p = 0.0652$), which could allow one to conclude in favor of H_0.

```
          Genetic Covariance Test via Bootstrap

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE     DC     PH
NGROW 5.7315 0.5267 0.2181 0.3335
DE    0.2515 0.0398 0.8272 0.2860
DC    0.0654 0.0207 0.0157 0.2074
PH    0.0712 0.0051 0.0023 0.0080

Mean Sq and Cross-Prods Ratios and p-values (upper triangular)
based on 400 estimates:
        NGROW      DE      DC      PH
NGROW 11.8192  0.0000  0.0000  0.0000
DE    13.7983 18.6856  0.0000  0.0000
DC    10.2558 23.5940 21.2446  0.0000
PH    13.6740 15.2891 16.7606 13.8862

Alternative hypothesis: two.sided
```

Figure 14 – Output of the R function designed to test genetic covariance via cross-products ratio based on bootstrap resampling

Using all the data set, empirical distributions of mean squares ratio are still well approximated by the F PDF and also by a Normal PDF (Figure 17), as expected since the following result[8] is observed:

$$F(\nu_1, \nu_2) \rightarrow^{\nu_1 X}_{\nu_2 \rightarrow \infty} \chi^2_{\nu_1} = Gamma(\nu_1/2, 2) \overset{\nu_1/2 \rightarrow \infty}{\rightarrow} Normal(\nu_1, 2\nu_1) \tag{53}$$

Unlike bootstrap (Figure 13), the simulation approach has provided a correct location of all the empirical distributions. Probably, simulations were less affected by the quality of the matrices $\mathbf{M}$ than the bootstrap.

Normal PDFs also provide a reasonable approximation to the distributions of mean cross-products ratio. The location parameter now achieved the unit for all pairs of variables, as well as symmetry around this value. These expected results were verified only in some cases with the bootstrap approach (Figure 13).

Likewise in the other approaches, all the p-values indicate strong ($p < 0.0001$) evidence against H_0 (Figure 18).

4.6 Further discussion

A problem found in the Wilks' and Pillai's approaches is the necessity of estimating $\boldsymbol{\Sigma_G}$. According to Cheverud (1988), Revell et al. (2010) and Proa et al. (2012), a phenotypic covariance matrix, estimated with large samples might approach $\boldsymbol{\Sigma_G}$ more accurately than genetic covariances estimated from small effective sample sizes, at least for morphometric data. Meyer e Kirkpatrick (2010) stated that in quantitative genetic anal-

[8] For more details see Casella e Berger (2002), p. 627.

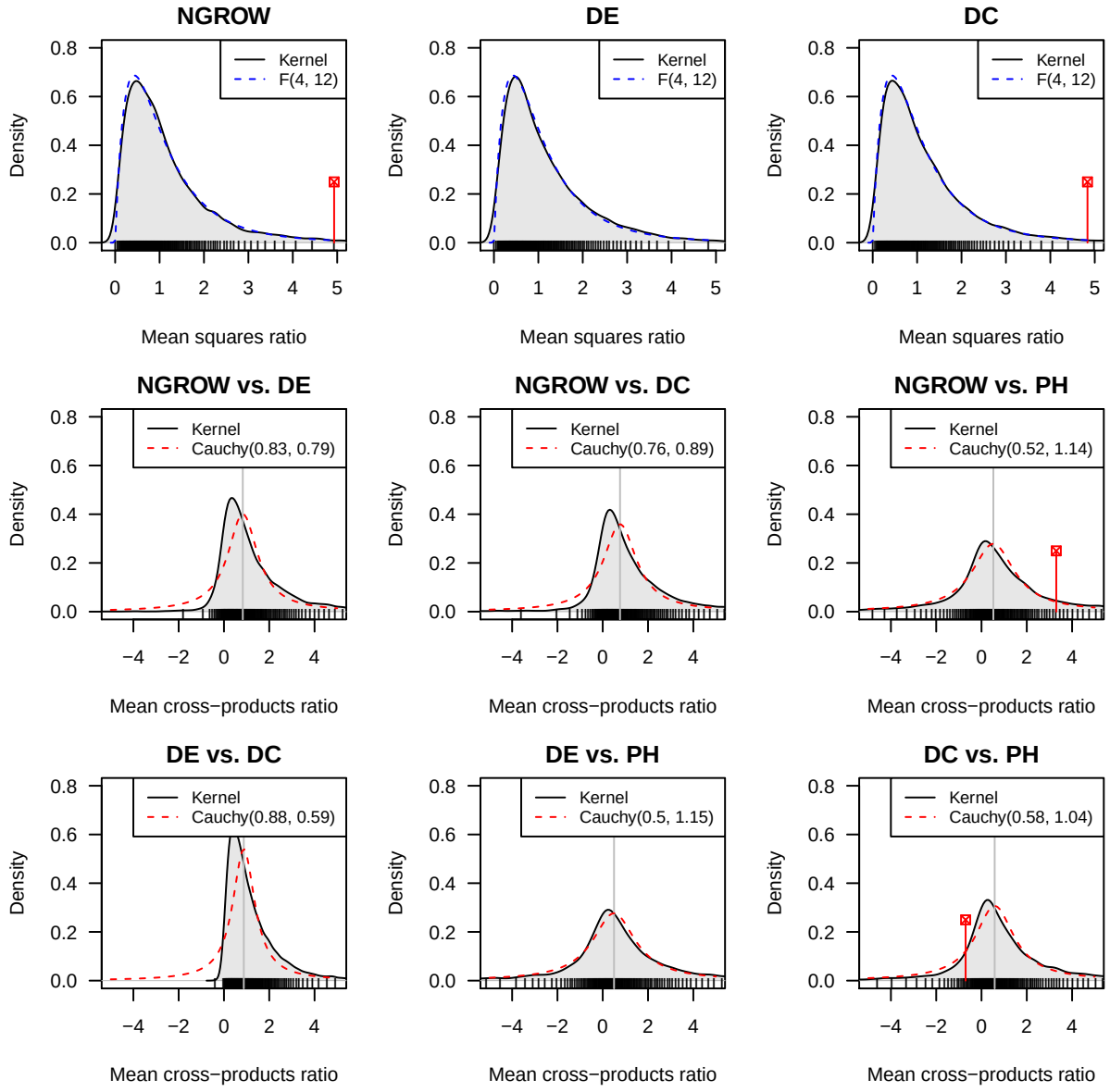

Figure 15 – Density of the simulated mean squares and cross-products ratios. Results obtained with part of the data

yses, we attempt to partition observed, overall (phenotypic) covariances into their genetic and environmental components. Typically, this results in strong sampling correlations between them. Hence, while the partitioning into sources of variation and estimates of individual covariance matrices may be subject to substantial sampling variances, their sum, i.e., the phenotypic covariance matrix, can generally be estimated much more accurately.

In Table 8 it is presented the estimates of genetic covariance, correlation and the p-values calculated through all the approaches described, for two sample sizes. Based on the p-values, two types of test can be identified: 1) those that do not take into account the residual covariance on its test statistic - Wilks, Pillai and 'Exact', and 2) those that do - tests via mean cross-products ratio based on bootstrap and Monte Carlo simulation. Using part of the data, i.e., 5 progenies and 4 environments, p-values of the first group

```
          Genetic Covariance Test via Wishart Simulation

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE      DC      PH
NGROW 4.5105 0.7641  0.6817  0.1268
DE    0.5148 0.1006  1.0186  0.1755
DC    0.1486 0.0332  0.0105 -0.1109
PH    0.0370 0.0077 -0.0016  0.0189

Mean Sq and Cross-Prods Ratios and p-values (upper triangular)
based on 9999 estimates:
        NGROW      DE      DC     PH
NGROW  4.9312  0.0166  0.0652 0.2248
DE    12.8014 21.6422  0.0015 0.0997
DC     6.3903 11.2840  4.8367 0.6025
PH     3.3012  8.2161 -0.7012 8.4603

Alternative hypothesis: two.sided
```

Figure 16 – Output of the R function designed to test genetic covariance via cross-products ratio based on Monte Carlo Simulation of Wishart matrices. Results obtained with part of the data

tend to be larger than those of the second group. Nevertheless, all tests are related and all of them were similarly affected by the sample size.

Figure 19 shows the p-values correlation matrix for the tests presented. Information of DE vs DC were excluded, due to missing value of the Wilks' test. All tests were highly correlated, specially inside their group, despite differences among magnitudes of p-values.

4.7 Extra examples

Three illustrative data sets were used in this section in order to evaluate the performance of the approaches under different conditions. Data sets were extracted from experiments involving different crops. All of them were carried out under a randomized block design.

1. *Maize*: 10 genotypes of maize evaluated at 3 blocks on the response variables: plant height (PH), grains yield (YI) and percentage of plant falling (FA).

2. *Garlic*: 89 accessions of garlic evaluated at 4 blocks on the response variables: bulb diameter (BD), bulb length (BL) and bulb yield (YI).

3. *Pepper*: 9 accessions of chili pepper evaluated at 2 blocks on the response variables: fruit length (FL), peduncle length (PL) and fruit weight (FW).

The **G** matrices presented the following condition numbers: 21.8, 266.3 and 449.9, respectively. Table 9 contains the estimates of genetic covariance, correlation and respective p-values according to the tests built through Wilks, Pillai, null correlation PDF and

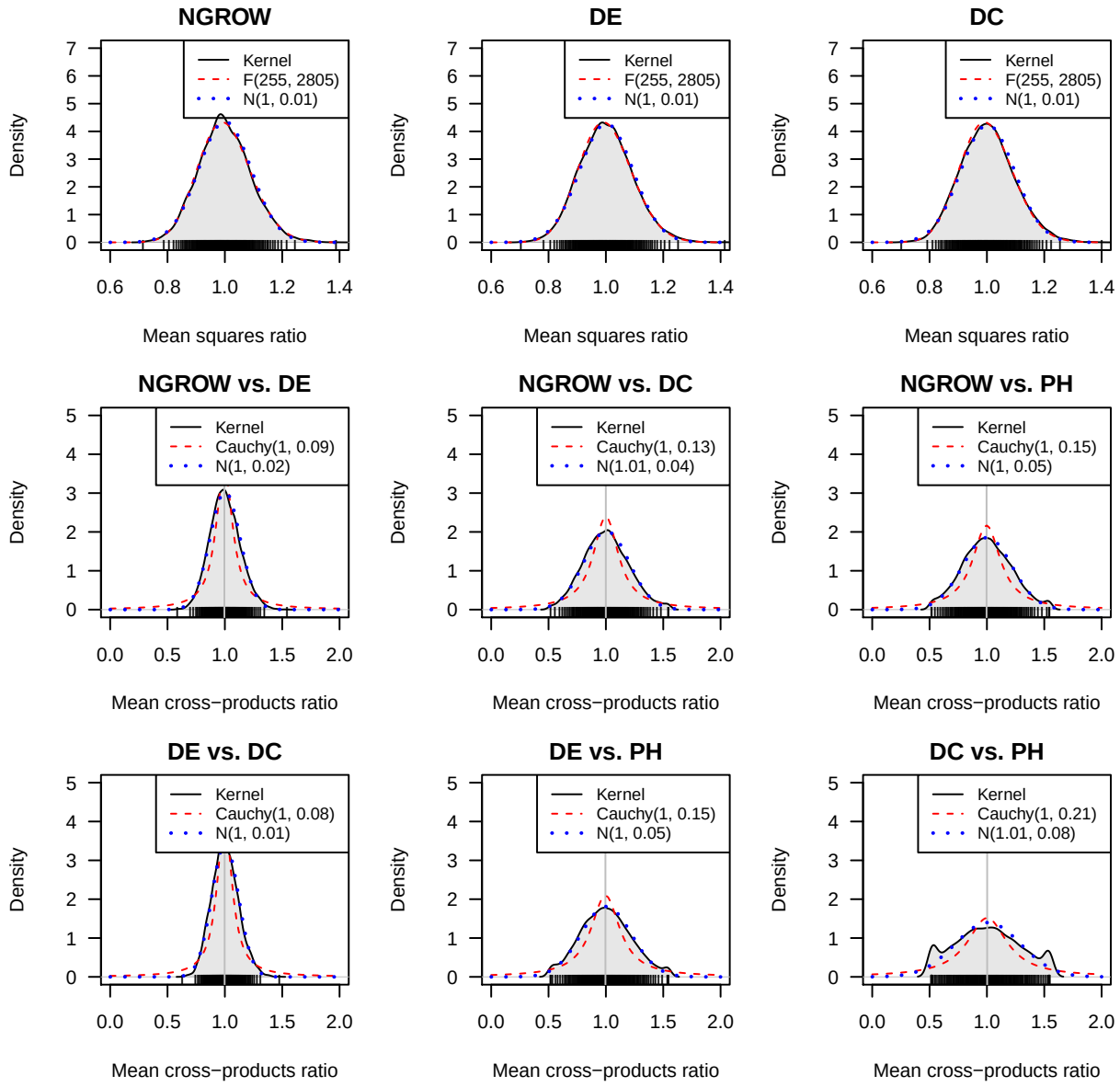

Figure 17 – Density of the simulated mean squares and cross-products ratios

mean cross-products ratio via bootstrap resampling and simulations of Wishart matrices. For all data sets, p-values obtained using Wilks and null correlation PDF are quite close to each other. Pillai's p-values are as closer to the latter two as there is less evidence against H_0. Moreover, there is agreement among these tests for all data sets.

Once more, the magnitude of the genetic covariance with respect to its relative residual covariance has been taken into account by the tests based on mean cross-products ratio. For the *Maize* data set, all tests appear to lead to the same conclusion. However, the test based on Wishart simulation does not. See that for PH vs YI the test based on bootstrap was more sensitive. For the *Garlic* data, there was total agreement among the tests, as expected since all genetic correlations are greater than 0.95. In addition, they may have been affected by the large sample size. For the *Pepper* data, the simulation approach

```
          Genetic Covariance Test via Wishart Simulation

Genetic (Co)variances and Correlations (upper triangular):
       NGROW     DE     DC     PH
NGROW 5.7315 0.5267 0.2181 0.3335
DE    0.2515 0.0398 0.8272 0.2860
DC    0.0654 0.0207 0.0157 0.2074
PH    0.0712 0.0051 0.0023 0.0080

Mean Sq and Cross-Prods Ratios and p-values (upper triangular)
based on 9999 estimates:
        NGROW      DE      DC      PH
NGROW 11.8192  0.0000  0.0000  0.0000
DE    13.7983 18.6856  0.0000  0.0000
DC    10.2558 23.5940 21.2446  0.0000
PH    13.6740 15.2891 16.7606 13.8862

Alternative hypothesis: two.sided
```

Figure 18 – Output of the R function designed to test genetic covariance via cross-products ratio based on Monte Carlo Simulation of Wishart matrices

Table 8 – Genetic covariances, correlations and associated p-values calculated through the approaches of Wilks' Lambda, Pillai's T_n, null correlation PDF (Exact) and mean cross-products ratio based on bootstrap resampling and Monte Carlo simulation

Data	Pairs	Cov.	Cor.	p-value				
				Wilks	Pillai	Exact	Bootstrap	Simulation
Incomplete	NGROW vs DE	0.5148	0.7641	0.0611	0.1265	0.0769	0.0000	0.0166
	NGROW vs DC	0.1486	0.6817	0.1139	0.1728	0.1359	0.0075	0.0652
	NGROW vs PH	0.0370	0.1268	0.7989	0.7997	0.8107	0.1675	0.2248
	DE vs DC	0.0332	1.0186	NA	0.0416	0.0005	0.0000	0.0015
	DE vs PH	0.0077	0.1755	0.7235	0.7256	0.7395	0.0875	0.0997
	DC vs PH	-0.0016	-0.1109	0.8240	0.8245	0.8344	0.5900	0.6025
Complete	NGROW vs DE	0.2515	0.5267	0.0000	0.0000	0.0000	0.0000	0.0000
	NGROW vs DC	0.0654	0.2181	0.0004	0.0005	0.0004	0.0000	0.0000
	NGROW vs PH	0.0712	0.3335	0.0000	0.0000	0.0000	0.0000	0.0000
	DE vs DC	0.0207	0.8272	0.0000	0.0000	0.0000	0.0000	0.0000
	DE vs PH	0.0051	0.2860	0.0000	0.0000	0.0000	0.0000	0.0000
	DC vs PH	0.0023	0.2074	0.0008	0.0009	0.0008	0.0000	0.0000

NA: not available

showed more sensitivity than the bootstrap's. For example, in the case involving FL vs FW ($r_G = 0.1470$), only the test based on simulation indicates evidence ($p < 0.05$) against

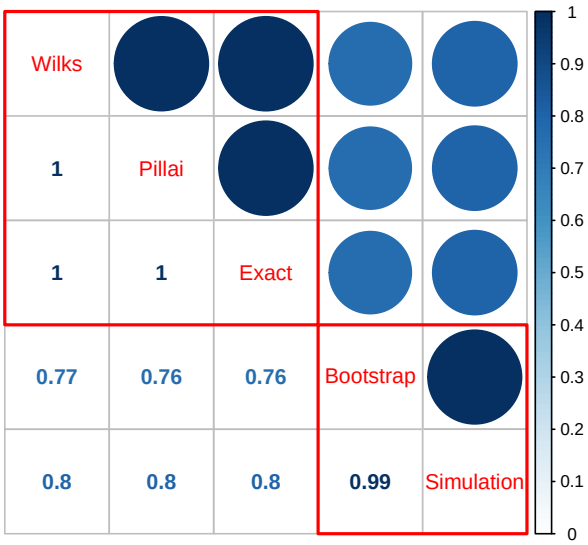

Figure 19 – Graphical representation of the correlation matrix among tests based on their p-values

Table 9 – Genetic covariances, correlations and associated p-values calculated through the approaches of Wilks' Lambda, Pillai's T_n, null correlation PDF (Exact) and mean cross-products ratio based on bootstrap resampling and Monte Carlo simulation from three illustrating data sets

Data	Pairs	Cov.	Cor.	p-value				
				Wilks	Pillai	Exact	Bootstrap*	Simulation
Maize	PH vs YI	-3830.67	-0.6289	0.0333	0.0592	0.0382	0.0450 (9.73)	0.0862
	PH vs FA	-4759.46	-0.6653	0.0218	0.0460	0.0255	0.0000 (31.31)	0.0000
	YI vs FA	358.42	0.1051	0.7518	0.7524	0.7583	0.4850 (1.24)	0.3809
Garlic	BD vs BL	8.0002	0.9709	0.0000	0.0000	0.0000	0.0000 (2.89)	0.0000
	BD vs YI	7.1163	0.9745	0.0000	0.0000	0.0000	0.0000 (3.05)	0.0000
	BL vs YI	5.8318	0.9857	0.0000	0.0000	0.0000	0.0000 (3.39)	0.0000
Pepper	FL vs PL	0.1183	0.1723	0.6024	0.6051	0.6124	0.1175 (5.38)	0.0731
	FL vs FW	1.4929	0.1470	0.6575	0.6592	0.6662	0.1725 (5.79)	0.0479
	PL vs FW	1.4987	0.7106	0.0119	0.0330	0.0143	0.0350 (19.45)	0.0000

*Values in round brackets represent mean cross-products ratio.

H_0. Note that the mean cross-products ratio is 5.79, meaning that the joint effect of the genetic and residual covariance on the phenotypic covariance is 5.79 times grater than the effect of the residual covariance only.

5 CONCLUDING REMARKS

5.1 Conclusions

Only under certain conditions does the probability density function of the product of two random Gaussian variables approximate the normal curve. Therefore, studying the distribution of a mean cross-products ratio as a quotient of two Gaussian variables is not suitable.

Wilks' and Pillai's statistics for testing independence of two sets of Gaussian variables can be used to test genetic covariance. Their approximations to a χ^2_1 distribution were checked for two scenarios of sample size, and using the degrees of freedom associated with the genetic factor as the number of independent observations.

We observed similarities among respective p-values calculated using both approaches and p-values obtained from the exact correlation PDF under the null hypothesis (H_0 : $\rho_{G(ij)} = 0$), which is equivalent to a Student's t-test for the correlation coefficient on $g-1$ degrees of freedom. In addition, we found that Wilks' Lambda is more sensitive than the two others, although they have provided similar conclusions about the null hypothesis.

The sample size affected the p-values of the three tests similarly. The accuracy of their inferences depends on the quality of the matrix $\mathbf{G}$ of (co)variance components, which might be related to the method of estimation.

Both procedures are invariant to scaling, allowing one to test either genetic covariance or correlation. Moreover, the test statistics presented can be applied in a more general way in order to test environmental covariances or other covariance components.

The F distribution can be reproduced by simulating Wishart matrices and bootstrapping experimental data. Thus, an intuitive test of genetic covariance can be build based on mean cross-products ratio.

Tests based on mean cross-products ratio are related to both the value of the genetic covariance and the magnitude of the latter relative to the residual covariance. In addition, both approaches (bootstrap and simulation) are more sensitive than the tests based on Wilks, Pillai statistics and null correlation PDF.

The performance of the tests based on mean cross-products ratio is related to the quality of the original data set in terms of the MANOVA assumptions. Moreover, the test statistic does not depend on the estimation of the matrix $\mathbf{\Sigma}_G$.

The test based on simulation of Wishart matrices is easier to implement, especially when dealing with complex experimental designs. Furthermore, it should be preferred when dealing with small sample sizes, say $n < 20$ experimental units.

When analyzing a large data set, 256 progenies and 12 environments, the mean cross-products ratios showed approximation to a Normal distribution.

5.2 Future directions

Here it was developed R functions that estimate $\mathbf{\Sigma}_{\boldsymbol{G}}$ through the method of moments. However, when dealing with unbalanced data, the method of restricted maximum likelihood (REML) must be preferred. For balanced data, it is known that both methods provide very similar estimates.

Development of a statistics based on Wilks' Lambda for testing genetic covariance taking into account the residual covariance.

Although the main interest in the tests presented here is elementwise, a p-value adjustment for multiple tests can be done through methods such as Bonferroni or Holm (HOLM, 1979).

Any concerns about the effects of other variables on the estimates of genetic covariances can be dealt with by using partial covariance matrices in place of the ordinary covariance matrices.

Power analysis for empirical tests may be developed following the approach presented by Silva et al. (2015), used to calculate the *simulated* power of Mantel's test.

Implementations in R language developed in this thesis should be published in the next version of the package **biotools** (SILVA, 2015).

REFERENCES

AJAYI, A. T.; ADEKOLA, M. O.; TAIWO, B. H.; AZUH, V. O. Character expression and differences in yield potential of ten genotypes of cowpea (*vigna unguiculata* L. Walp). **International Journal of Plant Research**, Rosemead, v. 4, p. 63–71, 2014.

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. 3rd. ed. New Jersey: John Wiley & Sons, 2003. 372 p.

AROIAN, L. A. The probability function of a product of two normal distributed variables. **The Annals of Mathematical Statistics**, Beachwood, v. 18, p. 256–271, 1947.

AROIAN, L. A.; TANEJA, V. S.; CORNWELL, L. W. Mathematical forms of the distribution of the product of two normal variables. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 7, p. 164–172, 1978.

BARROS, L. B.; MOREIRA, R. M. P.; FERREIRA, J. M. Phenotypic, additive genetic and environment correlations of maize landraces populations in family farm systems. **Scientia Agricola**, Piracicaba, v. 67, p. 685–691, 2010.

BARTLETT, M. S. Multivariate analysis. **Journal of the Royal Statistical Society B**, Chichester, v. 9, p. 176–197, 1947.

BERRY, K. J.; MIELKE-JR, P. W.; MIELKE, H. W. The fisher-pitman permutation test: an attractive alternative to the F test. **Psychological Reports**, Missoula, v. 90, p. 495–502, 2002.

BLOWS, M. W.; MCGUIGAN, K. The distribution of genetic variance across phenotypic space and the response to selection. **Molecular Ecology**, Malden, p. n/a–n/a, 2014.

BOIK, R. J. The fisher-pitman permutation test: A non-robust alternative to the normal theory f test when variances are heterogeneous. **British Journal of Mathematical and Statistical Psychology**, Chichester, v. 40, p. 26–42, 1987.

BOWMAN, A. W.; AZZALINI, A. **Applied smoothing techniques for data analysis: the kernel approach with S-plus illustrations**. New York: Oxford University Press, 1997. 204 p.

BOX, G. E. P.; MULLER, M. E. A note on the generation of random normal deviates. **The Annals of Mathematical Statistics**, Beachwood, v. 29, p. 610–611, 1958.

BROMILEY, P. A. Products and convolutions of gaussian probability density functions. **Tina-Vision Memo**, Manchester, v. 2003-003, p. n/a–n/a, 2003. University of Manchester.

BROWN, G. H. An empirical study of the distribution of the sample genetic correlation coefficient. **Biometrics**, Washington, v. 25, p. 63–72, 1969.

CAFLISCH, R. E. Monte carlo and quasi-monte carlo methods. **Acta Numerica**, Cambridge, v. 7, p. 1–49, 1998.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A. S. G. Bootstrap method applied over resampling levels in the estimation of genetic parameters of populations. **Scientia Agricola**, Piracicaba, v. 58, p. 785–793, 2001. (In Portuguese, with abstract in English).

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2nd. ed. Stanford: Duxbury Press, 2002. 660 p.

CHEVERUD, J. A comparison of genetic and phenotypic correlations. **Evolution**, Missouri, v. 42, p. 958–968, 1988.

COCHRAN, W. G.; COX, G. M. **Experimental design**. 2nd. ed. New York: John Wiley & Sons, 1966. 640 p.

COOK, R. D. On the mean and variance of the generalized inverse of a singular wishart matrix. **Electronic Journal of Statistics**, Beachwood, v. 5, p. 146–158, 2011.

CRAIG, C. C. On the frequency function of xy. **The Annals of Mathematical Statistics**, Beachwood, v. 7, p. 1–15, 1936.

DENG, H.; WICKHAM, H. **Density estimation in R**. [S.l.], 2011. Disponível em: <http://vita.had.co.nz/papers/density-estimation.html>.

DUTILLEUL, P.; CARRIERE, Y. Among-environment heterocedasticity and the estimation and testing of genetic correlation. **Heredity**, New York, v. 80, p. 403–4013, 1998.

EFRON, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics**, Beachwood, v. 7, p. 1–26, 1979.

EFRON, B.; TIBISHIRANI, R. J. **An introduction to the bootstrap**. New York: Chapman & Hall/CRC, 1993. 456 p.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4th. ed. Edingburgh: Longman Group Limited, 1996. 480 p.

FERREIRA, A.; CRUZ, C. D.; VASCONCELOS, E. S.; NASCIMENTO, M.; RIBEIRO, M. F.; SILVA, M. F. Use of the nonparametric bootstrap in evaluating genotypic, phenotypic and environmental correlations. **Acta Scientiarum Agronomy**, Maringa, v. 30, p. 657–663, 2008. (In Portuguese, with abstract in English).

FISHER, R. A. **The Design of Experiments**. Edingburgh: Oliver & Boyd, 1935. 250 p.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference**. 2nd. ed. London: Chapman and Hall/CRC, 2006. 242 p.

GAUTHIER, P.; POSSAMAI, D. **Efficient Simulation of the Wishart Model**. Available at SSRN, 2009. Disponível em: <http://ssrn.com/abstract=1474728>.

GELMAN, A.; CARLIN, J.; STERN, H.; RUBIN, D. **Bayesian data analysis**. London: CRC Press, 2004. 675 p.

GHOSH, M.; SINHA, B. K. A. Simple derivation of the wishart distribution. **The American Statistician**, Alexandria, v. 56, p. 100–101, 2002.

GOLUB, G. H.; LOAN, C. F. V. **Matrix Computations**. 3rd. ed. Baltimore: The Johns Hopkins University Press, 1996. 728 p.

GOOD, P. I. **Resampling methods: a pratical guide to data analysis**. 3rd. ed. Boston: Birkhauser, 2006. 218 p.

GOODNIGHT, C. J.; SCHWARTZ, J. M. A bootstrap comparison of genetic covariance matrices. **Biometrics**, Washington, v. 53, p. 1026–1039, 1997.

GUILLAUME, F.; WHITLOCK, M. C. Effects of migration on the genetic covariance matrix. **Evolution**, Missouri, v. 61, p. 2398–2409, 2007.

HENDERSON, C. R. Best linear unbiased prediction using relationship matrices derived from selected base populations. **Journal Dairy Science**, Champaign, v. 68, p. 443–448, 1985.

HILL, W. G.; THOMPSON, R. Probabilities of non-positive definite between-group or genetic covariance matrices. **Biometrics**, Washington, v. 34, p. 429–439, 1978.

HOJSGAARD, S.; HALEKOH, U.; ROBISON-COX, J.; WRIGHT, K.; LEIDI, A. A. **doBy: doBy - Groupwise summary statistics, general linear contrasts, population means (least-squares-means), and other utilities. R package version 4.5-5**. Vienna, Austria, 2012. Disponível em: <http: //cran.r-project.org/package=doBy>.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics**, v. 6, p. 65–70, 1979.

JOHNSON, N. L.; KOTZ, S. **Distributions in statistics: continuous multivariate distributions**. New York: John Wiley & Sons, 1972. 333 p.

KEMPTHORNE, O. **An introduction to genetic statistics**. Ames: Iowa State University Press, 1969. 545 p.

KOMINAKIS, A. P. Phenotypic correlations as substitutes to genetic correlations in dairy sheep and goats. **Journal of Animal Breeding and Genetics**, Malden, v. 120, p. 268–281, 2003.

KRZANOWSKI, W. J. **Principles of multivariate analysis: a user's perspective**. 2nd. ed. New York: Oxford University Press, 2000. 612 p.

LANDAU, D. P.; BINDER, K. **A Guide to Monte Carlo Simulations in Statistical Physics**. 3rd. ed. Cambridge: Cambridge University Press, 2009. 488 p.

LANDE, R. The genetic covariance between characters maintained by pleiotropic mutations. **Genetics**, Bethesda, v. 94, p. 203–215, 1980.

LEMOS, M. A.; GAMA, E. E. G.; OLIVEIRA, A. C.; ARAÃºJO, M. R. A. Genotypic, phenotypic and environmental correlations in progenies of corn. **Pesquisa AgropecuÃ¡ria Brasileira**, Brasilia, v. 27, p. 1563–1569, 1992. (In portuguese, with abstract in English).

LIU, B. H.; KNAPP, S. J.; BIRKES, D. Sampling distributions, biases, variances, and confidence intervals for genetic correlations. **Theoretical and Applied Genetics**, Heidelberg, v. 94, p. 9–19, 1997.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, 1998. 980 p.

MALIK, H. N.; MALIK, S. I.; HUSSAIN, M.; CHUGHTAI, S. U. R.; JAVED, H. I. Genetic correlation among various quantitative characters in maize (zea mays l.) hybrids. **Journal of Agriculture and Social Sciences**, Faisalabad, v. 1, p. 262–265, 2005.

MANGOLIN, C. A.; SOUZA-JR, C. L.; GARCIA, A. A. F.; GARCIA, A. F.; SIBOV, S. T.; SOUZA, A. P. Mapping qtls for kernel oil content in a tropical maize population. **Euphytica**, Dordrecht, v. 137, p. 251–259, 2004.

MANLY, B. F. J. **Randomization, Bootstrap and Monte Carlo Methods in Biology**. 3rd. ed. Boca Raton: Chapman & Hall/CRC, 2007. 480 p.

MANLY, B. F. J.; MCALEVEY, L.; STEVENS, D. A randomization procedure for comparing group means on multiple measurements. **British Journal of Mathematical and Statistical Psychology**, Leicester, v. 39, p. 183–189, 1986.

MANTEL, N. The detection of disease clustering and a generalized regression approach. **Cancer Research**, Philadelphia, v. 27, p. 209–220, 1967.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic Press, 1979. 521 p.

MEREDITH, W. R. **Quantitative genetics**. Madison: Cotton. Agron. Monogr. 24, 1984.

MEYER, K.; KIRKPATRICK, M. Better estimates of genetic covariance matrices by bending using penalized maximum likelihood. **Genetics**, Bethesda, v. 185, p. 1097–1110, 2010.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to linear regression analysis**. New York: John Wiley & Sons, 1982. 504 p.

MUNAWAR, M.; SHAHBAZ, M.; HAMMAD, G.; YASIR, M. Correlation and path analysis of grain yield components in exotic maize (zea mays l.) hybrids. **International Journal of Sciences: Basic and Applied Research**, v. 12, p. 22–27, 2013.

NEALE, M. C.; CARDON, L. R. **Methodology for Genetic Studies of Twins and Families**. Dordrecht: Springer, 1992. 496 p.

NIEDERREITER, H. **Random Number Generation and Quasi-Monte Carlo Methods**. Philadelphia: Society for Industrial and Applied Mathematics, 1992. 241 p.

PHAM-GIA, T. Exact distribution of the generalized wilks's statistic and applications. **Journal of Multivariate Analysis**, v. 99, p. 1698–1716, 2008.

PILLAI, K. C. S. Some new test criteria in multivariate analysis. **The Annals of Mathematical Statistics**, Beachwood, v. 26, p. 117–121, 1955.

PITMAN, E. J. G. Significance tests which may be applied to samples from any populations. **Supplement to the Journal of the Royal Statistical Society**, London, v. 4, p. 119–130, 1937.

PROA, M.; O'HIGGINS, P.; MONTEIRO, L. R. Type i error rates for testing genetic drift with phenotypic covariance matrices: a simulation study. **Evolution**, Missouri, v. 67, p. 185–195, 2012.

QUINN, G. P.; KEOUGH, M. J. **Experimental design and data analysis for biologists**. Cambridge: Cambridge University Press, 2010. 553 p.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

RAO, C. R. **Advanced statistical methods in biometric research**. New York: John Wiley & Sons, 1952. 390 p.

REEVE, E. C. R. The variance of the genetic correlation coefficient. **Biometrics**, Washington, v. 11, p. 357–374, 1955.

REIS, R. L.; M.UNIZ, J. A.; SILVA, F. F.; AQUINO, L. H. Additive genetic variance estimates in selected and unselected populations via monte carlo simulation using the r software. **Ciencia e Agrotecnologia**, Lavras, v. 33, p. 285–291, 2009.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informacão Tecnológica, 2002. 975 p.

REVELL, L. J.; MAHLER, D. L.; SWEENEY, J. R.; SOBOTKA, M.; FANCHER, V. E.; LOSOS, J. B. Nonlinear selection and the evolution of variances and covariances for continuous characters in an anole. **Journal of Evolutionary Biology**, St Andrews, v. 23, p. 407–421, 2010.

ROBERTSON, A. The sampling variance of the genetic correlation coefficient. **Biometrics**, Washington, v. 15, p. 469–485, 1959.

ROFF, D. A. The estimation of genetic correlations from phenotypic correlations: a test of cheverud's conjecture. **Heredity**, New York, v. 74, p. 481–490, 1995.

SCHATZOFF, M. Exact distributions of wilks's likelihood ratio criterion. **Biometrika**, Oxford, v. 53, p. 347–358, 1966.

SCHORK, N. J. Power calculations for genetic association studies using estimated probability distributions. **The American Journal of Human Genetics**, Houston, v. 70, p. 1480–1489, 2002.

SEGAL, M. R.; WIEMELS, J. L. Clustering of translocation breakpoints. **Journal of the American Statistical Association**, Alexandria, v. 97, p. 66–76, 2002.

SHEATHER, S. J. Density estimation. **Statistical Science**, Beachwood, v. 19, p. 588–597, 2004.

SILVA, A. R. **biotools: Tools for Biometry and Applied Statistics in Agricultural Science. R package version 2.0**. Vienna, Austria, 2015. Disponível em: <http://CRAN.R-project.org/package=biotools>.

SILVA, A. R.; DIAS, C. T. S.; CECON, P. R.; REGO, E. R. An alternative procedure for performing a power analysis of mantel's test. **Journal of Applied Statistics**, Leeds, v. 42, p. 1–9, 2015.

SIMONOFF, J. S. **Smoothing methods in statistics**. New York: Springer-Verlag, 1996. 340 p.

SPRINGER, M. D.; THOMPSON, W. E. The distribution of products of beta, gamma and gaussian random variables. **SIAM Journal on Applied Mathematics**, Philadelphia, v. 18, p. 721–737, 1970.

TAYLOR, C. C. Bootstrap choice of smoothing parameters in kernel density estimation. **Biometrika**, Oxford, v. 76, p. 705–712, 1989.

TENGAN, K. L. M.; OBENG-ANTWI, K.; AKROMAH, R. Genetic variances, heritability, and correlation studies on selected phenotypic traits in a backcross breeding program involving normal and opaque-2 maize. **Agriculture and Biology Journal of North America**, Milford, v. 3, p. 287–291, 2012.

TIMM, N. H. **Applied multivariate analysis**. New York: Springer, 2002. 800 p.

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. Ribeirão Preto: Sociedade Brasileira de Genética, 1992. 486 p.

VESTER, A.; WAAL, D. D. **Modelling the SOI volatility through the Wishart distribution @ONLINE**. 2015. Disponível em: <http://natagri.ufs.ac.za/dl/userfiles/Documents/00002/2093_eng.pdf>.

VIANA, M. L.; SUHRING, S.; MANLY, B. F. J. Application of randomization methods to study the association of *trichocereus pasacana (cactaceae)* with potential nurse plants. **Plant Ecology**, Dordrecht, v. 00, p. 1–5, 2000.

WAITT, D. E.; LEVIN, D. A. Genetic and phenotypic correlations in plants: a botanical test of cheverud's conjecture. **Heredity**, New York, v. 80, p. 310–319, 1998.

WALLER, L. A.; SMITH, D.; CHILDS, J. E.; REAL, L. A. Monte carlo assessments of goodness-of-fit for ecological simulation models. **Ecological Modelling**, v. 164, p. 49–63, 2003.

WAND, M. P.; JONES, M. C. **Kernel smoothing**. London: Chapman & Hall, 1995. 224 p.

WARE, R.; FRANK, L. A. D. Approximating the distribution for sums of products of normal variables. **Tech. Rep. UCDMS**, Canterbury, v. 15, p. n/a–n/a, 2003.

WEISSTEIN, E. W. **Correlation Coefficient–Bivariate Normal Distribution @ONLINE**. MathWorld–A Wolfram Web Resource, 2015. Disponível em: <http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html>.

WILKS, S. S. Certain generalizations in the analysis of variance. **Biometrika**, Oxford, v. 24, p. 471–494, 1932.

WILKS, S. S. On the independence of k sets of normally distributed statistical variables. **Econometrica**, New York, v. 3, p. 309–326, 1935.

WISHART, J. The generalized product moment distributions in samples from a normal multivariate population. **Biometrika**, Oxford, v. 20A, p. 32–52, 1928.

WOOTTON, R. J.; SMITH, C. **Reproductive biology of Teleost fishes**. New York: John Wiley & Sons, 2014. 496 p.

WORBY, C. J.; CHANG, H.; HANAGE, W. P.; LIPSITCH, M. The distribution of pairwise genetic distances: A tool for investigating disease transmission. **Genetics**, Bethesda, v. 198, p. 1395–1404, 2014.

YAMADA, Y. Genotype by environment interaction and genetic correlation of the same trait under different environments. **Japanese Journal of Genetics**, Shizuoka, v. 37, p. 498–509, 1962.

ZENG, L.; MEREDITH, E. R.; BOYKIN, D. L.; TALIERCIO, E. Evaluation of an exotic germplasm population derived from multiple crosses among *gossypium* tetraploid species. **The Journal of Cotton Science**, Baton Rouge, v. 11, p. 118–127, 2007.

APPENDICES

Appendix A - R code for performing analysis of variance according to a square lattice design, per environment, and to determine the least square means adjusted for block effects

```
> dados <- read.table("lattice16.csv", sep=";", header=TRUE)
> dados$fAmb <- factor(dados$Amb)
> dados$fRep <- factor(dados$Rep)
> dados$fBloco <- factor(dados$Bloco)
> dados$fTrat <- factor(dados$Trat)
> dados$PGCHA <- dados$PG * ((100 - dados$Umidade)/27040)

> require(car)
> require(doBy)

> I <- nlevels(dados$fAmb)
> J <- nlevels(dados$fTrat)
> nvar <- 12
> ls.mat <- matrix(NA, nrow=I*J, ncol=nvar)
> colnames(ls.mat) <- colnames(dados[,7:18])
> for(p in 1:nvar) {
+    for(i in 1:I) {
+       ls.mat[((i-1)*J+1):((i-1)*J+J), p] <- popMeans(aov(dados[dados$fAmb==i, p+6] ~
+       fRep + fRep/fBloco + fTrat,data = dados[dados$fAmb == i, ]),
+       effect = "fTrat")$Estimate
+    }
+ }
> trat <- rep(1:J, times = I)
> amb <- rep(1:I, each = J)
> write.table(x = cbind(amb, trat, ls.mat), file = "mediasGxA.csv",
+    sep = ";", row.names = FALSE)
```

Appendix B - R code for performing joint multivariate analysis of variance

```
> medias <- read.table("mediasGxA.csv", sep=";", header=TRUE)
> medias$f.env <- factor(medias$env)
> medias$f.prog <- factor(medias$progeny)

> M1 <- manova(cbind(NE, W500, LE, DE, DC, NROWS, NGROW, YP, PH, EH, MF, FF) ~
+   f.env + f.prog, data = medias)

> summary(M1, test = "Wilks")
            Df      Wilks approx F num Df den Df    Pr(>F)
f.env       11 0.00113222  223.743    132  22887 < 2.2e-16 ***
f.prog     255 0.00023922   11.016   3060  33546 < 2.2e-16 ***
Residuals 2805
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# testing for residual normality
> require(mvShapiroTest)
> mvShapiro.Test(residuals(M1))

        Generalized Shapiro-Wilk test for Multivariate Normality by
        Villasenor-Alva and Gonzalez-Estrada

data:  residuals(M1)
MVW = 0.9755, p-value < 2.2e-16
```

Appendix C - Function in R language for collinearity diagnosis

```
conditionNumber <-
function(m)
{
    eigval <- svd(m)$d
    cn <- max(eigval) / min(eigval)
    meaning <- NULL
    if (cn < 100) {
       meaning <- "weak collinearity"
    } else if (cn > 1000) {
       meaning <- "severe collinearity"
    } else {
       meaning <- "moderate to severe collinearity"
    }
    attr(cn, "meaning") <- meaning
    return(cn)
}
```

Appendix D - Function in R language for testing components of covariance via the *partial* Wilks' and Pillai's statistics

```
gencovtest1 <-
function(obj, geneticFactor, gcov = NULL,
    residualFactor = NULL, correctionFactor = 1,
    test = c("Wilks", "Pillai"))
{
    if (!inherits(obj, "manova"))
       stop("'obj' must be of class 'manova'")
    manov <- anova(obj)
    model <- obj$model
    SS <- summary(obj)$SS
    stopifnot(geneticFactor %in% names(model))
    dfg <- manov[geneticFactor, "Df"]
    Mg <- SS[[geneticFactor]] / dfg
    if (is.null(residualFactor)) {
        dfe <- df.residual(obj)
        Me <- SS[["Residuals"]] / dfe
    } else {
        stopifnot(residualFactor %in% names(model))
        dfe <- manov[residualFactor, "Df"]
        Me <- SS[[residualFactor]] / dfe
    }
    test <- match.arg(test)
    nvar <- nrow(Mg)
    if (is.null(gcov)) {
        nrep <- nrow(model) / nlevels(model[[geneticFactor]])
        gcov <- (Mg - Me) / (nrep * correctionFactor)
    } else {
        if (nrow(gcov) != nvar || ncol(gcov) != nvar)
           stop("'gcov' presents incompatible dimensions")
    }
    gcor <- cov2cor(gcov)

    # collinearity diagnosis
    cn <- conditionNumber(gcov)
    if (cn > 100) {
       mess <- paste("the genetic covariance matrix presents",
          attr(cn, "meaning"))
       warning(mess)
    }

    # Aux. function for Wilks' Lambda and Pillai's Tn
    teststat <- function(m, var1, var2)
    {
       m2 <- m[c(var1, var2), c(var1, var2)]
       Lambda <- det(m2) / prod(diag(m2))
       Tn <- m2[1, 2]^2 / prod(diag(m2))
       out <- list(Lambda = Lambda, Tn = Tn)
       return(out)
    }

    # test statistics, chi-sq approx and p-values
    pval <- matrix(NA, nvar, nvar,
       dimnames = list(rownames(Mg), colnames(Mg)))
    stat <- X2 <- pval
    for(i in 1:nvar){
```

```
      for(j in 1:nvar){
         if (i != j) {
            if (test == "Wilks") {
               stat[i, j] <- teststat(gcov, i, j)$Lambda
               X2[i, j] <- -dfg * log(stat[i, j])
            } else {
               stat[i, j] <- teststat(gcov, i, j)$Tn
               X2[i, j] <- dfg * stat[i, j]
            }
            pval[i, j] <- pchisq(X2[i, j], 1, lower.tail = FALSE)
         }
      }
   }

   # output
   out <- list(gcov = gcov, gcor = gcor,
      test = test, statistics = stat, X2 = X2,
      p.value = pval)
   class(out) <- "gencovtest"
   return(out)
}

# ---------------------------
# print method
print.gencovtest <-
function(x, digits = 4, ...)
{
    cat("\n          Genetic Covariance Test \n")
    cat("\nGenetic (Co)variances and Correlations (upper triangular):\n")
    GR <- lower.tri(x$gcov, diag = TRUE) * x$gcov +
       upper.tri(x$gcor) * x$gcor
    print(round(GR, digits))
    cat("\nChi-Sq (df = 1) approx. (", x$test, ") and
   p-values (upper triangular):\n", sep = "")
    X2.p <- lower.tri(x$X2, diag = TRUE) * x$X2 +
       upper.tri(x$p.value) * x$p.value
    print(round(X2.p, digits))
    invisible(x)
}
```

Appendix E - Test results for the experimental maize data using the *partial* Wilks' and Pillai's statistics

```
> wil <- gencovtest1(M1, "f.prog", test = "Wilks")
Warning message:
In gencovtest1(M1, "f.prog", test = "Wilks") :
  the genetic covariance matrix presents severe collinearity
> wil

          Genetic Covariance Test

Genetic (Co)variances and Correlations (upper triangular):
            NE      W500       LE      DE      DC   NROWS    NGROW         YP      PH      EH      MF      FF
NE      6.5256    0.2177   0.4884  0.3033  0.0389 -0.0289   0.6764     0.8267  0.4145  0.4251 -0.5816 -0.6783
W500    6.2075  124.6141   0.4378  0.2981  0.1341 -0.3265   0.1691     0.3735  0.3517  0.2227 -0.3492 -0.3247
LE      0.9888    3.8738   0.6282  0.2467  0.1363 -0.0260   0.7244     0.6348  0.4383  0.3855 -0.5048 -0.4818
DE      0.1545    0.6637   0.0390  0.0398  0.8272  0.6410   0.5267     0.6919  0.2860  0.2693 -0.5829 -0.5397
DC      0.0124    0.1876   0.0135  0.0207  0.0157  0.6785   0.2181     0.3957  0.2074  0.2159 -0.4115 -0.3595
NROWS  -0.0686   -3.3855  -0.0191  0.1187  0.0790  0.8629   0.1565     0.2726  0.0565  0.1292 -0.2188 -0.1948
NGROW   4.1364    4.5202   1.3744  0.2515  0.0654  0.3480   5.7315     0.8510  0.3335  0.3453 -0.6548 -0.6449
YP    625.4283 1234.7471 149.0034 40.8635 14.6868 74.9947 603.4157 87712.2529  0.4688  0.4536 -0.7596 -0.7848
PH      0.0944    0.3501   0.0310  0.0051  0.0023  0.0047   0.0712    12.3797  0.0080  0.8126 -0.1804 -0.2034
EH      0.0675    0.1545   0.0190  0.0033  0.0017  0.0075   0.0514     8.3487  0.0045  0.0039 -0.0728 -0.0943
MF     -2.9878   -7.8388  -0.8046 -0.2338 -0.1037 -0.4087  -3.1527  -452.4214 -0.0324 -0.0091  4.0446  0.9549
FF     -4.1560   -8.6943  -0.9158 -0.2581 -0.1081 -0.4339  -3.7029  -557.4733 -0.0435 -0.0140  4.6060  5.7521

Chi-Sq (df = 1) approx. (Wilks) and p-values (upper triangular):
            NE    W500       LE       DE       DC   NROWS    NGROW       YP       PH     EH       MF     FF
NE          NA  0.0004   0.0000   0.0000   0.5347  0.6443   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
W500   12.3791      NA   0.0000   0.0000   0.0315  0.0000   0.0065   0.0000   0.0000 0.0003   0.0000 0.0000
LE     69.4860 54.2706       NA   0.0001   0.0288  0.6783   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
DE     24.6074 23.7370  16.0124       NA   0.0000  0.0000   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
DC      0.3854  4.6251   4.7809 294.0248       NA  0.0000   0.0004   0.0000   0.0008 0.0005   0.0000 0.0000
NROWS   0.2132 28.7416   0.1721 134.9498 157.3080      NA   0.0119   0.0000   0.3666 0.0382   0.0004 0.0017
NGROW 155.9297  7.4013 189.6706  82.8611  12.4296  6.3215       NA   0.0000   0.0000 0.0000   0.0000 0.0000
YP    293.2822 38.3073 131.5193 166.1326  43.4163 19.6903 328.5304       NA   0.0000 0.0000   0.0000 0.0000
PH     48.0748 33.6744  54.4010  21.7665  11.2152  0.8152  30.0702  63.2683       NA 0.0000   0.0037 0.0010
EH     50.8380 12.9763  41.0369  19.1979  12.1697  4.2959  32.3808  58.7428 275.3833     NA   0.2441 0.1314
MF    105.2736 33.1530  74.9890 105.8687  47.3167 12.5039 142.7902 219.3778   8.4407 1.3566       NA 0.0000
FF    157.1981 28.4173  67.3434  87.7908  35.2859  9.8605 137.1098 244.0426  10.7766 2.2758 619.4521     NA

> pil <- gencovtest1(M1, "f.prog", test = "Pillai")
Warning message:
In gencovtest1(M1, "f.prog", test = "Pillai") :
  the genetic covariance matrix presents severe collinearity
> pil

          Genetic Covariance Test

Genetic (Co)variances and Correlations (upper triangular):
            NE      W500       LE      DE      DC   NROWS    NGROW         YP      PH      EH      MF      FF
NE      6.5256    0.2177   0.4884  0.3033  0.0389 -0.0289   0.6764     0.8267  0.4145  0.4251 -0.5816 -0.6783
W500    6.2075  124.6141   0.4378  0.2981  0.1341 -0.3265   0.1691     0.3735  0.3517  0.2227 -0.3492 -0.3247
LE      0.9888    3.8738   0.6282  0.2467  0.1363 -0.0260   0.7244     0.6348  0.4383  0.3855 -0.5048 -0.4818
DE      0.1545    0.6637   0.0390  0.0398  0.8272  0.6410   0.5267     0.6919  0.2860  0.2693 -0.5829 -0.5397
DC      0.0124    0.1876   0.0135  0.0207  0.0157  0.6785   0.2181     0.3957  0.2074  0.2159 -0.4115 -0.3595
NROWS  -0.0686   -3.3855  -0.0191  0.1187  0.0790  0.8629   0.1565     0.2726  0.0565  0.1292 -0.2188 -0.1948
NGROW   4.1364    4.5202   1.3744  0.2515  0.0654  0.3480   5.7315     0.8510  0.3335  0.3453 -0.6548 -0.6449
YP    625.4283 1234.7471 149.0034 40.8635 14.6868 74.9947 603.4157 87712.2529  0.4688  0.4536 -0.7596 -0.7848
PH      0.0944    0.3501   0.0310  0.0051  0.0023  0.0047   0.0712    12.3797  0.0080  0.8126 -0.1804 -0.2034
EH      0.0675    0.1545   0.0190  0.0033  0.0017  0.0075   0.0514     8.3487  0.0045  0.0039 -0.0728 -0.0943
MF     -2.9878   -7.8388  -0.8046 -0.2338 -0.1037 -0.4087  -3.1527  -452.4214 -0.0324 -0.0091  4.0446  0.9549
FF     -4.1560   -8.6943  -0.9158 -0.2581 -0.1081 -0.4339  -3.7029  -557.4733 -0.0435 -0.0140  4.6060  5.7521

Chi-Sq (df = 1) approx. (Pillai) and p-values (upper triangular):
            NE    W500       LE       DE       DC   NROWS    NGROW       YP       PH     EH       MF     FF
NE          NA  0.0005   0.0000   0.0000   0.5349  0.6443   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
W500   12.0834      NA   0.0000   0.0000   0.0323  0.0000   0.0069   0.0000   0.0000 0.0004   0.0000 0.0000
LE     60.8231 48.8843       NA   0.0001   0.0295  0.6783   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
DE     23.4574 22.6657  15.5200       NA   0.0000  0.0000   0.0000   0.0000   0.0000 0.0000   0.0000 0.0000
DC      0.3851  4.5834   4.7364 174.5026       NA  0.0000   0.0005   0.0000   0.0009 0.0006   0.0000 0.0000
NROWS   0.2131 27.1810   0.1720 104.7879 117.3974      NA   0.0125   0.0000   0.3670 0.0390   0.0005 0.0019
NGROW 116.6517  7.2950 133.7982  70.7455  12.1315  6.2437       NA   0.0000   0.0000 0.0000   0.0000 0.0000
```

YP	174.2678	35.5688	102.7534	122.0779	39.9214	18.9492	184.6903	NA	0.0000	0.0000	0.0000	0.0000
PH	43.8149	31.5457	48.9897	20.8634	10.9722	0.8139	28.3649	56.0303	NA	0.0000	0.0040	0.0012
EH	46.0910	12.6517	37.9051	18.4930	11.8839	4.2599	30.4092	52.4677	168.3975	NA	0.2447	0.1323
MF	86.2488	31.0883	64.9687	86.6421	43.1861	12.2023	109.3361	147.1266	8.3025	1.3530	NA	0.0000
FF	117.3381	26.8911	59.1847	74.2733	32.9533	9.6722	106.0549	157.0719	10.5521	2.2657	232.5332	NA

Appendix F - Function in R language for testing genetic correlation using the null correlation PDF

```
ect <-
function(obj, geneticFactor,
    residualFactor = NULL, correctionFactor = 1,
    alternative = c("two.sided", "less", "greater"))
{
    alternative <- match.arg(alternative)
    if (!inherits(obj, "manova"))
       stop("'obj' must be of class 'manova'")
    manov <- anova(obj)
    model <- obj$model
    SS <- summary(obj)$SS
    stopifnot(geneticFactor %in% names(model))
    dfg <- manov[geneticFactor, "Df"]
    Mg <- SS[[geneticFactor]] / dfg
    if (is.null(residualFactor)) {
        dfe <- df.residual(obj)
        Me <- SS[["Residuals"]] / dfe
    } else {
        stopifnot(residualFactor %in% names(model))
        dfe <- manov[residualFactor, "Df"]
        Me <- SS[[residualFactor]] / dfe
    }
    nrep <- nrow(model) / nlevels(model[[geneticFactor]])
    gcov <- (Mg - Me) / (nrep * correctionFactor)
    gcor <- cov2cor(gcov)
    nvar <- nrow(Mg)

    # r PDF
    fr <- function(r, Df) ((1-r^2)^((Df-2)/2)) / beta(.5, Df/2)

    # p-values
    pval <- matrix(NA, nvar, nvar,
       dimnames = list(rownames(gcov), colnames(gcov)))
    if (alternative == "two.sided") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- 2 * integrate(fr, Df = dfg,
                abs(gcor[i, j]), 1)$value
          }
       }
    } else if (alternative == "less") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- integrate(fr, Df = dfg, -1,
                gcor[i, j])$value
          }
       }
    } else if (alternative == "greater") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- integrate(fr, Df = dfg,
                gcor[i, j], 1)$value
          }
       }
    }
```

```
    # output
    out <- list(gcor = gcor, p.values = pval,
       alternative = alternative)
    class(out) <- "ect"
    return(out)
}

# ----------------------------
# print method
print.ect <-
function(x, digits = 4, ...)
{
    cat("\n          Exact Correlation Test \n")
    cat("\nGenetic Correlations and p-values (upper triangular):\n")
    Rp <- lower.tri(x$gcor, diag = TRUE) * x$gcor +
       upper.tri(x$p.values) * x$p.values
    print(round(Rp, digits))
    cat("\nAlternative hypothesis:", x$alternative, "\n")
    invisible(x)
}
```

Appendix G - Function in R language for testing components of covariance via mean cross-products ratio based on bootstrap resampling of a experimental data obtained from an experiment carried out under a randomized block design

```
gencovtest2 <-
function(obj, geneticFactor, repFactor,
    correctionFactor = 1, nboot = 400,
    alternative = c("two.sided", "less", "greater"))
{
    stopifnot(geneticFactor %in% names(obj$model))
    stopifnot(repFactor %in% names(obj$model))
    alternative <- match.arg(alternative)
    if (!inherits(obj, "manova"))
       stop("'obj' must be of class 'manova'")
    model <- obj$model
    SS <- summary(obj)$SS
    dfg <- anova(obj)[geneticFactor, "Df"]
    Mg <- SS[[geneticFactor]] / dfg
    dfe <- df.residual(obj)
    Me <- SS[["Residuals"]] / dfe
    ratio <- Mg / Me
    p <- nvar <- ncol(Mg)
    nrep <- nrow(model) / nlevels(model[[geneticFactor]])
    gcov <- (Mg - Me) / (nrep * correctionFactor)
    gcor <- cov2cor(gcov)

    # objects for bootstrapping
    GenLev <- levels(model[[geneticFactor]])
    nGen <- length(GenLev)
    EnvLev <- levels(model[[repFactor]])
    nEnv <- length(EnvLev)
    mBoot <- matrix(nrow = nboot, ncol = nGen)
    mcpG <- mcpR <- array(dim = c(p, p, nboot))
    fgen <- factor(rep(1:nGen, nEnv))
    frep <- gl(nEnv, nGen)

    # bootstrap
    pb <- winProgressBar(title = "Genetic Covariance Test via Bootstrap",
       label = "RESAMPLING PROGRESS", min = 0, max = nboot, width = 300L)
    i = 1
    repeat{
       mBoot[i, ] <- sample(GenLev, replace = TRUE)
       Ar1 <- array(dim = c(nEnv, p, nGen))
       for(j in 1:nGen) {
          Ar1[,, j] <- subset(model,
             model[[geneticFactor]] == mBoot[i, j])[[1]]
       }
       Yboot <- apply(Ar1, 2, rbind)
       Mboot <- manova(Yboot ~ frep + fgen)
       mcpG[,, i] <- summary(Mboot)$SS[["fgen"]] / dfg
       mcpR[,, i] <- summary(Mboot)$SS[["Residuals"]] / dfe
       setWinProgressBar(pb, i,
          label = sprintf("RESAMPLING PROGRESS (%.0f%%)", 100*i/nboot))
       i <- i + 1
       if (i > nboot) break()
    }
    stat <- windata(mcpG / mcpR, p = 0.01)
```

```
    colnames(stat) <- rownames(stat) <- colnames(Mg)

    # p-values
    pval <- matrix(NA, nvar, nvar,
       dimnames = list(rownames(Mg), colnames(Mg)))
    if (alternative == "two.sided") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             if (i != j) {
                pval[i, j] <- mean(stat[i, j, ] -
                   median(stat[i, j, ]) >= abs(ratio[i, j]) -
                   median(stat[i, j, ])) +
                   mean(stat[i, j, ] -
                   median(stat[i, j, ]) <= -abs(ratio[i, j]) -
                   median(stat[i, j, ]))
             } else {
                pval[i, j] <- mean(stat[i, j, ] >= ratio[i, j])
             }
          }
       }
       pval[pval > 1] <- 1
    } else if (alternative == "less") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- mean(stat[i, j, ] <= ratio[i, j])
          }
       }
       diag(pval) <- 1
    } else if (alternative == "greater") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- mean(stat[i, j, ] >= ratio[i, j])
          }
       }
    }

    # output
    out <- list(gcov = gcov, ratio = ratio,
       bootratio = stat, p.values = pval,
       alternative = alternative, dfg = dfg, dfe = dfe)
    class(out) <- "bootTest"
    Sys.sleep(0.5)
    close(pb)
    return(out)
}

# ------------------------------
# winsorized data
windata <-
function(x, p)
{
    if(length(p) != 1 || p < 0 || p > 0.5)
       stop('"p" deve ser um valor entre 0 e 0.5!')
    qx <- quantile(x, c(p, 1-p))
    x[x < qx[1]] <- qx[1]
    x[x > qx[2]] <- qx[2]
    return(x)
}
```

```
# ------------------------------
# print method
print.bootTest <-
function(x, digits = 4, ...)
{
    cat("\n          Genetic Covariance Test via Bootstrap \n")
    cat("\nGenetic (Co)variances and Correlations (upper triangular):\n")
    gcor <- cov2cor(x$gcov)
    GR <- lower.tri(x$gcov, diag = TRUE) * x$gcov +
       upper.tri(gcor) * gcor
    print(round(GR, digits))
    nsim <- dim(x$bootratio)[3L]
    cat("\nMean Sq and Cross-Prods Ratios and p-values (upper triangular) \nbased on",
        nsim, "estimates:\n")
    ratP <- lower.tri(x$ratio, diag = TRUE) * x$ratio +
       upper.tri(x$p.values) * x$p.values
    print(round(ratP, digits))
    cat("\nAlternative hypothesis:", x$alternative, "\n")
    invisible(x)
}
```

Appendix H - Function in R language for testing components of covariance via mean cross-products ratio based on Monte Carlo simulation

```
gencovtest3 <-
function(obj, geneticFactor,
residualFactor = NULL, correctionFactor = 1,
nsim = 9999, alternative = c("two.sided", "less", "greater"))
{
    if (!inherits(obj, "manova"))
       stop("'obj' must be of class 'manova'")
    alternative <- match.arg(alternative)
    manov <- anova(obj)
    model <- obj$model
    SS <- summary(obj)$SS
    stopifnot(geneticFactor %in% names(model))
    dfg <- manov[geneticFactor, "Df"]
    Mg <- SS[[geneticFactor]] / dfg
    if (is.null(residualFactor)) {
        dfe <- df.residual(obj)
        Me <- SS[["Residuals"]] / dfe
    } else {
        stopifnot(residualFactor %in% names(model))
        dfe <- manov[residualFactor, "Df"]
        Me <- SS[[residualFactor]] / dfe
    }
    nvar <- nrow(Mg)
    ratio <- Mg / Me
    nrep <- nrow(model) / nlevels(model[[geneticFactor]])
    gcov <- (Mg - Me) / (nrep * correctionFactor)
    gcor <- cov2cor(gcov)

    # Wishart simulation
    WG <- rWishart(nsim, dfg, Me) / dfg
    WR <- rWishart(nsim, dfe, Me) / dfe
    stat <- windata(WG / WR, p = 0.01)
    dimnames(stat) = list(rownames(Mg), colnames(Mg), NULL)

    # p-values
    pval <- matrix(NA, nvar, nvar,
       dimnames = list(rownames(Mg), colnames(Mg)))
    if (alternative == "two.sided") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             if (i != j) {
                pval[i, j] <- mean(stat[i, j, ] -
                   median(stat[i, j, ]) >= abs(ratio[i, j]) -
                   median(stat[i, j, ])) +
                   mean(stat[i, j, ] -
                   median(stat[i, j, ]) <= -abs(ratio[i, j]) -
                   median(stat[i, j, ]))
             } else {
                pval[i, j] <- mean(stat[i, j, ] >= ratio[i, j])
             }
          }
       }
       pval[pval > 1] <- 1
    } else if (alternative == "less") {
       for(i in 1:nvar) {
```

```
          for(j in 1:nvar) {
             pval[i, j] <- mean(stat[i, j, ] <= ratio[i, j])
          }
       }
       diag(pval) <- 1
    } else if (alternative == "greater") {
       for(i in 1:nvar) {
          for(j in 1:nvar) {
             pval[i, j] <- mean(stat[i, j, ] >= ratio[i, j])
          }
       }
    }

    # output
    out <- list(gcov = gcov, ratio = ratio,
       simratio = stat, p.values = pval,
       alternative = alternative,
       dfg = dfg, dfe = dfe)
    class(out) <- "genCovTest"
    return(out)
}

# -------------------------
# winsorized data
windata <-
function(x, p)
{
    if(length(p) != 1 || p < 0 || p > 0.5)
       stop('"p" deve ser um valor entre 0 e 0.5!')
    qx <- quantile(x, c(p, 1-p))
    x[x < qx[1]] <- qx[1]
    x[x > qx[2]] <- qx[2]
    return(x)
}

# -------------------------
# print method
print.genCovTest <-
function(x, digits = 4, ...)
{
    cat("\n          Genetic Covariance Test via Wishart Simulation \n")
    cat("\nGenetic (Co)variances and Correlations (upper triangular):\n")
    gcor <- cov2cor(x$gcov)
    GR <- lower.tri(x$gcov, diag = TRUE) * x$gcov +
       upper.tri(gcor) * gcor
    print(round(GR, digits/2))
    nsim <- dim(x$simratio)[3L]
    cat("\nMean Sq and Cross-Prods Ratios and p-values (upper triangular) \nbased on",
       nsim, "estimates:\n")
    ratP <- lower.tri(x$ratio, diag = TRUE) * x$ratio +
       upper.tri(x$p.values) * x$p.values
    print(round(ratP, digits))
    cat("\nAlternative hypothesis:", x$alternative, "\n")
    invisible(x)
}
```