Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Precision of Principal Component Analysis in a Larger Phenotypic Data Sets of Maize

Alberto Cargnelutti Filho (alberto.cargnelutti.filho@gmail.com)
Federal University of Santa Maria (UFSM)
Marcos Toebe
Federal University of Santa Maria (UFSM)

Research Article

Keywords: Zea mays, multivariate analysis, resampling.

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1321792/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

In maize, research with multiple variables has been carried out and the principal component analysis (PCA) has been used in order to reduce the data dimensionality. The objective of this work was to study the impact of observations number and correlation matrix coefficients (r) magnitude in the eigenvalue estimates precision of the PCA, with simulated and real larger phenotypic data sets of maize. Seven data files were simulated, formed by twelve variables and 6340 observations. The first data file was simulated with 66 r of the correlation matrix equal to 0.35. The remaining six files were simulated with 66 r equal to 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95. Real data file with twelve variables of 6340 maize plants with r in the interval $|0.01 \leq r \leq 0.99|$ was used. For the eight cases, PCA were performed in 3000 resamples with replacement, for sample sizes of 12 to 1000 observations. Insufficient sampling generates inaccurate and biased principal components (PC1 and PC2) eigenvalues estimates, and samples with a high observations number allow reliable PCA. The precision of the PC1 and PC2 eigenvalues estimates increases with the highest observations number. The precision of the PC1 eigenvalues estimates increases with larger r magnitudes, and the PC2 eigenvalues decreases.

1. Introduction

Maize is the cereal with the highest production volume worldwide, with the three largest producers in decreasing order being United States of America, China and Brazil (USDA, 2021). In maize experiments, multiple variables have been evaluated and the patterns of association between them have been measured using correlation coefficients and/or complementary analyzes (Cargnelutti Filho et al., 2010; Toebe et al., 2015; Olivoto et al., 2018; Cargnelutti Filho and Toebe, 2020).

Due to the difficulty of interpreting experiments with a high number of variables, in many cases, the principal components analysis (PCA) has been used, which allows to reduce the dimensionality of multivariate data through the generation of new variables (components) aiming at maximizing information in the first few components (Stauffer et al., 1985; Abdi and Williams, 2010). According to Ramachandran and Aschheim (2005) and Abdi and Williams (2010), PCA is one of the oldest and best-known techniques of multivariate analysis, that transforms a data set having a large number of inter-related variables to a new set of uncorrelated variables (principal components). According to Abdi and Williams (2010), PCA is probably the most popular multivariate statistical technique, used by almost all scientific disciplines.

In PCA, Stauffer et al. (1985) compared principal components derived from sets of real data and from randomly generated data in sample size scenarios. Osborne and Costello (2004) evaluated the effect of sample size and subject to item ratio and Ramachandran and Aschheim (2005) evaluated the effect of sample size on the accuracy with which the mode shapes of vibration can be computed. Kocovsky et al. (2009) evaluated the effect of sample size on the stability of PCA in truss-based fish morphometrics and Shaukat et al. (2016) evaluated the impact of sample size on the eigenstructure in an environmental database. Already Gañan-Cardenas and Correa-Morales (2021) studied the influence of sample size to test the equality for the smallest eigenvalues and Björklund (2019) described the main precautions that must be taken when using the PCA technique. Finally, Dochtermann and Jenkins (2011) studied the implication of small sample size in multivariate methods.

In maize research's, PCA was applied to evaluate the relationship of climatic indices and yields variables (Meyer et al., 1991), in classification of Italian maize germplasm (Brandolini and Brandolini, 2001), for predicting biomass and grain yields (Shukla et al., 2004), to assess aflatoxin contamination (Yao et al., 2011), to evaluated yield contributing variables (Bharathiveeramani and Prakash, 2012) and for an automatic system to kernel inspection (Valiente-González et al., 2014). PCA was also applied to characterize maize hybrids for water shortage (Guimarães et al., 2014), to evaluated sixty inbreds lines (Sandeep et al., 2017), to characterize grain yield and other variables in different maize hybrids grown under heat and drought stress (Ali et al., 2015), to predicting flowering time, yield, and kernel dimensions by analyzing aerial images (Wu et al., 2019), to evaluated plant nutrient traits in baby maize (Magudeeswari et al., 2019) and to characterize fifty-six Algerian maize populations (Belalia et al., 2019) and twenty-six sweet maize genotypes (Hemavathy, 2020).

Although the PCA technique is widely used in studies with maize and other crops, no researches were found relating the effect of the correlation coefficients magnitude in the correlation matrix and the sample size required in the PCA for maize variables and/or in simulated data. Some studies of sample size for estimate maize correlation coefficients, developed via resampling with replacement, indicated that the greater the correlation between two pairs of variables, the less is the need of sample size to estimate the correlation, in a given precision (Cargnelutti Filho et al., 2010; Toebe et al., 2015; Olivoto et al., 2018). As the PCA can be obtained from the correlation matrix, matrices with a high degree of correlation are expected to be well estimated even with few observations and, on the other hand, matrices with low degrees of correlation may require larger sample sizes (Cargnelutti Filho et al., 2010; Toebe et al., 2015; Olivoto et al., 2018).

The quality of the correlation matrix estimation may interfere in the principal components estimation, and highly correlated matrices can result in first components with high eigenvalue scores and well estimated, to the detriment of the last components and vice versa. Furthermore, matrices with the same correlation mean, but with varying degrees of correlation between pairs of variables, may have a different impact on the estimation of the principal components. In this sense, the objective of this work was to study the impact of the number of observations

and the magnitude of the correlation matrix coefficients in the eigenvalue estimates precision of the principal component analysis, with simulated data and real data of maize (*Zea mays* L.).

2. Material And Methods

Seven data files were simulated with p = 12 variables (X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11 and X12) and n = 6340 observations, using the *mvrnorm* function of the *MASS* package in R software (R Core Team, 2021). The variables were simulated with a multivariate normal distribution, with a mean zero and standard deviation one, and conditioned to all 66 Pearson's linear correlation coefficients (r) between the variables in each data file being equal. Thus, the correlation matrices between the 12 variables in files 1, 2, 3, 4, 5, 6 and 7 showed 66 r values equal to 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, respectively.

The eighth data file was formed by 12 variables measured in 6340 maize plants (*Zea mays* L.). These real data were obtained from experiments conducted in the 2008/2009 (first experiment) and 2009/2010 (second experiment) agricultural seasons, in the experimental area of the Plant Science Department, at the Federal University of Santa Maria, Santa Maria, State of Rio Grande do Sul, Brazil (29º42'S, 53º49'W, at 95 m altitude). In the first experiment were evaluated 361 plants of the single hybrid P32R21, 373 plants of the three-way hybrid DKB566, and 416 plants of the double cross hybrid DKB747. In the second experiment were evaluated 1777 plants of the single hybrid 30F53, 1693 plants of the three-way hybrid DKB566, and 1720 plants of the double cross hybrid DKB747.

In all the 6340 plants, the following variables were measured: plant height at harvest (PH, in cm), ear insertion height (EIH, in cm), ear weight (EW, in g), number of grain rows per ear (NR), ear length (EL, in cm), ear diameter (ED, in mm), cob weight (CW, in g), cob diameter (CD, in mm), hundred grains weight (HGW, in g), number of grains per ear (NGE), grain length (GL, in mm), calculated as the difference between the diameters of ear and cob divided by two, and grain yield (GY, in g per plant).

In the eight data files, consisting of 12 columns (variables) and 6340 rows (observations), principal component analysis (PCA) was performed from Pearson's linear correlation matrix between the variables. The correlation matrix, for the maize data file, was chosen due to the different variable measurement scales.

For each data file, 989 sample sizes (number of observations) were planned, with the initial sample size of 12 observations (in this study considered as a reference, i.e., minimum size required for principal component analysis) and the other sample sizes obtained with the increment of an observation. Thus, the planned sample sizes were n = 12, 13, 14, ..., 1000 observations. Thus, sample sizes of 12 to 1000 observations were planned. For each sample size planned, in each data file, 3000 resamples with replacement were obtained. In each resample, the eigenvalues estimate of the first two principal components (PC1 and PC2) were obtained. Thus, for each planned sample size, 3000 estimates of the PC1 and PC2 eigenvalues were obtained.

Based on 3000 eigenvalue estimates for each sample size and principal component (PC1 and PC2), the 97.5% percentile, the mean, the 2.5% percentile and the coefficient of variation (CV, in %) were determined. The percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{2.5\%}$), and coefficient of variation (CV, in %), for n = 12 and n = 1000 observations were presented in a table and the others were plotted in graphs for better visual representation. The statistical analysis was performed using Microsoft Office Excel and the R software (R Core Team, 2021).

3. Results

3.1. Eigenvalues of the principal components

In the seven data files, simulated to generate the correlation matrix with the 66 coefficients (r) equal to 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, it's possible to investigate the impact of the observations number and the magnitude of the correlation matrix coefficients in the eigenvalues estimates precision of the principal components, in extreme situations, that is, with all equal coefficients in the matrix. While, for situations with different values of r, it's possible to investigate based on real data, in which the magnitude of the correlation coefficients, in absolute values, fluctuated in the range $| 0.01 \leq r \leq 0.99 |$ and the mean of r was 0.42 (Table 1).

In the simulated data files, the eigenvalue estimates of the first principal component (PC1) increased gradually, with the magnitude increase of the correlation matrix coefficients (Table 1). The eigenvalues estimate of the others principal components (PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, PC12) decreased with the increase in the magnitude of the correlation matrix coefficients. Consequently, a similar response pattern was obtained in relation to the variance (in %) and the accumulated variance (in %). Therefore, as expected, the greater the magnitude of the correlation matrix coefficients, the greater the ability of the first principal component to explain the total variation of the data and, consequently, the greater the effectiveness of the principal component analysis.

In the maize real data file, with an average r of 0.42, the estimate of the PC1 eigenvalue (6.08) was relatively higher than the simulated data file with r = 0.45 (5.95) (Table 1). The presence of coefficients with high magnitudes (for example, EW and NGE = 0.90; EW and GY = 0.99; NGE and GY = 0.91) in the correlation matrix of the real data, explains the higher estimate of the PC1 eigenvalue. If a data file were simulated with the 66 coefficients of the correlation matrix equal to 0.42, the eigenvalue of PC1 would be 5.62 and the other eleven eigenvalues of the principal components would be equal to 0.58. Therefore, it can be inferred that for matrices with the same average of r (for example, r = 0.42), the eigenvalue estimates of PC1 will be relatively higher in those matrices with different values r (in this study, PC1 = 6.08) compared to matrices with equal values of r (in this study, PC1 = 5.62).

3.2. Number of observations

In the simulated data files, with r equal to 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, the PC1 eigenvalues estimates, with the 3640 observations, were 4.85, 5.95, 7.05, 8.15, 9.25, 10.35 and 11.45, respectively (Table 1). The averages of the 3000 PC1 eigenvalues estimates obtained from 12 observations (the smallest sample size used in this study) were 5.17, 6.08, 7.08, 8.07, 9.13, 10.25 and 11.40. In the other hand, the averages of the 3000 PC1 eigenvalues estimates obtained from 1000 observations (largest sample size used) were 4.85, 5.95, 7.05, 8.15, 9.25, 10.35 and 11.45 (Table 2 and Fig. 1). Therefore, the difference between the average of 3000 resamples and the eigenvalue obtained from 3640 observations decreases with the increase in the number of observations, which shows an increase in accuracy. Consequently, an insufficient number of observations generates biased estimates of PC1 eigenvalues. Similar reasoning is valid for the second principal component (PC2) (Tables 1 and 2 and Fig. 2).

In relation to PC1, in the eight data files, there was a reduction of the 97.5% percentile, an increase of the 2.5% percentile (Fig. 1), and a reduction in the coefficient of variation (Fig. 3) with an increase in the number of observations. Similar reasoning is valid for the second principal component (PC2) of the maize real data file (Figs. 2 and 4). As for the seven files of simulated data, in relation to PC2, there was a reduction of the 97.5% percentile, 2.5% percentile and coefficient of variation with the increase in the number of observations (Figs. 2 and 4). So, it can be inferred that with the increase in the number of observations there is an increase in the precision of the eigenvalues estimates of PC1 and PC2 (i.e., decrease in CV) and, therefore, principal components analysis with a high number of observations should be encouraged. Although the dimensioning of the observations number is not the principal focus of this study, visually, there is a tendency to stabilize the variation coefficient with approximately 270 observations (Figs. 3 and 4), which would indicate smaller gains from this number of observations onwards.

3.3. Magnitude of the correlation matrix coefficients

In relation to PC1, in the simulated data files with r equal to 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, the averages of the 3000 CV estimates, obtained from 12 observations were 21.00%, 19.07%, 17.50%, 15.00%, 11.44%, 7.20% and 2.77% and from 1000 observations were 2.84%, 2.22%, 1.95%, 1.52%, 1.10%, 0.64% and 0.21%, respectively (Table 2 and Fig. 3). Therefore, there was a decrease in the CV with an increase in the magnitude of the correlation matrix coefficients. Thus, it can be inferred that the precision of the eigenvalue estimates of PC1 increases (i.e., decrease in CV) with the increase of the magnitude of the correlation matrix coefficients.

In relation to PC2, in the simulated data files with r equal to 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, the averages of the 3000 CV estimates, obtained from 12 observations were 18.32%, 21.39%, 26.93%, 32.47%, 37.93%, 44.59% and 53.94% and from 1000 observations were 2.84%, 3.03%, 3.51%, 3.91%, 4.25%, 4.52%, 4.90%, respectively, that is, there was an increase in CV (i.e., decrease in precision) with the increase in the magnitude of the correlation matrix coefficients (Table 2 and Fig. 4).

4. Discussion

4.1. Eigenvalues of the principal components in maize

In the evaluation of 26 climatic divisions, Meyer et al. (1991) found that the first three components obtained from eighteen variables, had accumulated variance from 64.6–82.0%, depending on the climate region under study. The authors concluded that PCA is a powerful statistical tool for evaluating the relationship between crop yield and climatic variables. In the classification of Italian corn germplasm based on 562 accessions, Brandolini and Brandolini (2001) evaluated seventeen phenological, morphological and geographic variables. The authors found positive and negative correlations in the range of $|0.01 \leq r \leq 0.86|$ and the absolute r mean was 0.31. In PCA, the authors verified eigenvalues of 5.97, 2.95, 1.81, 1.39, 0.96 and 0.87 and cumulative variance of 35%, 52%, 63%, 71%, 77% and 82%, respectively, for the first six PCA components. According to Stauffer et al. (1985), generally, components with eigenvalues lower than $1/p \times 100\%$ of the total variance, where p represents the number of correlated variables, are not considered because they represent less information than those expressed in a single variable.

In seventeen traits of 144 inbred lines, Bharathiveeramani and Prakash (2012) found that the first five components had eigenvalues greater than 1, jointly explaining 78.36% of the total variance, with variance of 35.35%, 16.32%, 10.38%, 9.08% and 6.73%, respectively, for PC1, PC2, PC3, PC4 and PC5. To characterize corn hybrids for water shortage, Guimarães et al. (2014) used five variables and found that in the vegetative stage, the first two components explained 99.52% of the total variance (PC1 = 98.21% and PC2 = 1.31%). In the flowering stage, the first two components explained 85.08% of the total variance (PC1 = 60.05% and PC2 = 25.03%) and, in the grain swelling stage the first two components explained 98.52% of the total variance (PC1 = 91.48% and PC2 = 7.04%). Ali et al. (2015) evaluated sixteen variables in twelve F1 single cross-maize hybrids and four crop growing seasons and found that the first four components had eigenvalues greater than 1, with variance of 43.5% and 24.4%, respectively, for PC1 and PC2. Sandeep et al. (2017) evaluated twelve variables in sixty inbreds lines and found that the first three components presented 82.41% of the total variance, with 58.36%, 16.11% and 7.94% for PC1, PC2 and PC3.

In twelve baby corn genotypes, Magudeeswari et al. (2019) verified positive and negative correlations among six traits in the range of $|0.003 \leq r \leq 0.848|$ and the absolute r mean was 0.44. The authors verified that the first three principal components together accounted for 87.49% of variability, with eigenvalues of 2.90, 1.27 and 1.07 and variance of 48.37%, 21.23% and 17.89%, respectively for PC1, PC2 and PC3. In fifty-six Algerian maize populations, Belalia et al. (2019) evaluated fourteen agro-morphological traits and eighteen simple sequence repeat (SSR) markers and found that the first two components explained 55.44% of the total variation (PC1 = 43.04% and PC2 = 12.40%). In twenty-six sweet corn genotypes, Hemavathy (2020) evaluated thirteen quantitative and qualitative traits and verified positive and negative genotypic correlations in the range of $|0.01 \leq r \leq 0.971|$, with most non-significant and low magnitude correlations ($r \leq 0.50$) and absolute r mean of only 0.21. Six of the thirteen principal components showed eigenvalues greater than 1, with variance of 24.61%, 17.97%, 14.26%, 10.41%, 8.60% and 7.78%, respectively, for PC1, PC2, PC3, PC4, PC5 and PC6.

4.2. Number of observations and magnitude of the correlation matrix coefficients

Stauffer et al. (1985) compared principal components derived from real and random data. The authors simulated scenarios from 25 to 400 observations and verified that the variance associated with principal components from real data was relatively constant over all sample sizes. According to the authors, the first two components obtained from real data were considerably higher to those obtained from random data and, due to the low variability of components in random data, a sample of 20 or more elements would be adequate to estimate the variance of the components in random data. Through the analysis of the graphs presented by Stauffer et al. (1985), it can be seen that the mean value of the component's variance fluctuated between the sample sizes. Also, as the sample size increased, the amplitude of the 95% confidence interval for the principal components decreased, indicating a precision gain. For random data, the percent variance decreased with larger samples, mainly in the first component, with greater decreases in the smaller sample sizes and subsequent stabilization. Adopting the largest scenario assessed (Fig. 3 - Stauffer et al. 1985), it can be seen that the first three components have mean variance stabilized from n = 200 to 300 elements.

Ramachandran and Aschheim (2005) evaluated the effect of sample size on the vibration computed accuracy and identified that scenarios with principal components with less explanatory variance required a larger sample size. According to the authors, when the sample size increases, the errors become small and finally reach a constant value. In this sense, according to Björklund (2019), the robustness of the principal components increases with increasing sample size. Also, Kocovsky et al. (2009) verified that in small sample sizes, eigenvalues for the first principal components were unstable and inflated, recommending a minimum subject to item ratio from 3.5 to 8.0 to increase the eigenvalues and eigenvectors stabilization. Finally, Osborne and Costello (2004) emphasize that in many cases the sample size is greater than 1000 and concluded regarding sample size, that more is always better.

5. Conclusions

Insufficient sampling generates inaccurate and biased principal components (PC1 and PC2) eigenvalues estimates, and samples with a high observations number allow reliable PCA.

The precision of the PC1 and PC2 eigenvalues estimates increases with the highest observations number.

The precision of the PC1 eigenvalues estimates increases with larger r magnitudes, and the PC2 eigenvalues decreases.

Declarations

Data availability:

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Acknowledgments

We thank the Brazilian National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), for research Grant to the first author (process number 304652/2017-2); and to those who assisted in carrying out the experiment and in data collection.

Author contributions

Alberto Cargnelutti Filho and Marcos Toebe contributed equally in the conceptualization, data curation, formal analysis, investigation, methodology, resources, software, supervision, visualization, roles/writing - original draft and writing - review and editing.

References

1. Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comp Stat, 2, 433-459. https://doi.org/10.1002/wics.101
2. Ali, F., Kanwal, N., Ahsan, M., Ali, Q., Bibi, I., Niazi, N. K., 2015. Multivariate analysis of grain yield and its attributing traits in different maize hybrids grown under heat and drought stress. Scientifica, 2015, 563869. http://dx.doi.org/10.1155/2015/563869
3. Belalia, N., Lupini, A., Djemel, A., Morsli, A., Mauceri, A., Lotti, C., Khelifi-Slaoui, M., Khelifi, L., Sunseri, F., 2019. Analysis of genetic diversity and population structure in Saharan maize (*Zea mays* L.) populations using phenotypic traits and SSR markers. Genetic Resources and Crop Evolution, 66, 243-257. https://doi.org/10.1007/s10722-018-0709-3
4. Bharathiveeramani, B., Prakash, M., 2012. Factor analysis for yield contributing traits in maize (*Zea mays* L.). Electronic Journal of Plant Breeding, 3, 998-1001.
5. Björklund, M., 2019. Be careful with your principal components. Evolution, 73, 2151-2158. https://doi.org/10.1111/evo.13835
6. Brandolini, A., Brandolini, A., 2001. Classification of Italian maize (*Zea mays* L.) germplasm. Plant Genetic Resources Newsletter, 126, 1-11. https://www.bioversityinternational.org/fileadmin/_migrated/uploads/tx_news/Plant_Genetic_Resources_Newsletter_719.pdf
7. Cargnelutti Filho, A., Toebe, M., 2020. Reference sample size for multiple regression in corn. Pesquisa Agropecuária Brasileira, 55, e01400. https://doi.org/10.1590/s1678-3921.pab2020.v55.01400
8. Cargnelutti Filho, A., Toebe, M., Burin, C., Silveira, T.R., Casarotto, G., 2010. Sample size for estimating the Pearson correlation coefficient among corn characters. Pesquisa Agropecuária Brasileira, 45, 1363-1371. https://doi.org/10.1590/S0100-204X2010001200005
9. Dochtermann, N.A., Jenkins, S.H., 2011. Multivariate methods and small sample sizes. Ethology, 117, 95-101. https://doi.org/10.1111/j.1439-0310.2010.01846.x
10. Gañan-Cardenas, E., Correa-Morales, J.C., 2021. Comparison of correction factors and sample size required to test the equality of the smallest eigenvalues in principal component analysis. Revista Colombiana de Estadistica, 44, 43-64. https://doi.org/10.15446/rce.v44n1.83987
11. Guimarães, P.S., Bernini, C.S., Pedroso, F.K.J.V., Paterniani, M.E.A.G.Z., 2014. Characterizing corn hybrids (*Zea mays* L) for water shortage by principal components analysis. Maydica, 59, 72-79.
12. Hemavathy, A.T., 2020. Principal component analysis in sweet corn (*Zea mays* L. saccharata.). Forage Research, 45, 264-268.
13. Kocovsky, P.M., Adams, J.V., Bronte, C.R., 2009. The effect of sample size on the stability of principal components analysis of truss-based fish morphometrics. Transactions of the American Fisheries Society, 138, 487-496. https://doi.org/10.1577/T08-091.1
14. Magudeeswari, P., Sastry, E.V.D., Devi, T.R., 2019. Principal component (PCA) and cluster analyses for plant nutrient traits in baby corn (*Zea mays* L.). Indian Journal of Agricultural Research, 53, 353-357. https://doi.org/10.18805/IJARe.A-5042
15. Meyer, S.J., Hubbard, K.G., Wilhite, D.A., 1991. The relationship of climatic indices and variables to corn (maize) yields: A principal components analysis. Agricultural and Forest Meteorology, 55, 59-84. https://doi.org/10.1016/0168-1923(91)90022-I
16. Olivoto, T., Lúcio, A.D., Souza, V.Q.de, Nardino, M., Diel, M.I., Sari, B.G., Krysczun, D.K., Meira, D., Meier, C., 2018. Confidence interval width for Pearson's correlation coefficient: a Gaussian-independent estimator based on sample size and strength of association. Agronomy Journal, 110, 503-510. https://doi.org/10.2134/agronj2017.09.0566
17. Osborne, J.W., Costello, A.B., 2004. Sample size and subject to item ratio in principal components analysis. Practical Assessment, Research and Evaluation, 9, 1-9. https://doi.org/10.7275/ktzq-jq66
18. R Core Team. 2021. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at: <http://www.R-project.org>. Accessed on: Mar. 15 2021.
19. Ramachandran, J., Aschheim, M.A., 2005. Sample size and error in the determination of mode shapes by principal components analysis. Engineering Structures, 27, 1951-1967. https://doi.org/10.1016/j.engstruct.2005.06.020

20. Sandeep, S., Bharathi, M., Reddy, V.N., 2017. Principal component analysis in inbreds of maize (*Zea mays* L.). International Journal of Pure & Applied Bioscience, 5, 2008-2013. http://dx.doi.org/10.18782/2320-7051.5762
21. Shaukat, S.S., Rao, T.A., Khan, M.A., 2016. Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. Ekologia Bratislava, 35, 173-190. https://doi.org/10.1515/eko-2016-0014
22. Shukla, M.K., Lal, R., Ebinger, M., 2004. Principal component analysis for predicting corn biomass and grain yields. Soil Science, 169, 215-224. https://doi.org/10.1097/01.ss.0000122521.03492.eb
23. Stauffer, D.F., Garton, E.O., Steinhorst, R.K., 1985. A comparison of principal components from real and random data. Ecology, 66, 1693-1698. https://doi.org/10.2307/2937364
24. Toebe, M., Cargnelutti Filho, A., Lopes, S.J., Burin, C., Silveira, T.R.da, Casarotto, G., 2015. Sample size in the estimation of correlation coefficients for corn hybrids in crops and accuracy levels. Bragantia, 74, 16-24. https://doi.org/10.1590/1678-4499.0324
25. USDA - United States Department of Agriculture, 2021. World Agricultural Production. 37 p. (USDA. Circular Series WAP 4-21). Available at: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf>. Accessed on: Apr. 25 2021.
26. Valiente-González, J.M., Andreu-García, G., Potter, P., Rodas-Jordá, T., 2014. Automatic corn (*Zea mays*) kernel inspection system using novelty detection based on principal component analysis. Biosystems Engineering, 117, 94-103. https://doi.org/10.1016/j.biosystemseng.2013.09.003
27. Wu, G., Miller, N.D., de Leon, N., Kaeppler, S.M., Spalding, E.P., 2019. Predicting *Zea mays* flowering time, yield, and kernel dimensions by analyzing aerial images. Frontiers in Plant Science, 10, 1251. https://doi.org/10.3389/fpls.2019.01251
28. Yao, H., Hruska, Z., Kincaid, R., Ononye, A., Brown, R.L., Bhatnagar, D., Cleveland, T.E., 2011. Selective principal component regression analysis of fluorescence hyperspectral image to assess aflatoxin contamination in corn. Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. https://doi.org/10.1109/WHISPERS.2011.6080970

Tables

Table 1

Estimates of Pearson linear correlation coefficients (r) between the 66 pairs of twelve variables in maize cultivars (*Zea mays* L.) measured in 6340 plants (real data file). Estimates of the variance (eigenvalues) of twelve principal components (PC1, PC2, ..., PC12) in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file with r average of correlation matrix = 0.42.

Variables[(1)]	PH	EIH	EW	NR	EL	ED	CW	CD	HGW	NGE	GL	GY
PH	1	0.66	0.34	0.23	0.23	0.31	0.20	0.10	0.20	0.32	0.30	0.34
EIH	0.66	1	0.06	0.10	-0.03	0.06	0.11	0.01	-0.02	0.06	0.07	0.04
EW	0.34	0.06	1	0.41	0.76	0.83	0.66	0.50	0.57	0.90	0.63	0.99
NR	0.23	0.10	0.41	1	0.25	0.59	0.26	0.37	0.04	0.49	0.44	0.41
EL	0.23	-0.03	0.76	0.25	1	0.53	0.77	0.50	0.35	0.69	0.27	0.71
ED	0.31	0.06	0.83	0.59	0.53	1	0.47	0.54	0.54	0.76	0.80	0.84
CW	0.20	0.11	0.66	0.26	0.77	0.47	1	0.71	0.38	0.48	0.05	0.56
CD	0.10	0.01	0.50	0.37	0.50	0.54	0.71	1	0.37	0.36	-0.07	0.44
HGW	0.20	-0.02	0.57	0.04	0.35	0.54	0.38	0.37	1	0.20	0.38	0.56
NGE	0.32	0.06	0.90	0.49	0.69	0.76	0.48	0.36	0.20	1	0.65	0.91
GL	0.30	0.07	0.63	0.44	0.27	0.80	0.05	-0.07	0.38	0.65	1	0.69
GY	0.34	0.04	0.99	0.41	0.71	0.84	0.56	0.44	0.56	0.91	0.69	1
Data files	Correlation	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
		Variance (eigenvalues)											
Simulated	0.35	4.85	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
Simulated	0.45	5.95	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
Simulated	0.55	7.05	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
Simulated	0.65	8.15	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Simulated	0.75	9.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Simulated	0.85	10.35	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Simulated	0.95	11.45	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Real	0.42	6.08	1.68	1.46	1.02	0.80	0.33	0.28	0.19	0.14	0.02	0.00	0.00
		% of variance											
Simulated	0.35	40.42	5.42	5.42	5.42	5.42	5.42	5.42	5.42	5.42	5.42	5.42	5.42
Simulated	0.45	49.58	4.58	4.58	4.58	4.58	4.58	4.58	4.58	4.58	4.58	4.58	4.58
Simulated	0.55	58.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75
Simulated	0.65	67.92	2.92	2.92	2.92	2.92	2.92	2.92	2.92	2.92	2.92	2.92	2.92
Simulated	0.75	77.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
Simulated	0.85	86.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25
Simulated	0.95	95.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
Real	0.42	50.67	14.03	12.13	8.49	6.67	2.76	2.36	1.61	1.15	0.13	0.00	0.00
		Cumulative % of variance											
Simulated	0.35	40.42	45.83	51.25	56.67	62.08	67.50	72.92	78.33	83.75	89.17	94.58	100.00
Simulated	0.45	49.58	54.17	58.75	63.33	67.92	72.50	77.08	81.67	86.25	90.83	95.42	100.00
Simulated	0.55	58.75	62.50	66.25	70.00	73.75	77.50	81.25	85.00	88.75	92.50	96.25	100.00
Simulated	0.65	67.92	70.83	73.75	76.67	79.58	82.50	85.42	88.33	91.25	94.17	97.08	100.00
Simulated	0.75	77.08	79.17	81.25	83.33	85.42	87.50	89.58	91.67	93.75	95.83	97.92	100.00
Simulated	0.85	86.25	87.50	88.75	90.00	91.25	92.50	93.75	95.00	96.25	97.50	98.75	100.00

Simulated	0.95	95.42	95.83	96.25	96.67	97.08	97.50	97.92	98.33	98.75	99.17	99.58	100.00
Real	0.42	50.67	64.70	76.83	85.32	91.99	94.75	97.11	98.72	99.87	100.00	100.00	100.00

(1) PH: plant height at harvest; EIH: ear insertion height; EW: ear weight; NR: number of grain rows per ear; EL: ear length; ED: ear diameter; CW: cob weight; CD: cob diameter; HGW: hundred grains weight; NGE: number of grains per ear; GL: grain length; and GY: grain yield.

Table 2

Percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{97.5\%}$), and coefficient of variation (CV, in %) for 3000 eigenvalues estimates of the first two principal components (PC1 and PC2). Estimates obtained from 3000 resamples with replacement for n = 12 and 1000 observations in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file of twelve variables in maize cultivars (*Zea mays* L.), measured in 6340 plants, with r average of correlation matrix = 0.42.

Data files	Correlation	$P_{97.5\%}$	Mean	$P_{2.5\%}$	CV(%)	$P_{97.5\%}$	Mean	$P_{2.5\%}$	CV(%)
		n = 12				n = 1000			
		PC1							
Simulated	0.35	7.40	5.17	3.20	21.00	5.12	4.85	4.57	2.84
Simulated	0.45	8.20	6.08	3.77	19.07	6.20	5.95	5.69	2.22
Simulated	0.55	9.22	7.08	4.44	17.50	7.32	7.05	6.78	1.95
Simulated	0.65	10.05	8.07	5.40	15.00	8.39	8.15	7.90	1.52
Simulated	0.75	10.58	9.13	6.61	11.44	9.44	9.25	9.04	1.10
Simulated	0.85	11.21	10.25	8.30	7.20	10.47	10.35	10.22	0.64
Simulated	0.95	11.76	11.40	10.59	2.77	11.49	11.45	11.40	0.21
Real	0.42	8.26	6.46	4.66	14.49	6.30	6.09	5.88	1.78
		PC2							
Simulated	0.35	2.73	2.00	1.32	18.32	0.81	0.77	0.73	2.84
Simulated	0.45	2.53	1.76	1.11	21.39	0.69	0.65	0.61	3.03
Simulated	0.55	2.34	1.48	0.81	26.93	0.57	0.53	0.50	3.51
Simulated	0.65	2.07	1.19	0.57	32.47	0.45	0.41	0.38	3.91
Simulated	0.75	1.65	0.88	0.41	37.93	0.32	0.30	0.27	4.25
Simulated	0.85	1.14	0.54	0.23	44.59	0.19	0.18	0.16	4.52
Simulated	0.95	0.44	0.19	0.07	53.94	0.07	0.06	0.05	4.90
Real	0.42	3.34	2.32	1.47	21.08	1.82	1.69	1.58	3.55

Figures

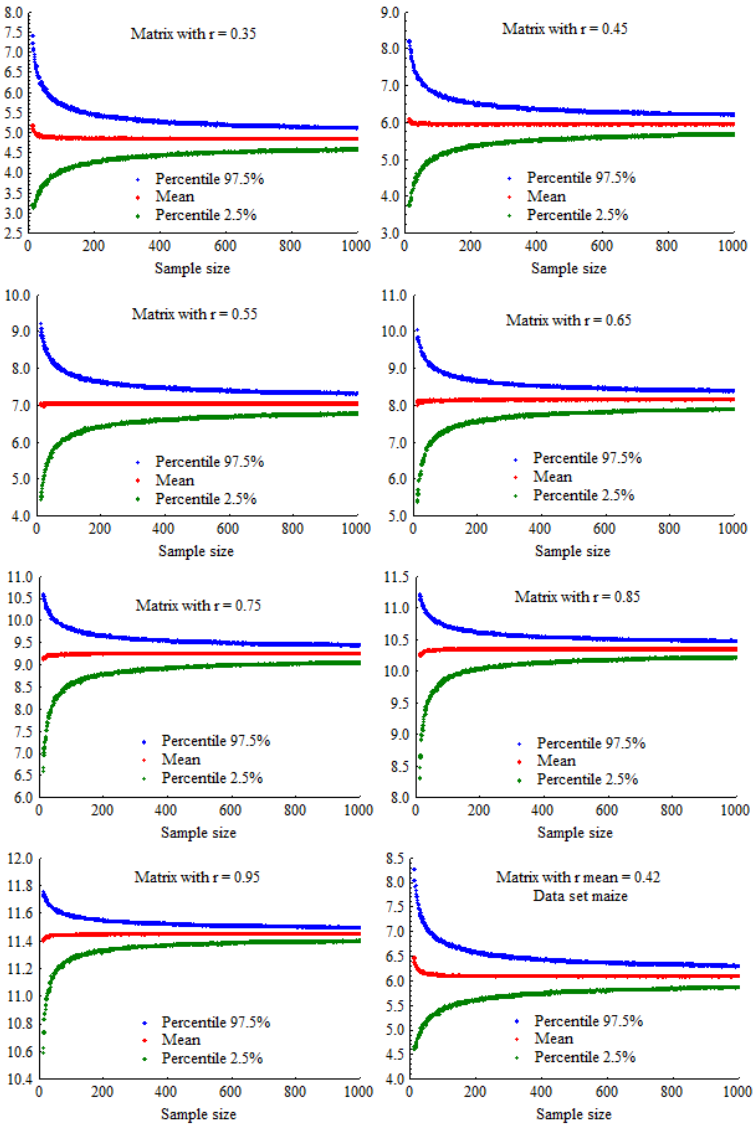

Figure 1

Percentile 97.5%, mean, and percentile 2.5% for 3000 eigenvalues estimates of the first principal component (PC1), based on resampling in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file of twelve variables in maize cultivars (*Zea mays* L.), measured in 6340 plants, with r average of correlation matrix = 0.42.

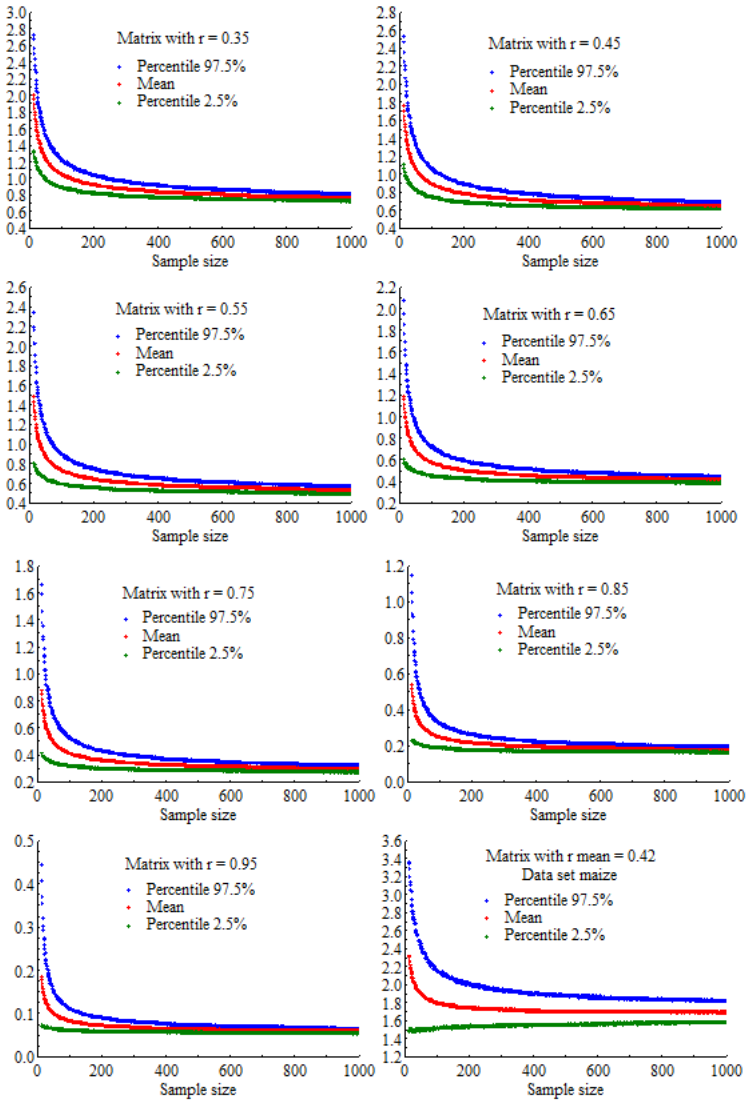

Figure 2

Percentile 97.5%, mean, and percentile 2.5% for 3000 eigenvalues estimates of the second principal component (PC2), based on resampling in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file of twelve variables in maize cultivars (*Zea mays* L.), measured in 6340 plants, with r average of correlation matrix = 0.42.

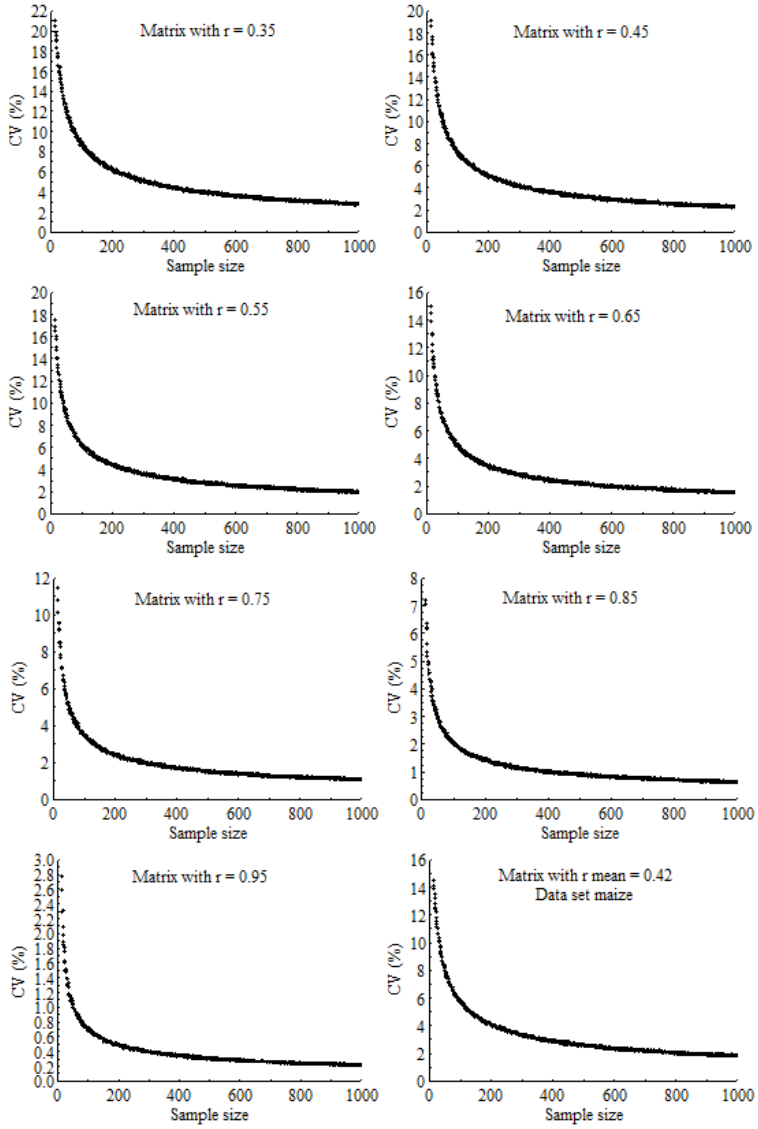

Figure 3

Coefficient of variation (CV, in%) for 3000 eigenvalues estimates of the first principal component (PC1), based on resampling in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file of twelve variables in maize cultivars (*Zea mays* L.), measured in 6340 plants, with r average of correlation matrix = 0.42.

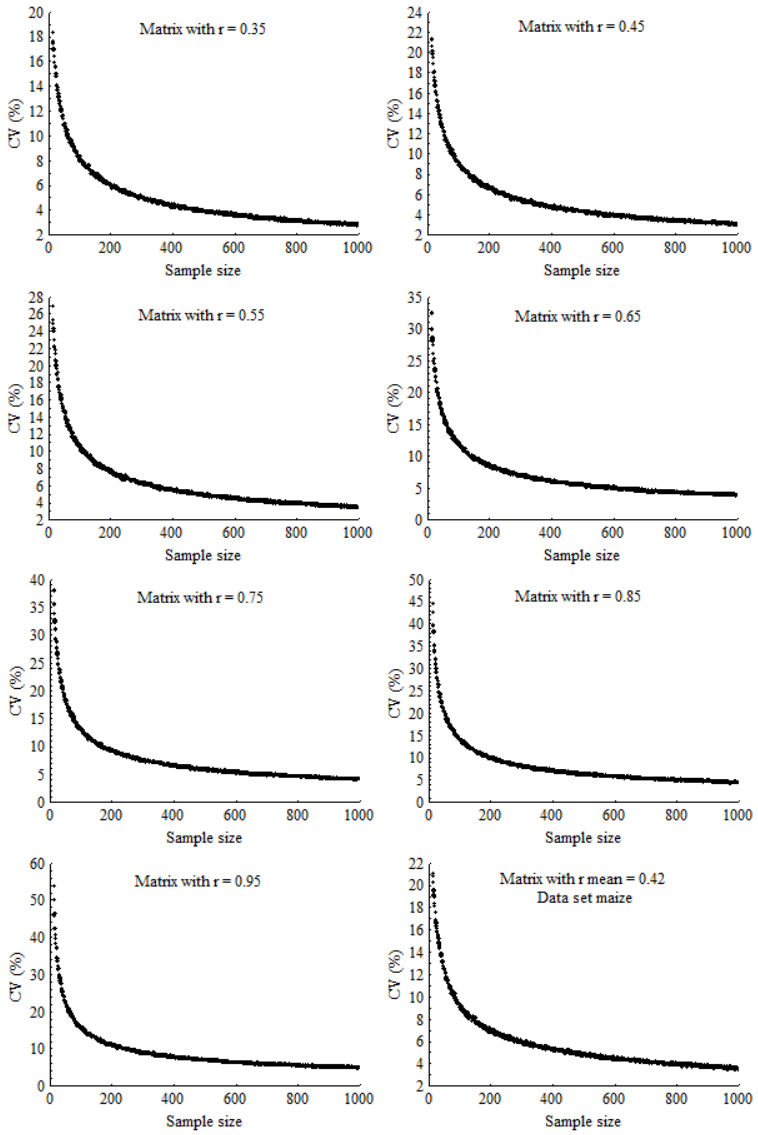

Figure 4

Coefficient of variation (CV, in%) for 3000 eigenvalues estimates of the second principal component (PC2), based on resampling in seven simulated data files with correlation matrices equal to r = 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95, and a real data file of twelve variables in maize cultivars (*Zea mays* L.), measured in 6340 plants, with r average of correlation matrix = 0.42.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.png