A RESAMPLING TEST FOR PRINCIPAL COMPONENT ANALYSIS OF GENOTYPE-BY-ENVIRONMENT INTERACTION

JOHANNES FORKMAN

Abstract. In crop science, genotype-by-environment interaction is often explored using the "genotype main effects and genotype-by-environment interaction effects" (GGE) model. Using this model, a singular value decomposition is performed on the matrix of residuals from a fit of a linear model with main effects of environments. Provided that errors are independent, normally distributed and homoscedastic, the significance of the multiplicative terms of the GGE model can be tested using resampling methods. The GGE method is closely related to principal component analysis (PCA). The present paper describes *i*) the GGE model, *ii*) the simple parametric bootstrap method for testing multiplicative genotype-by-environment interaction terms, and *iii*) how this resampling method can also be used for testing principal components in PCA.

1. Introduction

Forkman and Piepho (2014) proposed a resampling method for testing interaction terms in models for analysis of genotype-by-environment data. The "genotype main effects and genotype-by-environment interaction effects"(GGE) analysis (Yan et al. 2000; Yan and Kang, 2002) is closely related to principal component analysis (PCA). For this reason, the method proposed by Forkman and Piepho (2014), which is called the "simple parametric bootstrap method", can be used for testing principal components in PCA as well. The proposed resampling method is parametric in the sense that it assumes homoscedastic and normally distributed observations. The method is "simple", because it only involves repeated sampling of standard normal distributed values. Specifically, no parameters need to be estimated. The present paper describes how the simple parametric bootstrap method can be used for testing principal components in PCA.

The GGE model is used for analysis of complete series of crop variety trials, that is, for series such that all trials include all varieties. Analysis

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Dimensionality reduction, Principal component analysis, Resampling methods, Singular value decomposition.

of incomplete series of crop variety trials has its own difficulties (Forkman, 2013). Researchers want to explore genotype-by-environment interaction in order to define groups of genotypes that perform similar in varying environments, and to define clusters, so called mega-environments (Gauch, 1992), of environments in which genotypes give similar results.

Section 2 describes the GGE analysis through an example. In this example, an international series of maize variety trials is analyzed with specific focus on the interaction between varieties (i.e. genotypes) and locations (i.e. environments). Section 3 presents the simple parametric bootstrap method and illustrates this method using the same maize variety trials example. In Section 4, it is clarified that the GGE analysis is indeed a PCA, which implies that the simple parametric bootstrap method can be used for other applications than analysis of genotype-by-environment interaction.

Forkman and Piepho (2014) mainly focused on an analysis using the additive main effects and multiplicative interaction (AMMI) model, which was introduced by Mandel (1971) and advocated by Kemtpon (1984) and Gauch (1988). The AMMI analysis is is not a PCA, although closely related. The objective of the present paper is to show, through an explicit example, that the method proposed by Forkman and Piepho (2014) can also be used for the problem of dimensionality reduction in PCA.

2. GGE analysis

The dataset of Table 1 is an example of a complete series of crop variety trials. This dataset, which was also analyzed by Cornelius, Crossa and Seyedsadr (1996), includes yields from maize trials carried out by the international maize and wheat improvement center (CIMMYT). The study includes nine maize varieties (G1–G9) that were investigated in 20 environments (E1–E20). Varieties are genotypes, because differences between varieties are due to differences in genetic content. Similarly, trials represent varying environments.

In GGE analysis, effects of genotypes and effects of genotype-by-environment interaction are explored simultaneously. In the first step of the GGE analysis, the overall mean and the estimates of main effects of the environments are removed from the data. Since the series is complete, this is simply done by subtracting row means. The result is a matrix, $\hat{\mathbf{E}}$, of residuals from a fit of linear model with main effects of environments. Table 2 shows the matrix $\hat{\mathbf{E}}$ as computed from the dataset matrix of Table 1. Note that in Table 2, all rows sum to zero.

Let I and J denote the number of environments and genotypes, respectively, and let $M = \min(I, J-1)$. Through singular value decomposition, $\hat{\mathbf{E}}$ can be written as $\hat{\mathbf{E}} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Delta}}^{\mathrm{T}}$, where $\hat{\mathbf{\Gamma}} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_M)$ is an $I \times M$ matrix of estimated left-singular vectors, $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_M)$ is an $M \times M$

TABLE 1. Mean yields (kg/ha) of nine maize genotypes (G1–G9) investigated in 20 environments (E1–E20)

	G1	G2	G3	G4	G6	G6	G7	G8	G9
E1	3622	3426	3446	3720	3165	4116	3354	4529	3136
E2	3728	3919	4082	4539	4079	4878	4767	3393	4500
E3	5554	4937	5117	4542	6173	5205	5389	5248	3780
E4	4566	4963	5136	6030	5831	5980	4342	4442	5781
E5	4380	5201	4178	5672	5414	5591	4277	4476	5407
E6	6437	6036	6459	6678	6882	6916	6745	4986	5610
E7	2832	2515	3529	2998	3556	3949	3537	3088	3061
E8	6011	5278	4731	2516	2732	2983	4206	4484	3309
E9	4647	4714	5448	4864	5588	5603	4318	4001	5553
E10	3100	2972	2785	2843	2688	3024	2889	3353	2774
E11	4433	4349	4526	7117	5995	6150	5052	3713	6430
E12	6873	7571	7727	8385	8106	7637	7444	5816	8091
E13	6721	5627	6294	7332	7174	7262	5544	4117	6920
E14	5849	5932	5886	6439	6359	6380	5820	5522	6282
E15	4601	4126	4537	6331	6328	5961	4346	4321	4889
E16	5010	5196	5455	6351	6070	5730	5013	4551	5278
E17	4415	4211	4749	5161	5454	5807	3862	5243	4989
E18	3344	4415	4295	5618	4498	5333	5276	2940	5244
E19	1632	2282	3059	2233	3073	3011	3211	2634	2735
E20	4587	4396	5018	4988	5776	5088	4056	4806	4822

diagonal matrix of estimated singular values sorted from largest to smallest, and $\hat{\boldsymbol{\Delta}} = (\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_M)$ is a $J \times M$ matrix of estimated right-singular vectors. Environment and genotype principal components can be defined as

$$\hat{\boldsymbol{\gamma}}_1 \hat{\lambda}_1^c, \hat{\boldsymbol{\gamma}}_2 \hat{\lambda}_2^c, \ldots, \hat{\boldsymbol{\gamma}}_M \hat{\lambda}_M^c,$$

and

$$\hat{\boldsymbol{\delta}}_1 \hat{\lambda}_1^{1-c}, \hat{\boldsymbol{\delta}}_2 \hat{\lambda}_2^{1-c}, \ldots, \hat{\boldsymbol{\delta}}_M \hat{\lambda}_M^{1-c},$$

respectively, where $0 \leq c \leq 1$. For a discussion on how to choose c, see Jolliffe (2002).

It is common to display the first two principal components in a biplot, as in Figure 1. In this figure, principal component axis 1 (PC1) displays the values of $\hat{\boldsymbol{\gamma}}_1 \sqrt{\hat{\lambda}_1}$ and $\hat{\boldsymbol{\delta}}_1 \sqrt{\hat{\lambda}_1}$, whereas axis 2 (PC2) displays the values of $\hat{\boldsymbol{\gamma}}_2 \sqrt{\hat{\lambda}_2}$ and $\hat{\boldsymbol{\delta}}_2 \sqrt{\hat{\lambda}_2}$. Yan and Tinker (2006) explains how GGE biplots should be interpreted. Basically, genotypes that are close to each other in the biplot perform similar in varying environments. In Figure 1, it appears that genotypes G1, G2, G3 and G7 are similar to each other. Also, genotypes G4, G5, G6 and G9 are similar to each other. In performance, genotype G8 deviates from the other genotypes. One might wonder if these observed patterns are random or systematic. The simple parametric bootstrap method, described

TABLE 2. The matrix $\hat{\mathbf{E}}$ of residuals (kg/ha) from a fit of a linear model with main effects of environments

9.3	-186.7	-166.7	107.3	-447.7	503.3	-258.7	916.3	-476.7
-481.4	-290.4	-127.4	329.6	-130.4	668.6	557.6	-816.4	290.6
449	-168	12	-563	1068	100	284	143	-1325
-664.1	-267.1	-94.1	799.9	600.9	749.9	-888.1	-788.1	550.9
-575.1	245.9	-777.1	716.9	458.9	635.9	-678.1	-479.1	451.9
131.6	-269.4	153.6	372.6	576.6	610.6	439.6	-1319.4	-695.4
-397.4	-714.4	299.6	-231.4	326.6	719.6	307.6	-141.4	-168.4
1983.2	1250.2	703.2	-1511.8	-1295.8	-1044.8	178.2	456.2	-718.8
-323.7	-256.7	477.3	-106.7	617.3	632.3	-652.7	-969.7	582.3
163.6	35.6	-151.4	-93.4	-248.4	87.6	-47.4	416.6	-162.4
-874.2	-958.2	-781.2	1809.8	687.8	842.8	-255.2	-1594.2	1122.8
-643.7	54.3	210.3	868.3	589.3	120.3	-72.7	-1700.7	574.3
388.7	-705.3	-38.3	999.7	841.7	929.7	-788.3	-2215.3	587.7
-203.1	-120.1	-166.1	386.9	306.9	327.9	-232.1	-530.1	229.9
-447.9	-922.9	-511.9	1282.1	1279.1	912.1	-702.9	-727.9	-159.9
-396	-210	49	945	664	324	-393	-855	-128
-461.8	-665.8	-127.8	284.2	577.2	930.2	-1014.8	366.2	112.2
-1207.4	-136.4	-256.4	1066.6	-53.4	781.6	724.6	-1611.4	692.6
-1020.2	-370.2	406.8	-419.2	420.8	358.8	558.8	-18.2	82.8
-250.4	-441.4	180.6	150.6	938.6	250.6	-781.4	-31.4	-15.4

below, was developed for this question. The simple parametric bootstrap method can be used to test the significance of the principal components.

3. The simple parametric bootstrap method

Forkman and Piepho (2014) introduced the simple parametric bootstrap method for the GGE model. For hypothesis testing it is assumed that

$$\mathbf{E} = \mathbf{\Theta}_{(\kappa)} + \mathbf{R}, \tag{1}$$

where $\mathbf{E}$ is the matrix of true residuals. These are the residuals after subtraction of the actual intercept and the actual main effects of environments. In practice, $\mathbf{E}$ cannot be computed, because the true values of these parameters are not known; only $\hat{\mathbf{E}}$ can be computed (Table 2). Equation (1) is the null model, that is, the model under the null hypothesis. In (1), the fixed part of the null model is $\mathbf{\Theta}_{(\kappa)}$, and the random part is $\mathbf{R}$. The rank of $\mathbf{\Theta}_{(\kappa)}$ is κ. Thus, κ is the actual number of principal components. The matrix $\mathbf{R}$ is a matrix of independent N(0, σ^2) distributed errors.

The null hypothesis is $H_0 : \kappa = K$, and the alternative hypothesis is $H_1 : \kappa > K$. Hypotheses are tested sequentially: $K = 0, 1, 2, \ldots$ until a non-significant result is obtained. In order to test the significance of the

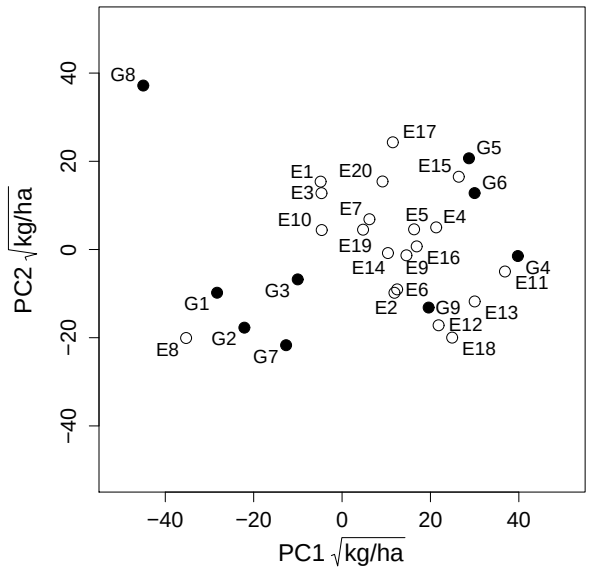

FIGURE 1. GGE biplot of the maize dataset (using $c = 0.5$)

$(K+1)$th component, the test statistic

$$T = \frac{\hat{\lambda}_{K+1}^2}{\sum_{k=K+1}^{M} \hat{\lambda}_k^2} \tag{2}$$

is used. For computation of the p-value, Forkman and Piepho (2014) proposed the "simple parametric bootstrap method":

1. Do the following a large number of times:
   i. Sample an $(I-K) \times (J-1-K)$ matrix of random standard normal values.
   ii. For this matrix, compute $T_b = \hat{\lambda}_1^2 / \sum_{k=1}^{L} \hat{\lambda}_k^2$.
2. Estimate the p-value as the frequency of T_b larger than T.

Note that this method is simple in the sense that no parameters need to be estimated. Still it is parametric, because it assumes the normal distribution. Forkman and Piepho (2014) called it a bootstrap method since they developed it as a simplified version of a full parametric bootstrap method that includes parameter estimation.

Table 3 presents the result of the simple parametric bootstrap method when applied to the maize example. First, the null hypothesis of no principal

components is tested. This null hypothesis is strongly rejected ($T = 0.640$, p-value $= 0.000$). Second, the null hypothesis of a model with a single principal component is tested. This null hypothesis cannot be rejected ($T = 0.319$, p-value $= 0.296$). In other words, it could not be inferred from the data that the actual model includes more than a single principal component. Since a non-significant result was obtained, significance testing is stopped at this stage.

The result of Table 3 sheds new light on the biplot of Figure 1. Since the second principal component is not significant, differences should be looked for mainly along the first principal component. It then appears that genotype G8 belongs to the same group as genotypes G1, G2, G3 and G7, because these genotypes are grouped on the left hand side of the first principal component axis.

TABLE 3. Sequential tests of the multiplicative terms of the maize dataset

	$K+1$	T	p-value	
Start $\rightarrow$	1	0.640	0.000	
	2	0.319	0.296	$\leftarrow$ Stop

4. The GGE analysis is a PCA

From Section 2 it might be obvious that the GGE analysis is nothing but a PCA. However, since in textbooks the PCA is usually presented slightly different, a few remarks might be helpful.

Let $\mathbf{X}$ be a column-wise mean-centered matrix. The singular values of $\mathbf{X}$, and then also of $\mathbf{X}^{\mathrm{T}}$, can be denoted $\hat{\lambda}_1$, $\hat{\lambda}_2$,..., $\hat{\lambda}_M$. In the GGE analysis, as presented in the example of the present paper, the matrix $\hat{\mathbf{E}}$ was instead row-wise mean-centered. This causes no difficulty, because the singular values of $\hat{\mathbf{E}}$ are the same as the singular values of the column-wise mean-centered matrix $\hat{\mathbf{E}}^{\mathrm{T}}$.

PCA uses the covariance matrix $\mathrm{cov}(\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{X}/(J-1)$, where J is the number of observations. The eigenvalues of $\mathrm{cov}(\mathbf{X})$ are $(\hat{\lambda}_1^2,\ \hat{\lambda}_2^2,\ \ldots,\ \hat{\lambda}_M^2)/(J-1)$, and the $(K+1)$th principal component accounts for $T = \hat{\lambda}_{K+1}^2 / \sum_{k=K+1}^{M} \hat{\lambda}_k^2$ per cent of the residual sum of squares. This is exactly the test statistic (2) that is used in the GGE analysis. Thus, large values of T indicate important principal components. The simple parametric bootstrap method, as presented in the present paper, can consequently be used for testing the significance of these components.

5. Conclusion

The GGE analysis is a PCA with environments as variables and genotypes as observations. PCA is a widely used method, with applications in all sorts

of different fields of research. Since the GGE analysis is a PCA, the simple parametric bootstrap method, which was developed for the GGE analysis, can be used also for other applications. Through this method, p-values can be computed for tests of principal components. However, it should be noted that the method assumes that random errors are independent, normally distributed and homoscedastic. When these requirements are fulfilled, the method performs well with regard to type I error and power (Forkman and Piepho, 2014), but this may not be the case otherwise. More research is needed to answer this question.

References

[1] P. L. Cornelius, J. Crossa, and M. S. Seyedsadr, *Statistical tests and estimators of multiplicative models for genotypeby- environment interaction*, in: Genotype-by-Environment Interaction, Boca Raton, FL, 1996, pp. 199-231.

[2] J. Forkman, *The use of a reference variety for comparisons in incomplete series of crop variety trials*. J. Appl. Statist. **40** (2013), 2681–2698.

[3] J. Forkman, and H. P. Piepho, *Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models*. Biometrics **70** (2014), 639–647.

[4] H. G. Gauch, *Model selection and validation for yield trials with interaction*. Biometrics **44** (1988), 705–715.

[5] H. G. Gauch, *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*, Elsevier, Amsterdam, 1992.

[6] I. Jolliffe, *Principal Component Analysis*, Springer, Secaucus, NJ, 2002.

[7] R. A. Kempton, *The use of biplots in interpreting variety by environment interactions*. J. Agric. Sci. Camb. **108** (1984), 123–135.

[8] J. Mandel, *A new analysis of variance model for non-additive data*. Technometrics **13** (1971), 1–18.

[9] W. Yan, L. A. Hunt, Q. Sheng, and Z. Szlavnics, *Cultivar evaluation and mega-environment investigation based on the GGE biplot*, Crop Sci. **40** (2000), 597–605.

[10] W. Yan, and M. S. Kang, *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*, CRC Press, Boca Raton, FL, 2002.

[11] W. Yan, and N. A. Tinker, *Biplot analysis of multi-environment trial data: principles and applications*. Can. J. Plant Sci. **86** (2006), 623–645.

Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Box 7043, 750 07 Uppsala, Sweden

E-mail address: johannes.forkman@slu.se