Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Department of Crop Production Ecology

A simple parametric bootstrap method for testing principal components in normally distributed data

LinStat 2014
Linköping, Sweden, August, 24-28 2014

Johannes Forkman, Swedish University of Agricultural Sciences (SLU)

Outline

1. The genotype main effects and genotype-by-environment interaction effect (GGE) model

2. The simple parametric bootstrap method

3. Principal component analysis (PCA)

The genotype main effects and genotype-by-environment interaction effect (GGE) model

Johannes Forkman

Genotypes and Environments

- Genotypes are usually varieties of some crop
- Environments are usually different locations

Yield (kg/ha)

$\mathbf{Y}$

$J = 9$ genotypes

$I = 20$ environments

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

GGE

- Explores effects of Genotypes and Genotype-by-Environment interaction simultaneously

- A singular value decomposition (SVD) on the matrix of residuals from a fit of a linear model with main effects of environments

Yan et al. (2000)

SLU

$\hat{\mathbf{E}}$

$I = 20$ environments	$J = 9$ genotypes									Sum
	9.3	-186.7	-166.7	107.3	-447.7	503.3	-258.7	916.3	-476.7	0
	-481.4	-290.4	-127.4	329.6	-130.4	668.6	557.6	-816.4	290.6	0
	449.0	-168.0	12.0	-563.0	1068.0	100.0	284.0	143.0	-1325.0	0
	-664.1	-267.1	-94.1	799.9	600.9	749.9	-888.1	-788.1	550.9	0
	-575.1	245.9	-777.1	716.9	458.9	635.9	-678.1	-479.1	451.9	0
	131.6	-269.4	153.6	372.6	576.6	610.6	439.6	-1319.4	-695.4	0
	-397.4	-714.4	299.6	-231.4	326.6	719.6	307.6	-141.4	-168.4	0
	1983.2	1250.2	703.2	-1511.8	-1295.8	-1044.8	178.2	456.2	-718.8	0
	-323.7	-256.7	477.3	-106.7	617.3	632.3	-652.7	-969.7	582.3	0
	163.6	35.6	-151.4	-93.4	-248.4	87.6	-47.4	416.6	-162.4	0
	-874.2	-958.2	-781.2	1809.8	687.8	842.8	-255.2	-1594.2	1122.8	0
	-643.7	54.3	210.3	868.3	589.3	120.3	-72.7	-1700.7	574.3	0
	388.7	-705.3	-38.3	999.7	841.7	929.7	-788.3	-2215.3	587.7	0
	-203.1	-120.1	-166.1	386.9	306.9	327.9	-232.1	-530.1	229.9	0
	-447.9	-922.9	-511.9	1282.1	1279.1	912.1	-702.9	-727.9	-159.9	0
	-396.0	-210.0	49.0	945.0	664.0	324.0	-393.0	-855.0	-128.0	0
	-461.8	-665.8	-127.8	284.2	577.2	930.2	-1014.8	366.2	112.2	0
	-1207.4	-136.4	-256.4	1066.6	-53.4	781.6	724.6	-1611.4	692.6	0
	-1020.2	-370.2	406.8	-419.2	420.8	358.8	558.8	-18.2	82.8	0
	-250.4	-441.4	180.6	150.6	938.6	250.6	-781.4	-31.4	-15.4	0

Singular value decomposition

$$\hat{\mathbf{E}} = \hat{\Gamma}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Delta}}^{\mathrm{T}}$$

$\hat{\Gamma}$ is a matrix of left-singular vectors: $\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_M$

$\hat{\mathbf{\Lambda}}$ is a diagonal matrix with singular values: $\hat{\lambda}_1, \hat{\lambda}_2, \ldots \hat{\lambda}_M$

$\hat{\mathbf{\Delta}}$ is a matrix of right-singular vectors: $\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_M$

where $M = \min(I, J-1)$

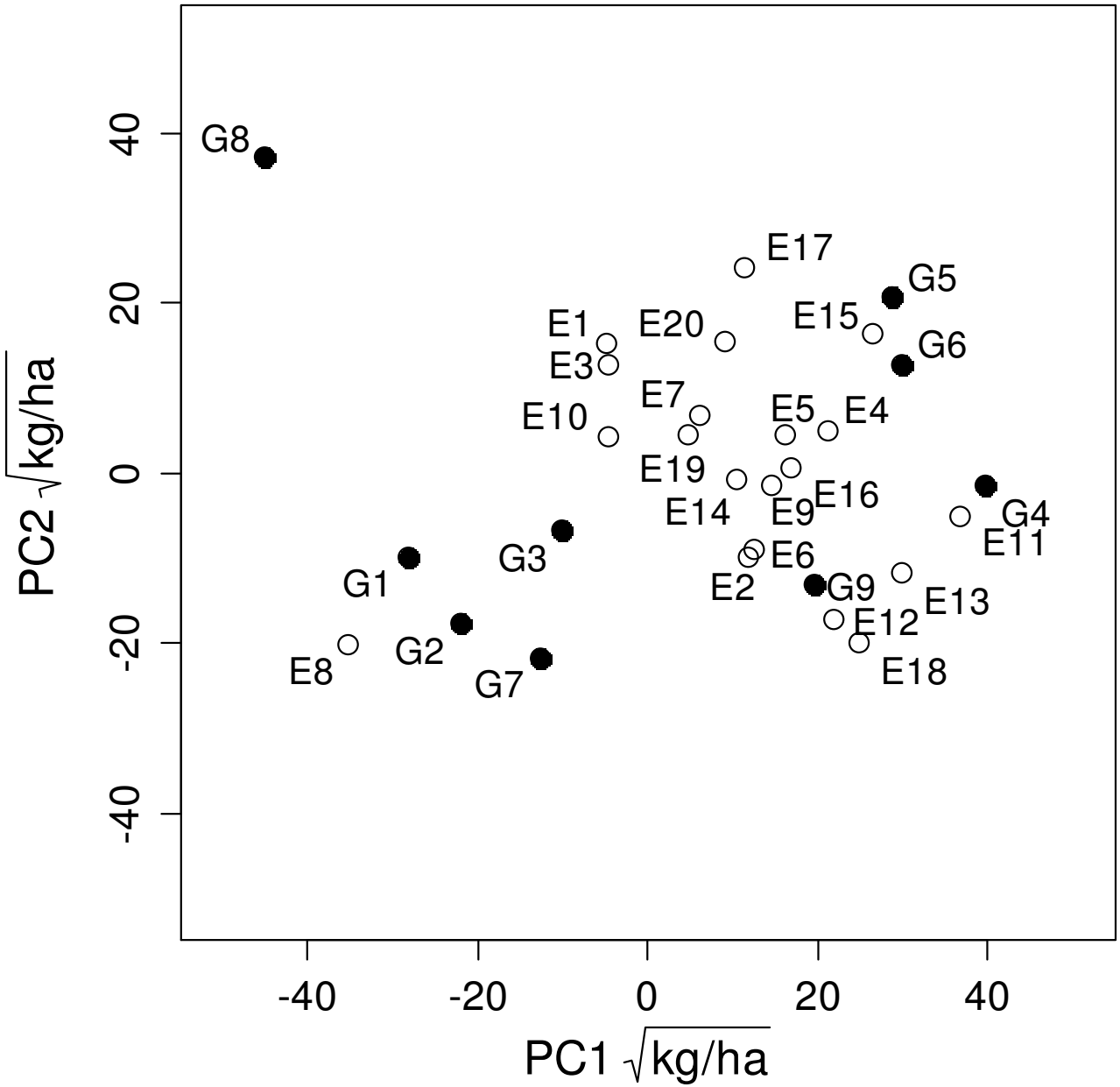
Principal components

Environment-PC: $\hat{\boldsymbol{\gamma}}_1\hat{\lambda}_1^c \quad , \quad \hat{\boldsymbol{\gamma}}_2\hat{\lambda}_2^c \quad , \ldots , \quad \hat{\boldsymbol{\gamma}}_M\hat{\lambda}_M^c$

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

$$0 \leq c \leq 1$$

GGE biplot



$(c = 0.5)$

A simple parametric bootstrap method

Johannes Forkman

BIOMETRICS DOI: 10.1111/biom.12162

Parametric Bootstrap Methods for Testing Multiplicative Terms in GGE and AMMI Models

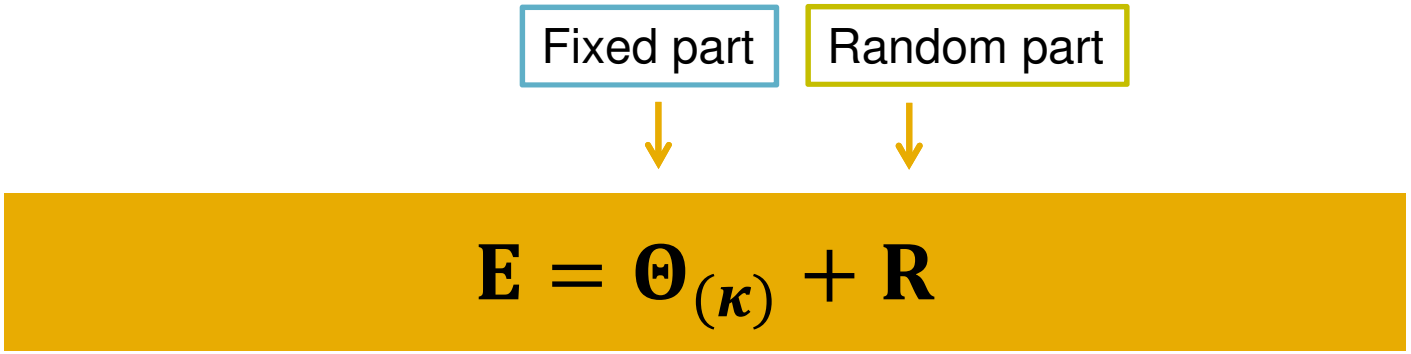
Johannes Forkman[1,*] and Hans-Peter Piepho[2]

[1]Department of Crop Production Ecology, Swedish University of Agricultural Sciences, PO Box 7043, 750 07 Uppsala, Sweden
[2]Institute of Crop Science, University of Hohenheim, 70 593 Stuttgart, Germany
*email: johannes.forkman@slu.se

SUMMARY. The *genotype main effects and genotype-by-environment interaction effects* (GGE) model and the *additive main effects and multiplicative interaction* (AMMI) model are two common models for analysis of genotype-by-environment data. These models are frequently used by agronomists, plant breeders, geneticists and statisticians for analysis of multi-environment trials. In such trials, a set of genotypes, for example, crop cultivars, are compared across a range of environments, for example, locations. The GGE and AMMI models use singular value decomposition to partition genotype-by-environment interaction into an ordered sum of multiplicative terms. This article deals with the problem of testing the significance of these multiplicative terms in order to decide how many terms to retain in the final model. We propose parametric bootstrap methods for this problem. Models with fixed main effects, fixed multiplicative terms and random normally distributed errors are considered. Two methods are derived: a *full* and a *simple* parametric bootstrap method. These are compared with the alternatives of using approximate F-tests and cross-validation. In a simulation study based on four multi-environment trials, both bootstrap methods performed well with regard to Type I error rate and power. The simple parametric bootstrap method is particularly easy to use, since it only involves repeated sampling of standard normally distributed values. This method is recommended for selecting the number of multiplicative terms in GGE and AMMI models. The proposed methods can also be used for testing components in principal component analysis.

KEY WORDS: AMMI; Genotype–environment interaction; GGE; Multi-environment trials; Principal component analysis; Singular value decomposition.

The null model



The rank of $\mathbf{\Theta}_{(\kappa)}$ is κ

$\mathbf{R}$ is a matrix of independent $\mathrm{N}(0, \sigma^2)$ distributed errors

κ is the actual number of principal components

The null hypothesis

$$H_0: \kappa = K$$

$$H_1: \kappa > K$$

Test sequentially: $K = 0, 1, 2, \ldots$
until a non-significant result is obtained.

Test statistic

To test the significance of the $(K+1)$th component, use

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

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.

The simple parametric bootstrap method

Why is it **simple**?

- We don't need to estimate any parameters

Why is it **parametric**?

- We still have to assume normal distribution

Results for the maize dataset

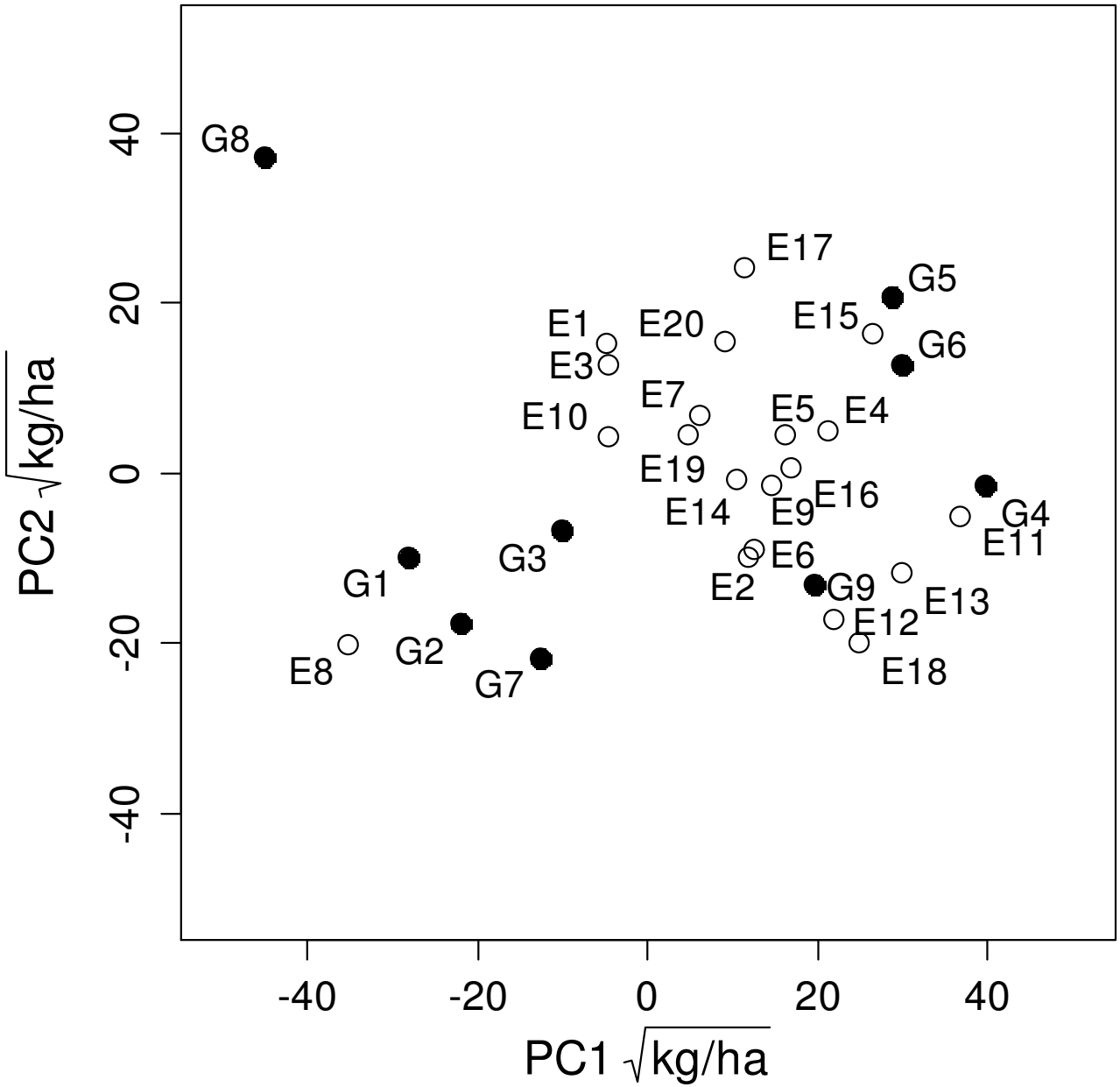
Start

$K+1$	T	p-value
1	0.640	0.000
2	0.319	0.296

Stop

The first principal component was significant, but the second was not.

GGE biplot



$(c = 0.5)$

Principal component analysis

Johannes Forkman

PCA

- Let $\mathbf{X}$ be a column-wise mean-centered matrix
- The singular values of $\mathbf{X}$ (and $\mathbf{Y} = \boldsymbol{X}^{\mathrm{T}}$) can be denoted

$$\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_M.$$

- 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 $(\hat{\lambda}_1^2, \hat{\lambda}_2^2, \ldots, \hat{\lambda}_M^2)/(J-1)$
- The $(K+1)$th principal component accounts for

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

per cent of the residual sum of squares.

Conclusion

- The simple parametric bootstrap method can be used to test the significance of the principal components

Requirement

- Random errors are independent, normally distributed and homoscedastic

Summary

- The GGE analysis is a PCA with environments as variables and genotypes as observations
- The significance of the principal components can be tested using the simple parametric bootstrap method

Summary

Thank you for your attention

- The GGE analysis is a PCA with environments as variables and genotypes as observations
- The significance of the principal components can be tested using the simple parametric bootstrap method

Forkman, J., and Piepho, H. P. (2014). *Biometrics*. In press.

Yan et al. (2000). *Crop Science* 40, 597-605.