) An easy introduction to biplots for multi-environment trials

Article · January 2012

CITATIONS
6

READS
1,093

1 author:

Ric Coe
Consultative Group on International Agricultural Research
169 PUBLICATIONS **6,262** CITATIONS

SEE PROFILE

An Easy Introduction to Biplots for Multi-Environment Trials

Brigid McDermott (br.mcdermott@gmail.com), Statistical Services Centre, University of Reading, UK

Ric Coe (r.coe@cgiar.org), Statistical Services Centre, University of Reading, UK and World Agroforestry Centre, Kenya

7 April 2012

Foreword

This manual is a simple and non-mathematical introduction to the use of biplots for understanding multi-environment trial data. The focus is on the GGE biplots (Yan and Kang, 2003) though many of the concepts explained are equally useful for understanding other biplots. The authors claim no originality for either content or presentation. As this is just an introductory exposition to the subject, readers looking for more information are directed to some of the excellent materials referenced. However, most other introductions to biplots begin, logically, with a presentation of the singular value decomposition model and many agriculture students can find this intimidating. Before any discussion of the math involved, this guide takes students from familiar line graphs, through simple genotypes plotted on axes representing environments to biplots on rotated axes and simple ideas of projection. The intent is to give readers an intuitive feel for what is happening in the construction of a biplot and an appreciation of the usefulness of biplots so that they will be motivated and confident enough to learn more.

The authors wish to thank CIMMYT and Mr. Titus Kosgei, an MSc graduate at the University of Nairobi, for making available a portion of their maize trial data for use as an example in this manual. The authors also wish to thank Maxwell Mkondiwa (Unda College, Malawi) and Rebecca Selvarajah-Jaffery (GreenInk) for commenting on content and language respectively.

1 Getting Started

1.1 Why do I have to learn about biplots?

Well, the short answer is that you don't. You can be a plant breeder without using biplots; after all, many varieties were released before biplots were used. However, as a plant breeder you need to be able to understand lots of information from your trials relating to line entries (genotypes*), plot locations, seasons and crop traits. It is hard to make sense of hundreds of numbers on a spreadsheet. You need many methods to summarize this data and help you make decisions, like which genotypes preformed best in which environments and which lines to advance to the next generation. You will also probably want to understand the nature of genotype by environment (GxE) interaction so you can respond to it. As breeders become more aware of the social side of variety development and deployment, they are thinking that 'environment' can include social dimensions. Then we can use GxE ideas to understand such things as how different social groups prefer and use different genotypes, and end up with a similar data analysis problem. Biplots are just one tool to provide you with this summary information. We believe they are very powerful tools for drawing meaning from your plant breeding data. We hope you will decide to read on and judge their usefulness for yourself.

1.2 Graphs help to answer breeding questions

Many types of graphs, not just biplots, can be useful in the plant breeding process. Graphs help to visualize relationships between measurements on things like genotypes, environments*, crop traits and genetic structure. Think about some of the questions you might want to answer as a plant breeder:

- Which genotypes yield better than others in a given location?
- Which genotype or group of genotypes yield best in which environments or group of environments?
- Does the same genotype, or group of genotypes, yield best at a given location over a number of seasons?
- Do the same genotypes tend to do well over a range of environments or do different genotypes consistently perform better in different environments?
- In how many and in which locations* should I plant my breeding trials?
- For how many years do I need to run my variety trials at the same location?
- Do some locations require separate variety development?
- Can we group environments into those which are similar for variety performance?
- Which genotypes are best for specific plant traits or for a combination of plant traits?
- Does plant trait performance differ by environment?

Graphs of data from breeding trials can help you pull the information from the trial data to answer these questions. Other methods such as analysis of variance, model fitting and estimation are still important, but graphs are a central part of analysis of GxE data. Let's look as some examples.

Box 1

**These definitions are modified slightly from common use and used in this guide to help simplify the explanations.*

Gentoype: Crop line, entry, or variety on which the breeder has or is collecting performance and trait information.

Location: A physical place that crop genotypes are planted and grown.

Environment: The combination of physical attributes of a location and the climatic and other attributes of a specific season (i.e. soil type, fertility, topography, temperature, rainfall, pest/disease challenge) that affect the plant growth. Thus, for the purposes of this manual, a genotype grown in a specific location for a specific season is subject to a single environment. Genotypes grown in the wet season of 2010 at location A and the wet season of 2011 at location A are being grown in two different environments. We can also think of the environment as including both the way the crop is managed and the social context in which it is grown and assessed. Sometimes the interaction of these with genotype is described as 'GxExMxC' with M standing for management and C for culture. But you can also simply include these as aspects of E.

1.3 A commonly used line graph

Figure1-1 displays a common point and line graph giving the yield of eight genotypes under consideration in 24 environments. This has been drawn from the data in Table 1-1. This is the sort of table that is generated by multi-environment trials. Each number in the table is usually the mean from several replicate plots in that environment. The analysis of data from each environment on its own is important, but not what we are exploring here. Further, each of those means has some uncertainty or 'error' attached to it (measured, for example, by a standard error), but we are not thinking about them either. Let's just see if we can understand the patterns in that large G by E table.

Table 1-1. Yields of eight genotypes in 24 environments (tonnes/ha)

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15
G1	0.79	2.11	6.21	0.56	4.06	2.13	5.61	0.20	3.32	3.01	5.12	0.77	0.78	0.62	2.00
G2	0.59	0.92	10.11	0.74	4.01	1.12	4.58	0.70	2.61	1.49	4.59	0.20	1.33	0.62	2.67
G3	1.18	3.44	4.08	0.50	6.91	4.27	6.87	0.30	4.57	2.88	6.61	0.59	0.97	1.02	1.45
G4	2.17	4.46	6.51	0.35	4.80	5.99	5.44	1.20	4.21	3.59	4.89	0.39	2.26	2.05	3.33
G5	0.99	4.83	6.71	1.43	5.83	2.95	3.66	0.50	1.54	2.55	5.70	1.83	0.39	1.44	1.66
G6	1.57	3.94	7.49	0.61	5.88	2.74	7.22	1.19	4.81	3.72	5.59	0.38	1.36	1.43	3.94
G7	2.38	4.74	9.94	1.31	5.74	5.59	9.01	0.70	4.47	3.59	4.85	3.48	2.35	1.23	3.22

G8	**0.98**	**2.56**	**12.02**	**3.31**	**5.41**	**2.03**	**5.92**	**0.30**	**4.39**	**4.19**	**6.10**	**0.39**	**1.97**	**0.62**	**4.93**

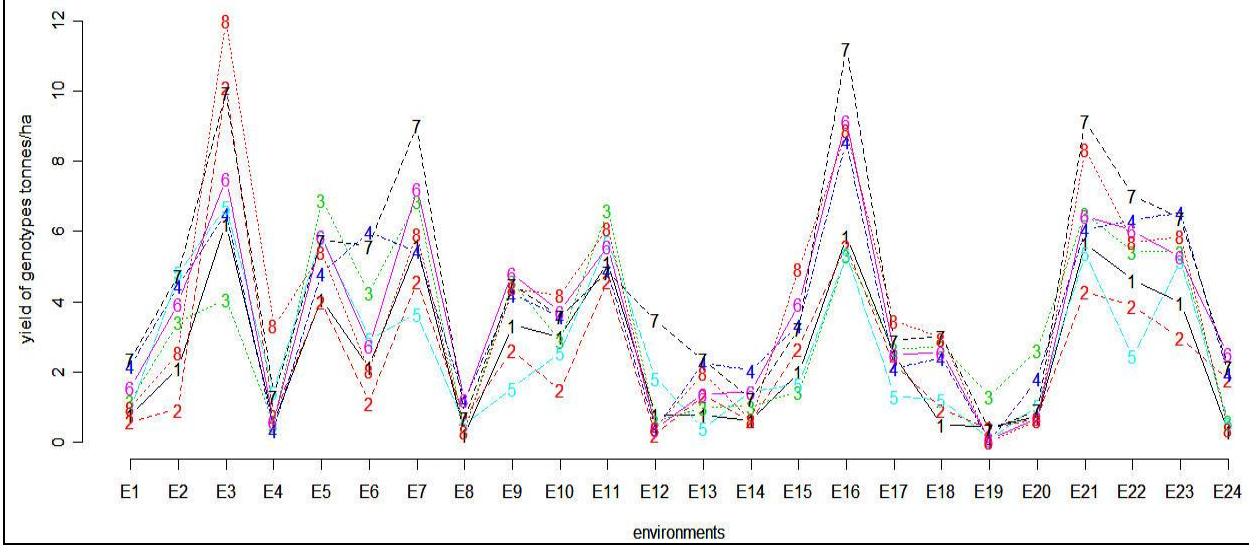
	E16	E17	E18	E19	E20	E21	E22	E23	E24
G1	**5.87**	**2.50**	**0.49**	**0.42**	**0.73**	**5.66**	**4.61**	**3.98**	**0.32**
G2	**5.58**	**2.22**	**0.93**	**0.40**	**0.68**	**4.28**	**3.89**	**2.98**	**1.78**
G3	**5.32**	**2.65**	**2.73**	**1.31**	**2.63**	**6.51**	**5.42**	**5.43**	**0.60**
G4	**8.55**	**2.10**	**2.39**	**0.03**	**1.81**	**6.05**	**6.31**	**6.53**	**1.94**
G5	**5.36**	**1.32**	**1.19**	**0.08**	**1.03**	**5.38**	**2.45**	**5.18**	**0.54**
G6	**9.15**	**2.50**	**2.56**	**0.08**	**0.68**	**6.46**	**6.02**	**5.28**	**2.52**
G7	**11.21**	**2.89**	**3.01**	**0.33**	**0.91**	**9.15**	**7.03**	**6.39**	**2.13**
G8	**8.89**	**3.45**	**2.95**	**0.01**	**0.62**	**8.35**	**5.71**	**5.86**	**0.38**

In Figure1-1, you can easily see which genotypes do well in which environments, and which environments tend to have higher yields or lower yields on average. It is clear that there is some genotype by environment interaction because you can see that, for example, genotype 3 did well in environment 11 but poorly in environment 15. When genotypes actually change ranking from environment to environment this is often called "cross-over" effect. For more details on cross-over effect, see Box 2 below.

You can also see from Figure1-1 which environments have the most variable yields. Notice that environments E3, E7, E16, E21 and E22, have a wide range between their lowest and highest yields. You will note too that those environments with a wide range of yields tend to have high average yields. A difference in yield of 1.0 t/ha means something different depending on whether you are talking about E3 or E14. Increasing the yield of the lowest yielding variety in E14 by a tonne a hectare will bring it almost equal with the highest yielding variety in that environment, whereas you would need to increase the poorest yield in E3 by about 8 tonnes/ha to equal the best performer in that environment.

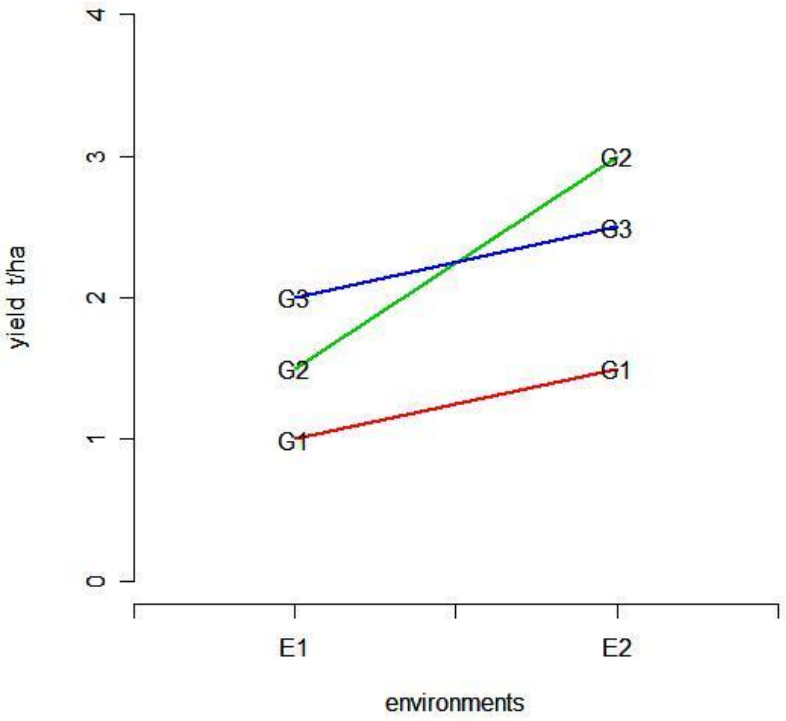
With a bit more difficulty you see which genotypes have similar yields and which environments have similar performances of genotypes. However, these comparisons would become very difficult to see if there were even more genotypes displayed on the graph.

Figure1-1. Yield of eight genotypes in 24 environments.



Box 2. Genotype by Environment Interaction and Cross-Over Effect

Figure 1-2 Plot showing Cross-Over Interaction between G2 and G3

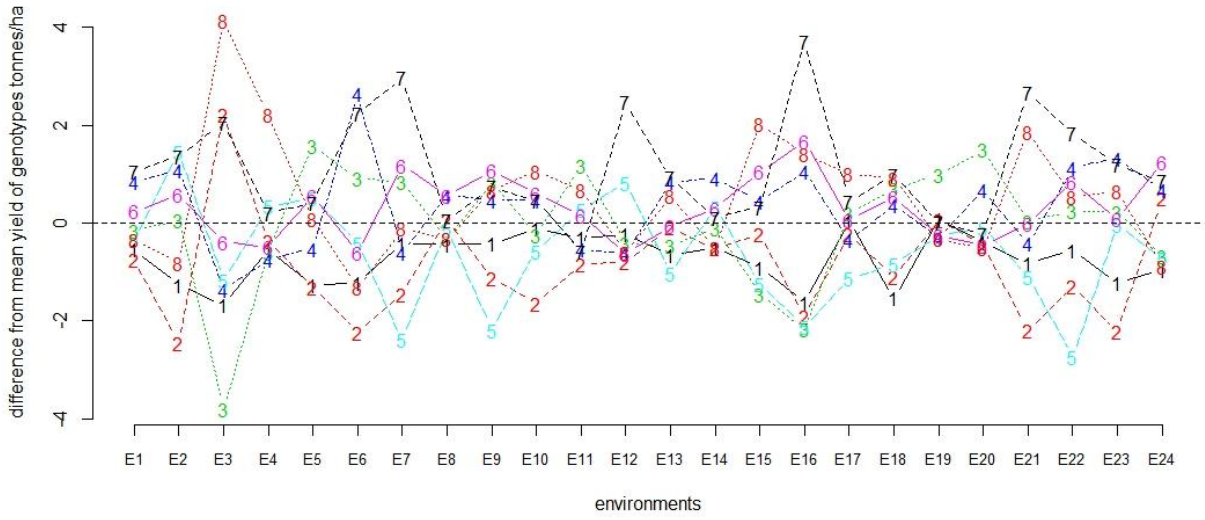


With a simpler example as in Figure 1-2 we can see some of the features that need to be understood in larger, messier sets of data such as Figure 1-1. The three genotypes here each do better in E2 than in E1. When you look at G1 and G2 you can see there is a GxE interaction. G2 is better than G1 in both environments but the difference is larger in E2. This shows up as the G1 and the G2 lines on the graph not being parallel. There is also an interaction of G2 and G3 with environment. G3 beats G2 in E1, but their ranking is reversed in E2 where G2 beats G3. This is shown as the G2 and G3 lines on the graph crossing each other. For this reason it is sometimes called a 'cross-over interaction'. If looking for the best variety (based on this trait only) of G2 and G1, you would select G2 in both environments. But if looking for the best of G2 and G3 you would select G3 in E1 but G2 in E2.

Suppose the high yields in some environments were due to good soils while environments with poor yields had unusually poor soils. We know yield will be generally lower on poor soils and that the environmental variation is not what is of interest. In that case we might not want to consider the relative performances of environments and wish to remove this soil effect from the graph. To do this, all we have to do is subtract the mean yield over all genotypes at each location from the yield of each genotype at that location. This will give a mean yield at all environments of zero. Now our graph in Figure 1-3 shows the relative performance of the genotypes at each location and removes the environment to environment

variation. We can immediately see which genotypes are yielding above and below average at a given environment, as well as the ranking of genotypes by yield at each environment.

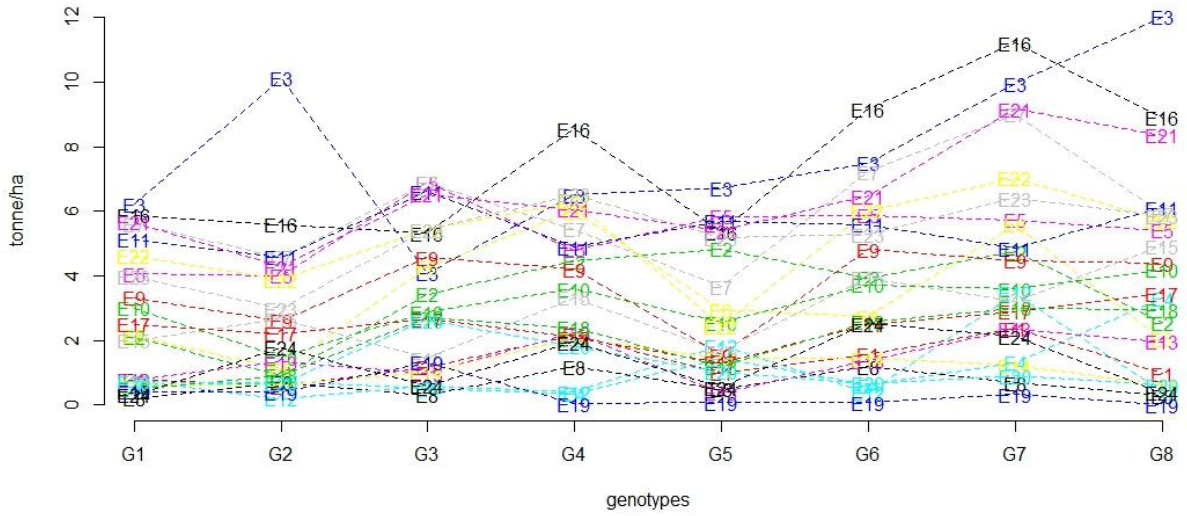
Figure 1-3. Environment centered genotype yields by environment



The standard line graph of yield of genotypes versus environments is very useful. It shows us the ranking by yield of the genotypes at each environment, with a feel for the size of the difference between genotype yields. We can easily pick out in which environments a particular genotype yields well and in which it performs poorly.

What if we want to answer questions about environments? In which environments are there a wide range of yields, so that the trials held in that environment really discriminate between genotype performances? In which environments is the performance of genotypes very similar so we may decide to plant trials in only one of the locations associated with these environments next season. The first question may be answered by examining the graph in Figure 1-3, but in judging similarity of performance between two or more environments, the graph in Figure 1-4 is helpful where the yield for each environment is plotted against genotype. Here we can see there are a number of low yielding environments crowding the lowest quarter of the graph in which there is little difference in the yield of the eight genotypes under consideration. By contrast the environments 3, 7, 16, and 21 have visible differences in the performance of different genotypes, and it would be interesting to explore why this might be so.

Figure 1-4. Yield genotypes in environments by genotype



In all of these graphs we have been looking at the attribute of yield only. We don't have a way of examining yield, date of flowering, leaf area index and other attributes on the same graph. It would be nice if we could also make the comparison between the yield for similar genotypes and similar environments easily on the same graph. For example, is it easy to spot which genotypes respond to environment in a similar way in Figure 1-3**,** or which environments are similar in terms of genotype performance in Figure 1-4?

To accomplish this we need to move to a different type of graph that is designed to show this sort of information. That is what the biplot is. First we will look at a very simple example to show how the biplot is related to more familiar graphs, and then look at real examples and show how to interpret them.

Box 3. Definition of a Biplot

The "Bi" in biplot refers to the display to two different types of things, like genotypes and environments, on the same graph. It does not refer to the fact that the graph has two axes. A biplot is a two dimensional representation of a matrix, or table, like the table of mean yields by genotype and environment in Table 1-1

1.4 A very simple example

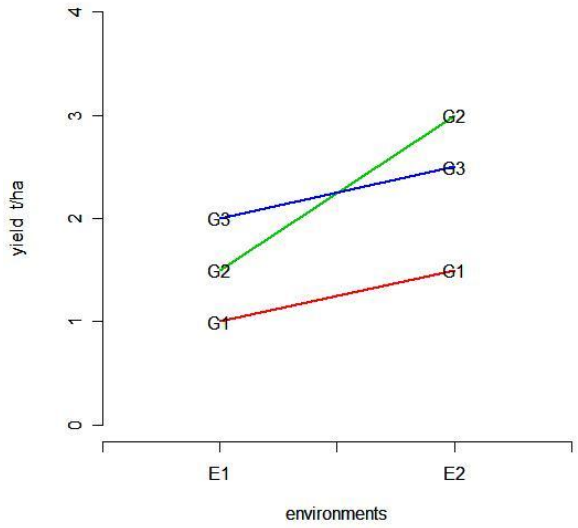
Let's look at a simple example that shows us the performance of three genotypes in two different environments: E1 andE2. We will use the same data as in the simple example in Box 2 on GxE interaction and cross-over. Also, we will assume one yield value for every genotype by environment combination so that we ignore any variation between yields for different replications (plots) for the same genotype at the same location.

Table 1-2 gives us the data and Figure 1-5 plots this data using the same type of line plot we saw in Section 1.2. With so few genotypes it is easy to see the performance of the three genotypes at the two locations.

Table 1-2. Yield of genotypes G1, G2 and G3 (t/ha) at sites E1 and E2

	Sites	
Genotypes	E1	E2
G1	1	1.5
G2	1.5	3
G3	2	2.5

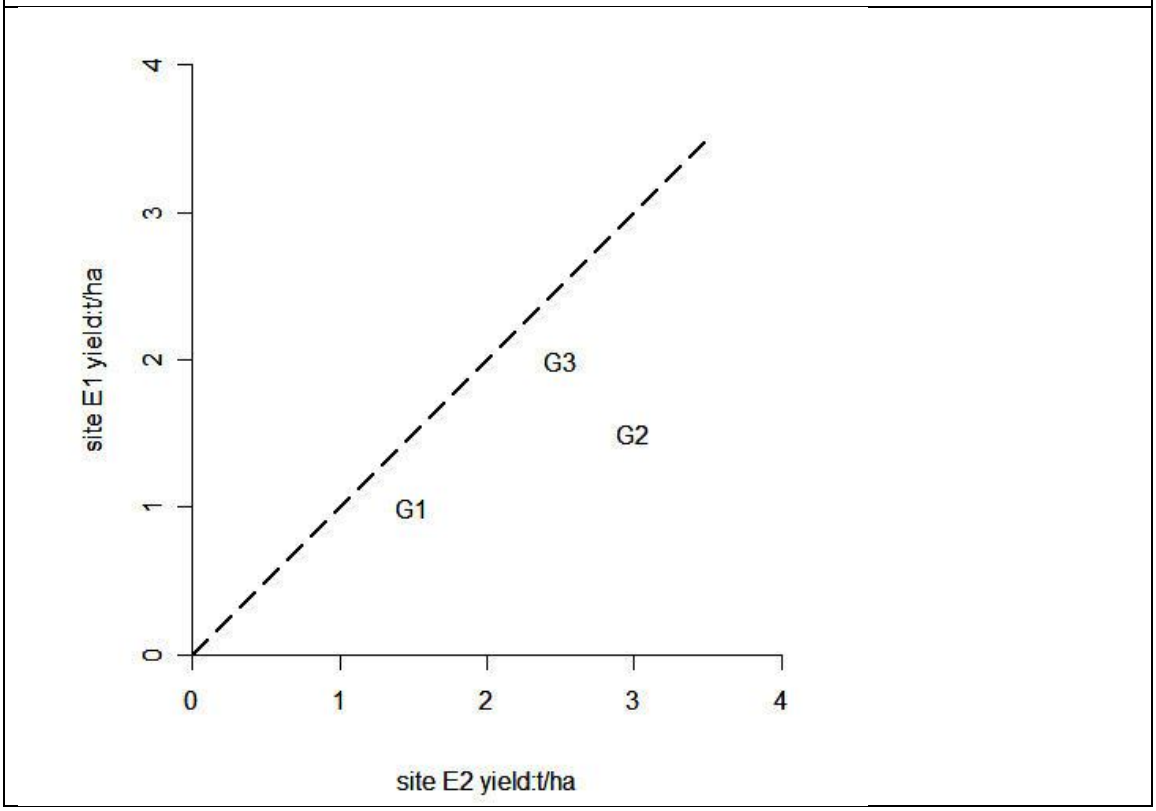
Figure 1-5. Line graph of yield of genotypes G1, G2 and G3 (t/ha) in sites E1 and E2



In Figure 1-6 we present a different plotting of the yield of each genotype with the "x" co-ordinates defined by the yield at site E2, and the "y" coordinates defined by the yield at site E1. The diagonal line across the plot shows us the points on the graph where the yield would be equal in sites E1 and E2. Thus if a genotype marker was plotted right on this dashed line, we would know that it had produced the same yield at the two sites. The closer a genotype marker is to the dashed line, the more equal their performance in the two environments. When a genotype's performance tends to be the same, consistently, over a number of environments we say that the genotype's performance is stable. In Figure 1-6 G1 and G3 are more stable than G2.

This graph makes it very easy to see which genotype yielded the highest in which environment. Since all genotypes are to the right on the dashed line, all genotypes yield better at site E2. If one of the genotype markers had been to the left of the dashed line we would know instantly that that genotype did better in E1 than in E2. The further to the right the point is on the horizontal axis, the higher the yield in site E2. If the point is further towards the top on the vertical scale, the higher the yield at site E1. We know there is cross-over effect if the ranking of the genotypes in the horizontal and vertical directions is different. The vertical ranking in E1 (lowest to highest yield) is G1, G2, G3, whereas the horizontal ranking in E2 is G1, G3 and G2. Again, we see a cross-over effect for G2 and G3 for these two environments.

Figure 1-6. The yield of the three genotypes plotted on axes defined by the sites E2 and E1



We imagined for this simple example that the soil structure was better at site E2 than at site E1 and so all genotypes yielded better at the second site. Let us add the mean (green line) of the yield for E2 and the mean of the yield at E1 (blue line) in Figure 1-7. Now we can see which genotypes are yielding above or below average at a given environment.

The main effect of E2 can be thought of as the mean yield at E2 minus the mean yield at E1 [2.33-1.5=0.833]; we add it to the plot as the red diagonal line. What this line is saying is that on average we would expect the yield of any genotype at E2 to be 0.833 tonnes more than its yield at E1. If there is little interaction the plotted genotype points will tend to lie along this red line. In fact, what we see is that G1 and G3 performed a little less well in E2 than we would expect given the main effect of E2, while G2 performed much better in E2 than we would expect based on the main effect of E2.

Generally we are not very interested in the main effects of the different environments, but rather the relative performance of the genotypes at each environment. If we subtract the mean yield at each site from the genotype yields at that site, we can shift the graph so that we center the graph around the origin of the plot (0,0 point). The data with the site means subtracted are given in Table 1-3, with the resulting graph in Figure 1-8.

Figure 1-7. Yield of three genotypes plotted against environments with environment means (blue and green lines) and main effect of environment 2 (red line) added to the plot.

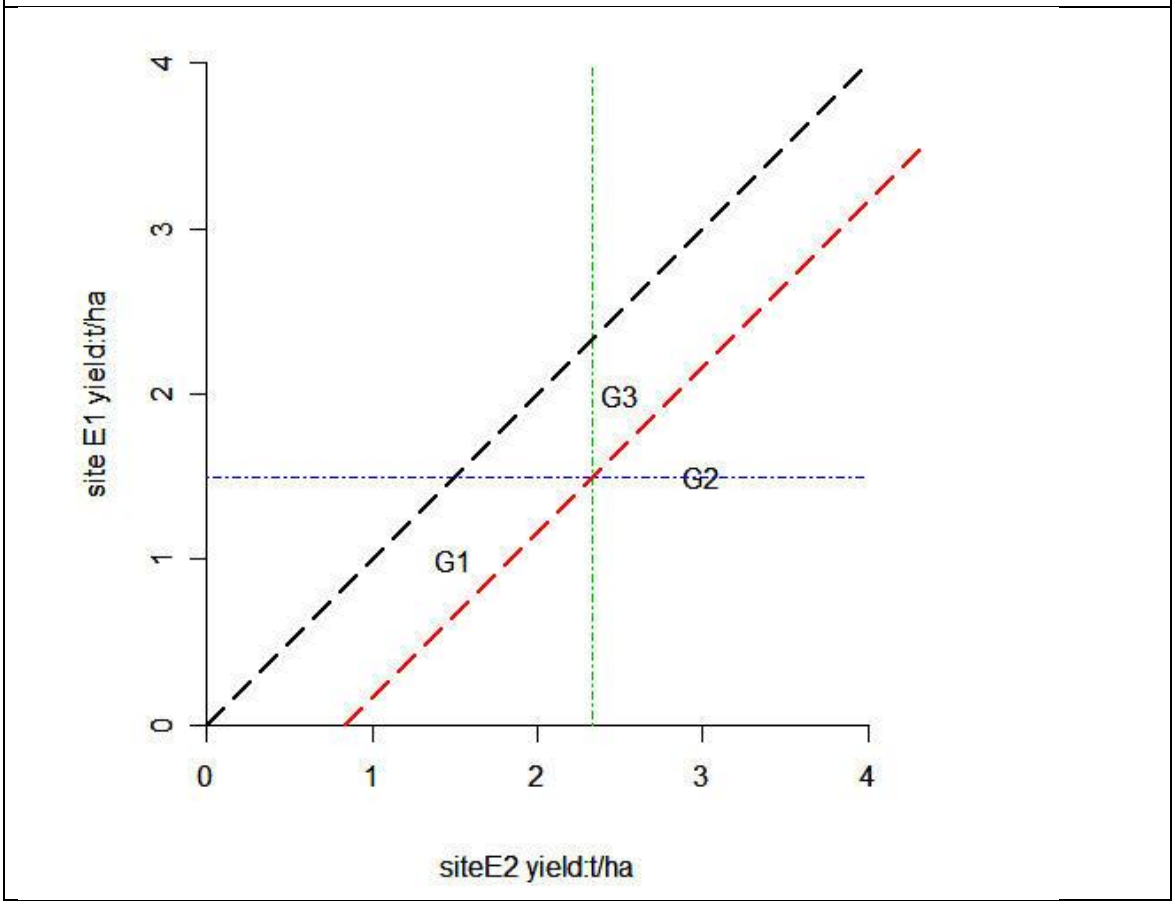
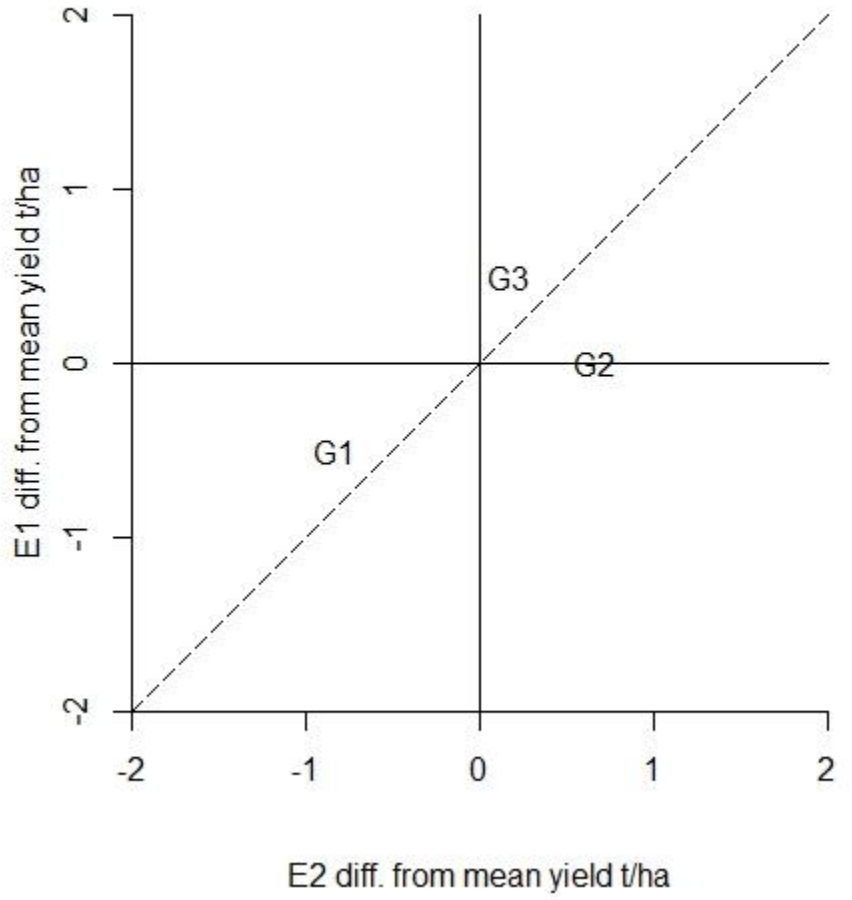


Table 1-3. Yield data with environment means subtracted from each column

	Sites	
Genotypes	E1	E2
G1	-0.5	-0.83
G2	0.0	0.67
G3	0.5	0.17

Figure 1-8. Yield of three genotypes plotted against environments with environment effects removed. The dotted line describes points where a genotype's yield difference from the mean would be the same in both environments.



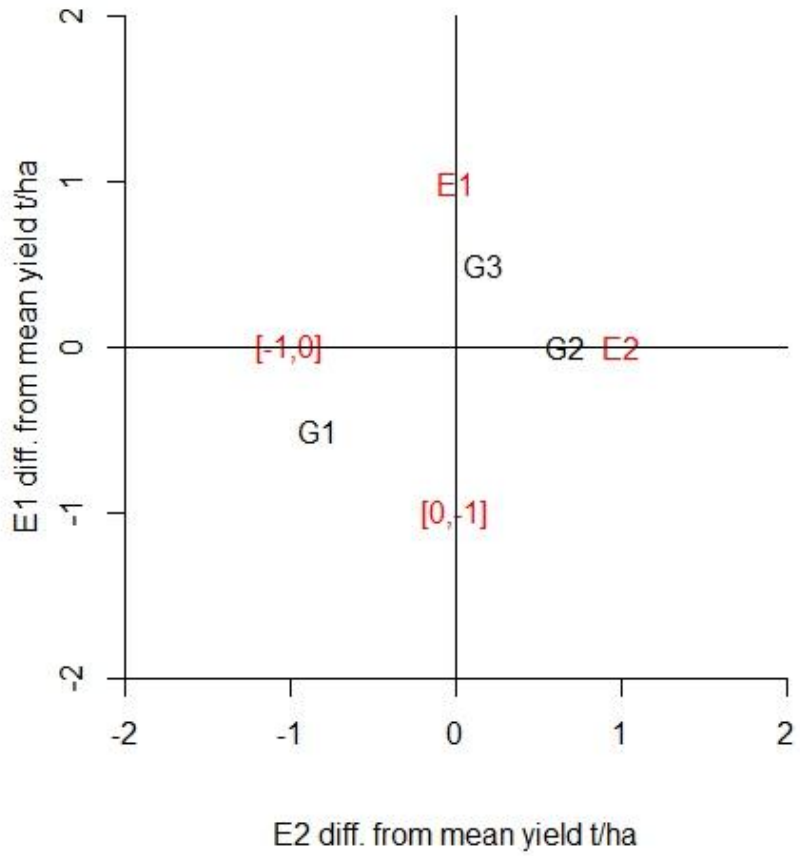
Although we have shifted the position of our points on the graph we can still draw all the conclusions we did with the un-centred plot. G1 performs below average at both sites while G2 gives the average yield at E1, but performs better than average by over half a tonne an hectare in E2. G1 and G3 are more stable than G2 because they are closer to the diagonal line, so G2 is showing more GxE interaction. We can see that there is a cross-over effect since the ranking of the genotypes is different in the horizontal and vertical directions.

1.5 Moving towards the biplot

We have seen that we can plot multiple genotypes against two axes and shift the points in the plot without losing information. What we have done so far is very useful if you have only two locations, but what are we to do with three or more environments? We might possibly graph in three dimensions but how could we graph with 24 axes for 24 environments? There are two more tricks we need to use to be able to handle many locations. The first of these is to rotate the axes. We continue to show what that means with the simple example.

Let us start by identifying points on our current E1 and E2 axes. The marker "E1" is placed at (0,1) and "E2" at (1,0); the corresponding points (-1,0) and (0,-1) are marked for reference in Figure 1-9

Figure 1-9. Yield of three genotypes plotted against environments with environment effects removed.

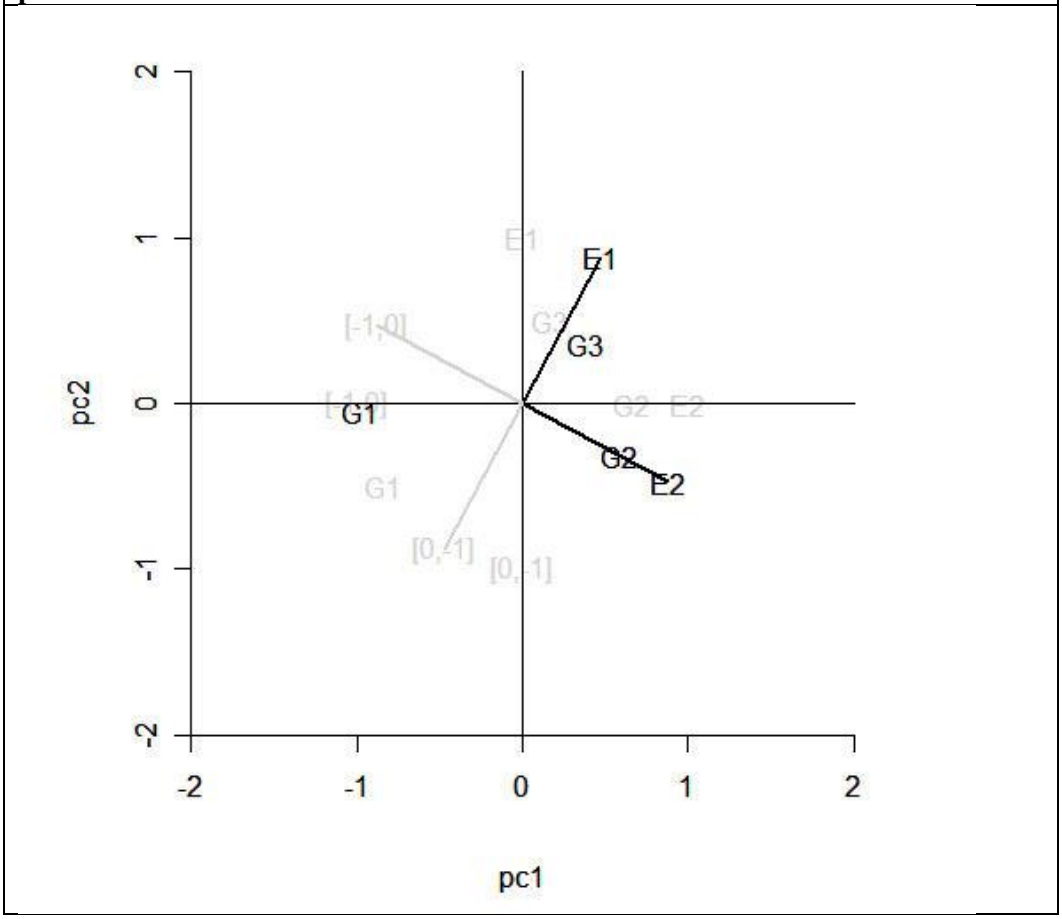


Now imagine rotating the whole diagram by taking the whole page and twisting it to the right. None of the information we can read from it changes if we look at the rotated version in Figure 1-10. So we could

draw the graph this way, with these new horizontal and vertical axes, and showing where the original axes were with those E1 and E2 markers.

Thus, we can make all the same conclusions as we did with the earlier plot. We can continue to make comparisons between the genotype points and the shifted environmental axes. We see that G1 yields below average in both environments. We see that G3 yields highest in E1, while G2 yields highest in E2, and therefore we have a cross-over effect.

Figure 1-10. Environment axes and genotype points from Figure 1-9 rotated and plotted on new axes



1.6 Starting to use the biplot

The plot in Figure 1-10 is a biplot because both the genotypes and the environments are identified on the same plot. Let's stay with our practice of using simple data sets and look at four genotypes in three environments. If we try to do the same thing as we did with two environments we have a problem. We had an axis for each environment, so now we need 3 axes. That gives a three dimensional graph that we cannot draw on paper – it looks like the graphs above but with a third axes coming up out the plane of the paper. So we introduce our last data transformation trick.

If the data were plotted in 3 dimensions, the points would form a cloud (rather than a scatter on the paper). If you looked at the cloud from a different direction it might look different. For example, in order

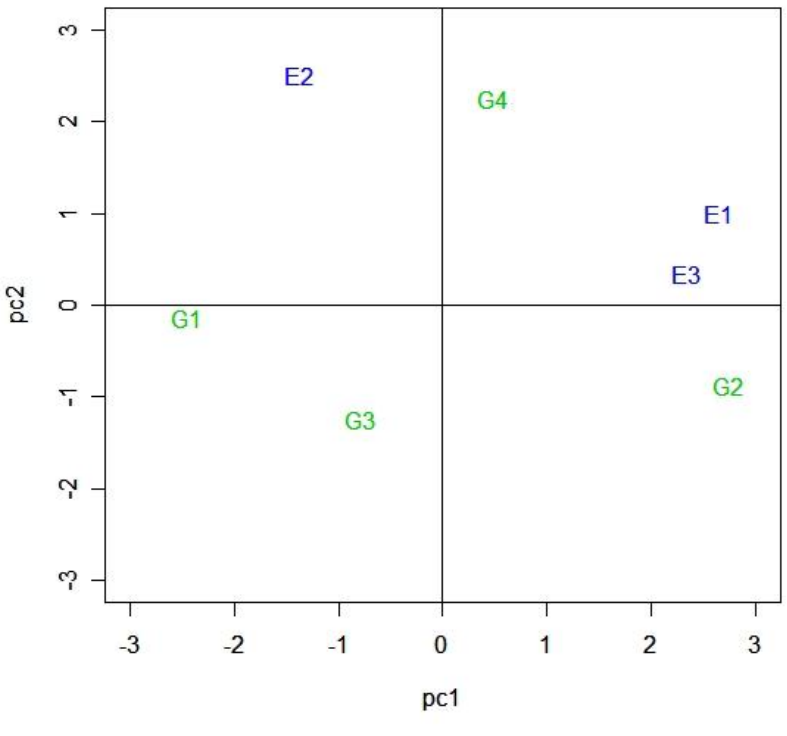
to draw the two dimensional graph, we choose to look at the cloud from the direction that shows as much variability as possible. If we had marked the original E axes, then we can also see where they are when viewed from this direction. Hence we can draw the G and E points on a graph again – a biplot, as they are both represented on the same plot. The two dimensional plot is an approximation. When we look at the cloud from one particular direction then some of the pattern is hidden. Whichever biplot software you use will use a mathematical technique so that you look at the cloud in the direction that gives you the best two dimensional approximation of the cloud of points. In fact, the labels on the axes "pc1" and "pc2" refer to first and second principal components which come from the technique used to get the best two dimensional view. We will discuss this more in Section 3.2.1

The data are given in Table 1-4 and the biplot for these data in Figure 1-11 below. Again it is a biplot because the graph contains information for both genotypes and environments on the same axes. You can see that we could easily add points indicating more genotypes and more environments.

Table 1-4. Simulated yield data for three genotypes at four environments

	E1	E2	E3
G1	7	14	2
G2	20	5	14
G3	10	9	6
G4	17	16	10

Figure 1-11. A biplot for the data in Table 1-4

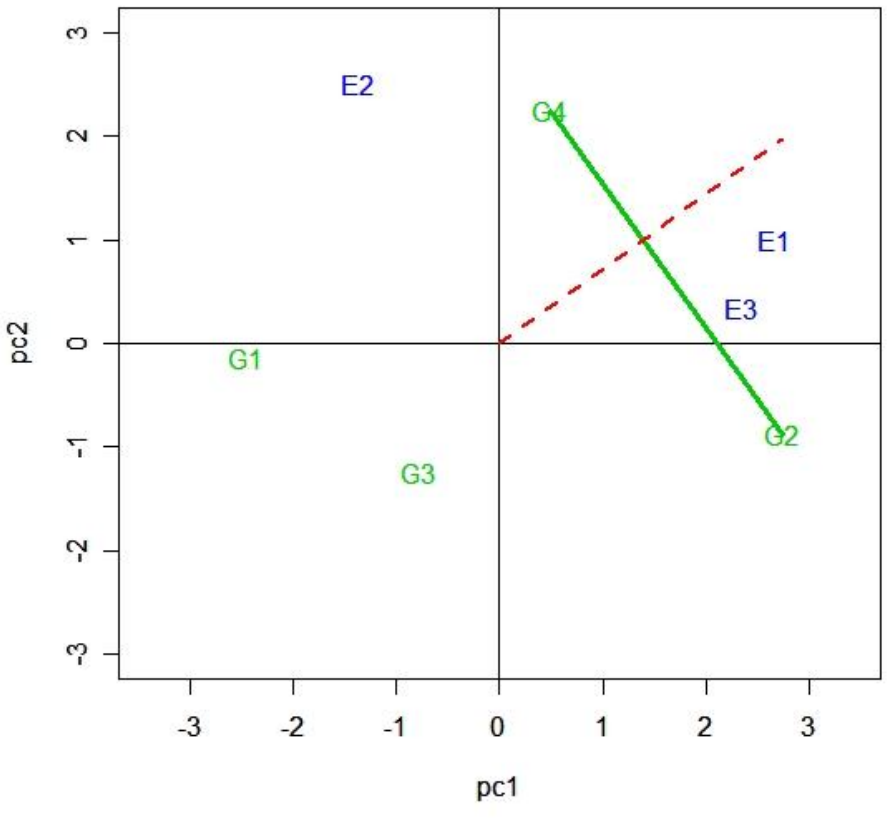


Just as with our earlier example of three genotypes in two environments, the main effect of environment is removed from the points before plotting. We are looking at the relative performance of each of the genotypes in each of the environments. The graph does not give us, at least directly, the absolute yield of

a given genotype in a given environment, but it does give us a good feel for which genotypes are doing well in which environments. It can also give us a feel for which environments are similar in terms of the performance of genotypes, and which environments differ in the genotypes giving the highest yield. For example, E1 and E3 are the most similar environments (they are closest together on the graph) and G1 and G3 are the most similar genotypes.

Actually, the biplot can clearly show us which genotypes did best in which environments. Let us look at the performance of G4 and G2 on the right hand side of Figure 1-11 by drawing a line between these two points as shown below in Figure 1-12. We then draw a line from the graph origin (0,0) that cuts the line between the genotypes at a ninety degree angle (i.e. the red dotted line is perpendicular to the green line). If any environment point lies on the red dotted line this means that genotypes G4 and G2 would be producing equal yields in that environment. Thus if an environment point lies to one side of the red line then this indicates that the closer genotype gave a higher yield in that environment. You can see from Figure 1-12 that G2 was higher yielding in both E1 and E3 than G4.

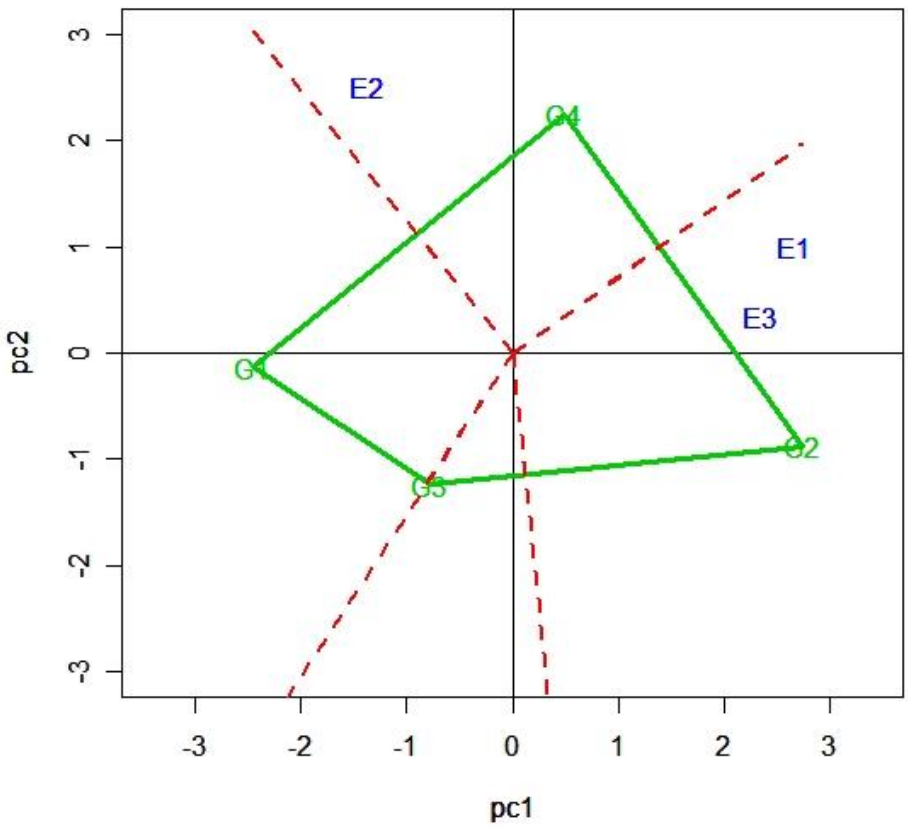
Figure 1-12. Added lines to Figure 1-11. A biplot for the data in Table 1-4 to compare the performance of G2 and G4 in the available environments. The red dashed line marks environments for which the yield of G2 and G4 would be equal.



We could continue comparing genotypes by drawing connecting segments between all the genotypes and then creating lines that cut these segments perpendicularly. This is done in Figure 1-13, producing what is known as a "what-won-where" biplot [Yan and Kang, 2003]. It is clear from Figure 1-13 that G2 gave the highest yields in E1 and E3, while G4 gave the highest yield in E2.

Please note: The perpendicular line does not necessarily have to cross the line connecting the genotypes near the middle of this connecting line. Notice the perpendicular line crossing the line between G1 and G3, the perpendicular separating (red) line is actually crossing almost on top of the G3 point. You may even see cases where the separating line is actually outside the two genotype points being compared but you have to imagine the extension of the line connecting the two genotype points to permit a line drawn from the origin to hit it at a ninety degree angle.

Figure 1-13. "What-Won-Where" Biplot for the environment centered data of Table 1-4



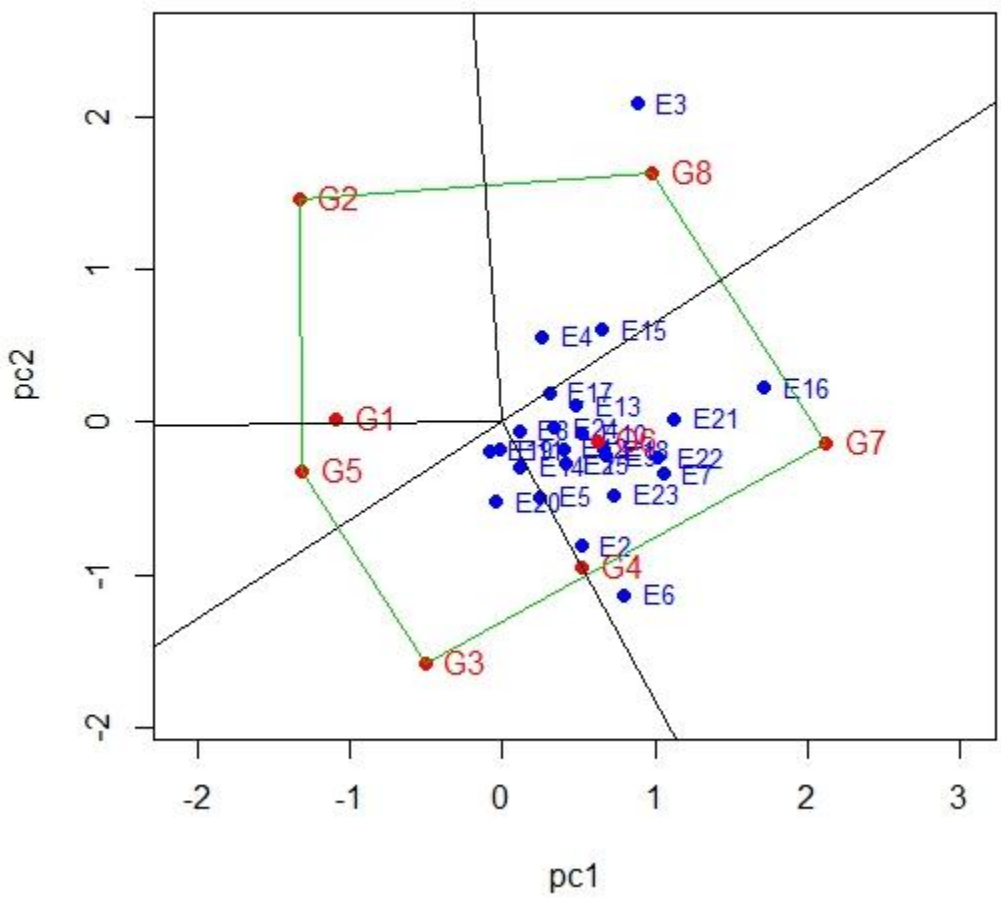
2 First Steps in using biplot information for breeding decisions

Let's move to using real data so we can begin to see why the biplot is so useful. In Figure 2-1 below we see a "what-won-where" biplot for the eight genotypes in 24 environments shown earlier in Section 1.3 in Table 1-1. For now we will continue to ignore the question of how we reduce data in multiple dimensions to be plotted in two dimensions; the data presented in the biplot is not exactly the same as that given in Table 1-1. Rather, it is a simplification of the data, allowing it to be plotted in two dimensions and illustrating some of its most important features. More details are discussed later in Section 2.1.1.

2.1 Choosing genotypes for environments

Figure 2-1 gives us a "what-won-where" biplot for the eight genotypes in 24 environments. You will notice that only the outer genotypes are connected. It is clear that G7 gives the highest expected yields in the majority of environments. Genotype G8 is the "winner" in environments E3, E4, and E15 while G3 wins in E5, E14, E11, E19 and E20, although this is admittedly a little hard so see as the environment points tend to crowd together. (Look for the points rather than the labels). So this biplot is showing us some cross over effects. If you refer back to Figure 1-3 you will see that the same plot predicts the winners pretty well. The only discrepancies are where there are very small differences between the top genotypes in an environment. So while the biplot is an approximation, it is a pretty good one, at least for looking at what-won-where.

Figure 2-1. "What-Won-Where" Biplot for environment centered data of Table 1-1



Remember that we said the dividing lines, drawn out from the origin, mark the points of equal yield between the genotypes they divide. Thus, as we see that E17 sits right on the dividing line between G8 and G7 we know that, according to this biplot, G8 and G7 are expected to give the same yield at E17. In the same way we see that the points for E5 and E14 are only slightly to the G3 side of the dividing line, between G3 and G7, so although G3 is expected to out yield G7 in these two environments it is not likely to out yield G7 by much. On the contrary E3 is far from the dividing lines between G8 and G7 and the dividing lines between G8 and G2. Here the biplot is telling us that, given the information in this one data set, it looks like G8 is uniquely adapted to the environment E3.

2.1.1 Formulating the model: What am I seeing?

How does this help us in our plant breeding decisions? It depends a great deal on whether the biplot is based on one season's data or whether the patterns we are seeing in this biplot are similar over a number of seasons. As in all agronomic research, observed phenotypic outcomes, like yield or pest damage score,

vary from plot to plot and season to season. We see any given phenotypic value, like a yield of 1.5 tonne/ha, as the outcome of the combination of the plant's genotype with characteristics of the environment, both physical and cultural, produced at a specific site during a specific period. Some genotypic effects are fairly strong and tend to be expressed repeatedly in multiple environments: we call these **genotype main effects**. Some environments have characteristics that tend to have a strong effect on the observed quantitative phenotypic response for a range of different genotypes: these are called **environment main effects**. Then we have quantitative phenotypic outcomes that can only be observed when a specific genotype is grown in a specific environment, giving us a response that is different than that we would have expected based on the genotype or environment main effects alone. When this response is observed repeatedly it is labelled **genotype by environment interaction effect**. Finally there are the components of response that are not repeatable. Examples of these include the effects of microclimate or localized soil structure on the yield of a plot, or the unusual combination of biological factors in a part of the field that escalates the pest challenge. They also include measurement errors or variation introduced by sampling. These unrepeatable interactions, often called **unexplained effects**, of genotype and environment are usually small and account for the small differences we see in, say yield, between treatment replications within the same trial. However, occasionally they can be quite large, dwarfing the usually dominant main effects.

Plant breeders often express this relationship mathematically with this model:

Response =M + G + E +GE + e

M is the overall mean of the response for all genotypes in all environments, G is the genotype main effect, E is the environment main effect, GE is the genotype by environment interaction and e is the unexplained/unrepeatable effects. You have probably created such models using the ANOVA and related analyses. If you are more used to seeing this model written in mathematical symbols, then look at Box 4.

With the results of your analysis you create a model, a conceptualization, of how some response is realized, i.e. **Plant height= G + E +GE**. This model doesn't give the height of a specific plant but estimates the average plant height of a given genotype in a given environment. How well this model estimates average plant height depends on:

i. how well this type of model fits the biological process that determines plant height;
ii. how much of the total variation in plant height is determined by the main effects and repeatable interaction;
iii. how well you were able to separate and identify the main effects and interaction, G + E + GE, from the unexplained/unrepeatable effects, e.

The biplot is a visualization of a specific model. The specifications of the model determine how we set up the biplot and, consequently, what the biplot shows us. Remember that so far we have decided to remove the main environment effects (E), by subtracting the environment means (M+E) from our model so that we have biplot models, **Response-M-E= G + GE**, portrayed in a way that we can visualize the relationships between genotypes and environments. Like the ANOVA models you have created before, the value of your biplot is determined by the three factors named above, plus how good your two-dimensional approximation is of higher dimensional data.

So we must decide on how well we have separated repeatable from unrepeatable effects in the biplot in Figure 2-1 when we attempt to use the information in the biplot to assist our breeding decisions. If the biplot is based on data from a single season, then we will be unsure if the smaller differences would be seen again in another season. For example, we might question whether G3 will continue to outperform G7 in E5, E14, E11, E19 and E20 in subsequent seasons since the biplot indicates that the expected yield for both genotypes is quite similar in these environments. The biplot indicates that yield is much larger for G8 in E3 than for the other genotypes but even this might be unrepeatable if it is based on a single replication where G8 was grown on a particularly fertile and favorably situated plot.

Unfortunately the biplot does not give us a picture of the size of **e**, the unexplained effects. These are measured by the standard error of genotype means in each environment, usually estimated from plot-to-plot replication in that environment. Research is ongoing on how to add this information to the biplot so we can get visual clues as to whether we have confidence that, say, G3 will consistently out-yield G7 in E5. Generally though, the more seasons and replications in the data used to create the biplot the more confident we are that we are looking at repeatable patterns.

Box 4. Model and ANOVA

You can write the model we gave earlier: **Response =M+ G + E +GE + e**

as $y_{ijk} = \mu + G_i + E_j + GE_{ij} + \epsilon_{ijk}$

- y_{ijk} is the response, such as yield or protein content, of the j[th]replication of genotype "i", in environment "j"
- μ is the overall mean of the responses
- G_i is the genotype effect ($\mu_i - \mu$): the mean response of genotype i minus the overall mean
- E_j is the environment effect ($\mu_j - \mu$): the mean response of environment j minus the overall mean
- GE_{ij} is the interaction effect ($\mu_{ij} - \mu_i - \mu_j + \mu$): the mean response for genotype i in environment j minus the mean response for genotype i , minus the mean response for environment j, plus the overall mean
- ε_{ijk} is the difference of the response in the j[th] replication (often the j[th] plot) from the mean response for genotype i in environment j

Often we assume that the ε_{ijk} are normally distributed with mean=0 and variance = $\sigma^2$when we do further analyses like ANOVA. You can see that we can only get an estimate of $\sigma^2$when we have replications of the genotypes in the environments

The table of data, Table 1-1, for which we make the biplot in Figure 2-1 has only one entry per genotype/environment combination, so the values are actually average yields for each genotype in each environment. So our models will be based on

$$average(y_{ijk}) = \mu + G_i + E_j + GE_{ij}$$

When we do analysis of variance we take the total sum of squares which is the addition of all the squares of the difference between each observation and the overall mean as shown in this equation

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{j=1}^{k} (y_{ijk} - \mu)^2$$

Then we break up this total sum of squares (SST) into the parts we think we can explain according to our model leaving behind the part than we cannot assign to a given cause. So according to the model we wrote at the top of this box we divide the total sum of squares as follows:

Source	Sum of Squares
Genotype	SS(genotype)
Environment	SS(environment)
GEinteraction	SS(interaction)
Unexplained	SSE (not available from average y_{ijk} model; needs replication at each environment)
	SST - μ

So when we do ANOVA we are implicitly making a model too.

2.1.2 Biplots based on multiple seasons data

What if we view the information in the biplot in Figure 2-1 as one example of a biplot coming from multiple seasons of biplots that all look very similar to Figure 2-1, so that we are more confident the what-wins-where pattern is repeatable? The genotype G7 is now seen to be particularly stable. That is, it continues to be the superior performer over a number of locations over a number of seasons. The genotypes G8 or G3 may be the top performer at a few locations but they are less stable than G7.

As a breeder you want to breed for high performance (i.e., high yield, high protein content, good taste) and that high performance should be the highest, or close to the highest, consistently in all locations within the geographical area for which your variety will be released (i.e. high stability). But, as discussed in the following paragraph, there might be times when looking for local adaptation is advantageous, and the biplots can help with that. Remember though, that the biplot does not give us directly a measure of the absolute performance (actual yields, actual disease scores). It shows us the relative performance so that we can see that one genotype does better than another in an environment or group of environments. The difference in performance between the top genotype and the second best genotype may be so small as to be trivial or large enough to be of economic importance. We would have to return to the original data (or make some calculations) to find out the actual yield or difference in performance between two genotypes.

The biplot in Figure 2-1 also indicates that we may have at least one opportunity. The genotype G8 appears to be particularly suited to the environment E3 and to a lesser extent, E4 and E15. Should we start a particular breeding effort to develop a variety for environments represented by E3, E4 and E15 based on the genotype G8? If we do so, we will be classifying the environment represented by E3 as a mega-environment. A **mega environment** is a location, or a group of locations, where the same genotypes (or group of genotypes) consistently perform the best over a number of seasons. (Yan, 2006) Put another way, two mega environments are distinguished by having the same crossover effect consistently observed over a number of seasons.

Whether or not we try and exploit this mega-environment difference by breeding for a variety for this mega-environment depends on answers to a number of questions. Does the region represented by E3, E4 and E15 contain enough farmers willing to pay for a variety developed for their area? Will the difference in performance of varieties based on genotype G8 be large enough to be economic or will it actually relate to a small, unprofitable, for example, yield increase of 5 kg per hectare, compared to the other available varieties. If the answer to these and similar questions is negative, we may choose to try and understand what traits make G8 suited to these locations and to try and incorporate some of these in the released varieties so that they perform reasonably well in these locations too. In other words, rather than recognize a new mega-environment and a new breeding program, we may choose to use the information to increase the stability of our final release over all the environments represented on our biplot. Whatever our decisions, the biplot can alert us to opportunities in our breeding program.

Box 5. Breeding Decisions presented by cross over effects

Biplot analysis of Multi-environment trial data. See the presentation from which the tables below were taken from Yan (May, 2006). This presentation discusses many of the concepts in this manual and more! www.ggebiplot.com/Biplot_Analysis_of_MET_Data_IITA.pps

Are the cross-over patterns repeatable?

- If yes
  - The target environment can be divided into multiple mega-environments
  - GE can be exploited by selecting for each mega-environment
- If no
  - The target environment cannot be divided into multiple mega environments
  - GE cannot be exploited
  - GE must be avoided by testing across locations and years.
  - the reason for the GE can be investigated to assist these processes (my addition)

Classify your target environment into one of three categories

	With Cross-Over GE	No Cross over GE
Repeatable	(2) **Multiple Mega-Environments** Select for specifically adapted genotypes for each Mega-environment	(1) **Single Simple Mega-Environment** A single test environment in a single location, suffices to select a single best variety
Not repeatable	(3) **Single Complex Mega-Environment** Select for generally adapted genotypes across the whole region across multiple years.	

2.1.3 Looking at a biplot based on one season's results

Which genotypes and locations breeders continue to use is a complex decision depending upon, among other things, what traits we are breeding for, what environments we are breeding for and what information we already have about the genotypes we are testing. That said, breeders constantly make decisions about the next generation trials based, in part, on the results of the last, single season. Figure 2-1 can be viewed as just the summary of the results of the previous season's results. As such, the breeder is obviously going to want to continue with G7, as it "wins" in the majority of the environments and very likely the other winning genotypes will be included in subsequent trials also. If the number of genotypes examined in the next season's trial must be reduced, any of the non-winning genotypes might be considered for the cut, with the possible exception of G4 whose performance is close enough to the winning line to suggest

it might be a winner in later trials. Now look back at the data in Table 1-1 and its visualization in Figure 2-1 and ask yourself if you would have made the same decisions with the raw data.

A biplot like Figure 2-1, based on one season's data, can also be useful in diagnosing the source of GxE interaction. In this example, E3 stands out as very different from other environments and G8 is uniquely adapted to it. By looking at the characteristics of E3 and G8 we might get insights into mechanisms. For example, suppose E3 suffered a long dry-spell at the end of the growing season, making the season much shorter than at other places. We might then see what was special about G8 that allowed it to thrive in such conditions. Did it flower early? Did it have deeper roots? Such information is clearly valuable in advancing a breeding program. The information from a single G and E is only good for generating hypotheses, but that is important. It means we can design the next trial to deliberately test those hypotheses. For example, if we think the adaptation to a short season is to do with deep rooting, then our next trial would deliberately compare deep and normal rooted genotypes in short and normal season environments. If in the biplot we had seen a group of environments similar to E3, and a group of genotypes similar to G8, then we could see whether they shared the characteristics of short season or deep rooting, adding weight to evidence for the hypothesis.

3 Understanding the Biplot

3.1 The relationship between genotype and environments

Suppose you had some data in a table or matrix form with genotypes on the rows and environments listed on the columns. Let us take the simple example in Table 3-1. For this example we want to have some negative numbers so we will pretend the results are some type of score that can take negative values.

Table 3-1. Ficticious score data for 4 genotypes grown in 3 environments

	E1	E2	E3
G1	7	-6	18
G2	8	-6	12
G3	-10	7	-10
G4	17	-12	18

Now suppose you could break this data down so the genotypes are represented by X and Y coordinates and the environments are also represented by X and Y coordinates as shown in Table 3-2.

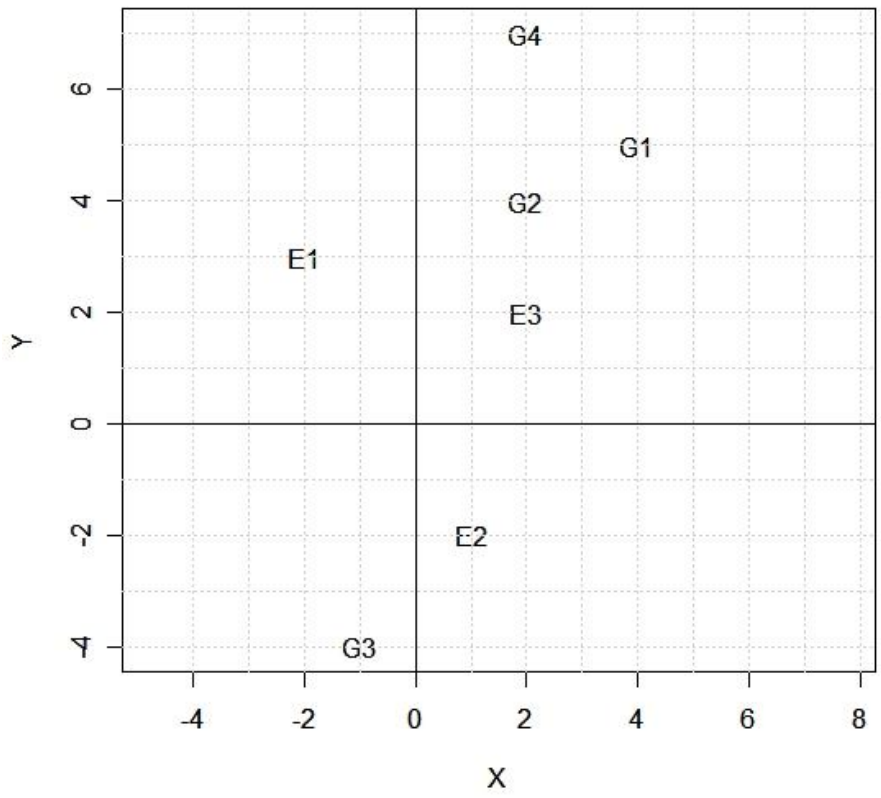
Table 3-2. Component Matrices for Table 3-1

	X	Y
G1	4	5
G2	2	4
G3	-1	-4
G4	2	7

	E1	E2	E3
X	-2	1	2
Y	3	-2	2

Now, as you might guess, it will be possible to plot both the genotypes and the environments on the same X-Y axes.

Figure 3-1. Plot of Coordinates in Table 3-2 and equation for inner product for G1*E1



Inner Product =

(x value of row * x value of column)+(y value of row * y value of column)

Value in Table 3-1 in cell G1xE1 =

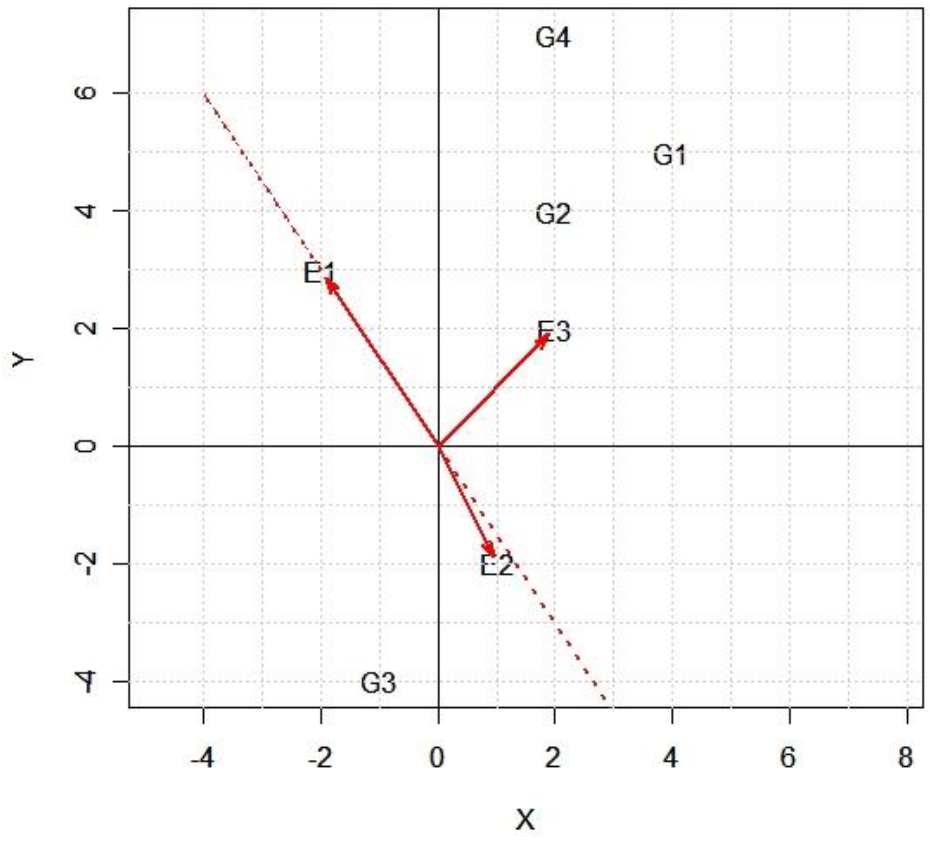
(x for G1 * x for E1) + (y for G1 * y for E1)

(4 * -2) + (5*3) = -8 + 15 = 7

You can see that G1 is plotted at x=4 and y=5 and that E1 is plotted at x=-2 and y=3. What is not so obvious is that something called the inner product will give us back the original score of the first variety, G1, at, E1, environment one. The inner product is calculated by multiplying the two x values together and added to the two y values multiplied together so that you have (4*-2) +(5*3)= -8+15=7. The value in Table 3-1 showing the score of G1 in E1 is 7. You can see that the other values will follow the same rules.

It is this property, and its corollary explained in the Box 6, that allows us to explore the relationship between genotypes and environments on the same graph. Suppose we draw arrows from the graph's origin to the points representing each of the three environments. We can also extend the lines indicated by the arrow with a dotted line in the positive and negative directions as we have done for E1 in Figure 3-2.

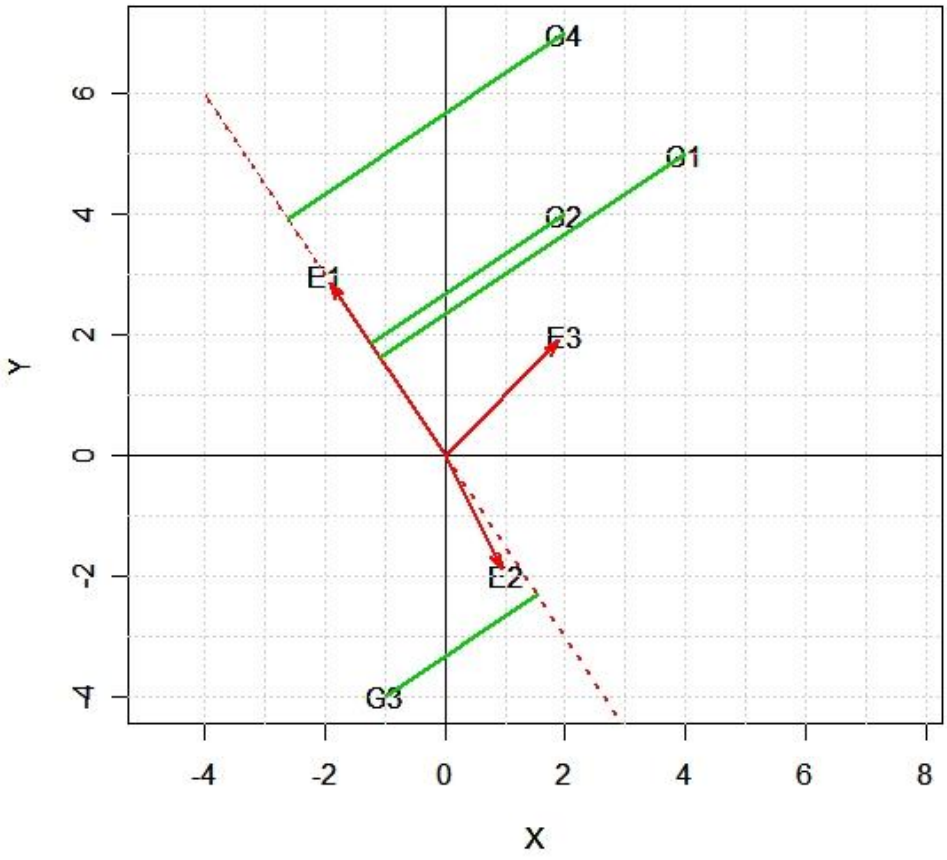
Figure 3-2. Biplot of data in Table 3-1 with arrow vectors marking environment points and the vector for E1 extended in both directions



On this biplot it is very easy to see the relationships between genotypes and environments. Suppose we concentrate on the line for environment 1 (E1). Now drop perpendicular lines from each of the points for the genotypes on to the line for environment 1(E1) as shown in Figure 3-3. These green lines are known as projections. It is easy to see that G4 is first along the positive direction of E1followed by G2, then G1 and then, taking a negative value, G3; which is exactly the ranking of the scores of these genotypes at E1. Look at the first column of Table 1-1. Even the scale of the relationship is correct as the green lines for G1 and G2 are close together when they hit the E1 line just as their scores would suggest. Pay close attention to the fact that the projection from G3 hits the E1 line on its extension in negative direction. The negative extension runs from the origin, (0,0), in the opposite direction from the arrow, and all projections

on this negative extension indicate negative scores in the original data. In fact, you could if you wish calculate the score values of each genotype at each environment in Table 3-1 as explained in Box 6.

Figure 3-3. Projections of genotype points onto the line of Environment 1



In Figure 3-3 you didn't really need the projection lines to see which genotypes were doing well in E2. This is often the case but sometimes it is not quite so obvious exactly where the projection line from the genotype will strike the environment line. Look at the projections of the genotypes onto the line indicated by environment 3 (E3) in Figure 3-4. You may need some practice "eye-balling" a perpendicular projection line but, it is still easy to see which genotypes did well in comparison to others.

Be careful not to get the idea of closeness to the environmental axis mixed up with the length of the projection on the axis. In Figure 3-4 the genotype G1 is closer that G4 to the E3 axis but both genotypes have the same projection, and thus the same score, in environment E3. Likewise G2 is closer to the E3 axis than G4, but its projection is shorter and therefore the score value of G2 in E3 is less than that of G4.

Look at some of the other relationships clearly displayed in these biplots. The genotypes G4, G1, G2 score highly in environments E1 and E3 although the relationship is not exactly the same as there is

crossover for G1 and G2. However, there is something entirely different happening in environment E2 with G3. Genotype G3 only has a positive score in E2, where all other genotypes have negative scores. The biplots clearly illustrate this big difference that might be lost in a large dataset.

Figure 3-4. Projections of the genotype points onto the line of Environment 2

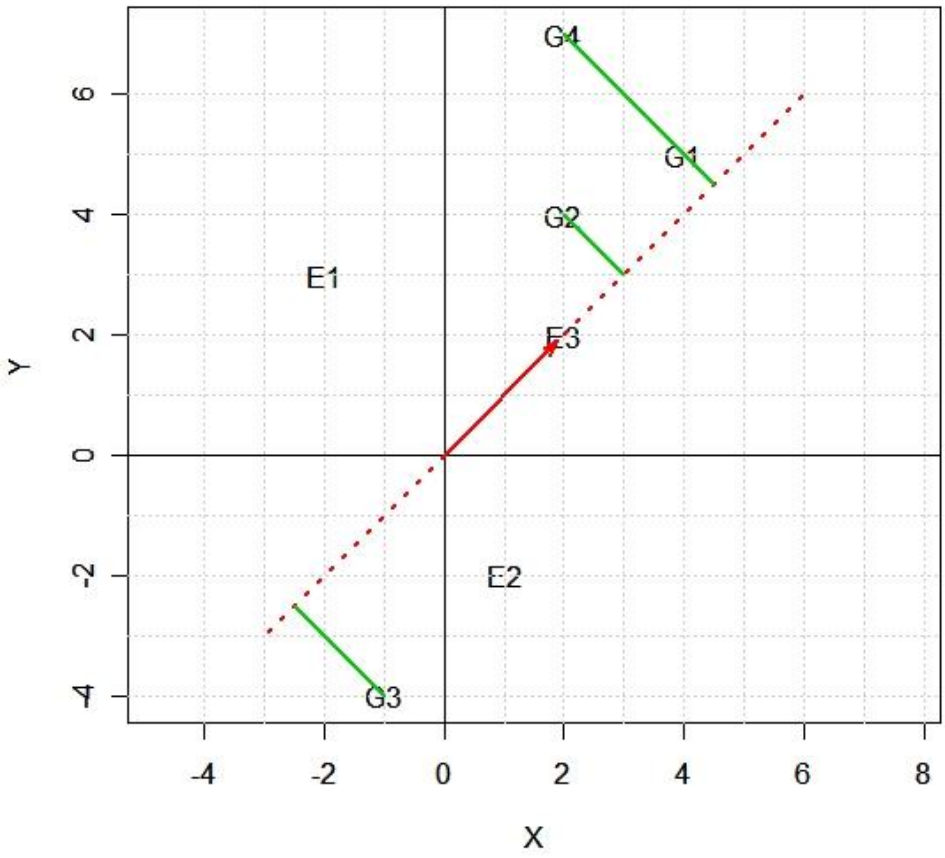
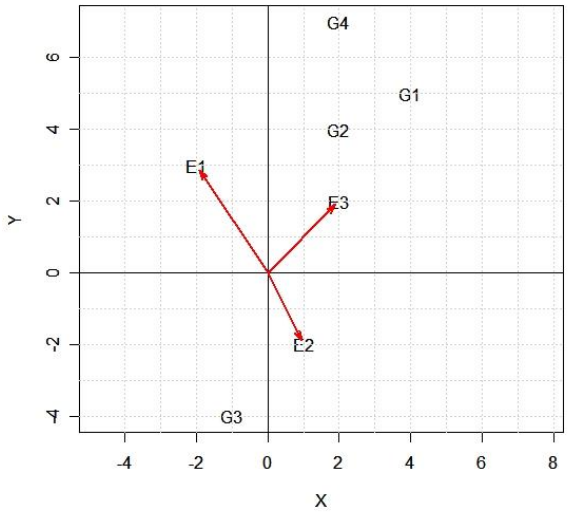
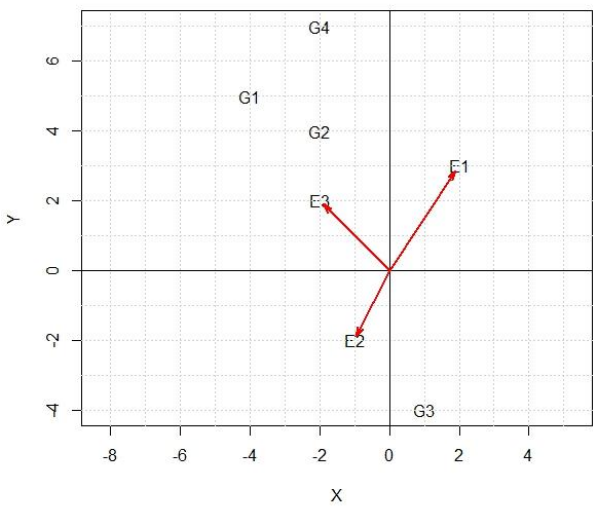


Figure 3-5 a. Biplot of data in Table 3-1 with arrow vectors marking environmental points	Figure 3-5 b. Biplot of data in Table 3-1 with arrow vectors marking environment points and signs of X axis reversed
	

Finally before we leave this section, it is important to make the point that the sign (negative or positive) of the values on the axes makes no difference to the interpretation of the biplots. Look at Figure 3-5 (a) and Figure 3-5 (b), and take some time to satisfy yourself that the relationship between the genotype points and the environmental points is the same, and that the projections of genotype points onto the environmental vectors (or vice versa) would be the same. What determines whether the original table value was negative or positive is whether the point projects onto the positive or negative direction of the environmental axis, as was explained above. Different software will give you different rotations so you may see the same biplot "flipped" either horizontally, vertically or both. Be aware that these are all the same biplots and will give you the same interpretations.

So now we have an understanding of the basic property of biplots that make them so useful for visualizing the relationship between two entities like genotypes and environments. However, in creating this example we have "cheated" a little because we have again returned to the situation where we can display the data perfectly in two dimensions. We know though, that we will not usually be able to do this. The next section discusses how to get the best two dimensional view of higher dimensional data.

Box 6. Explanation of inner product property

There is an actual mathematical relationship between the points on the biplot graph, the perpendicular projection of one point on the line formed by the other, and the actual value in the original data table.

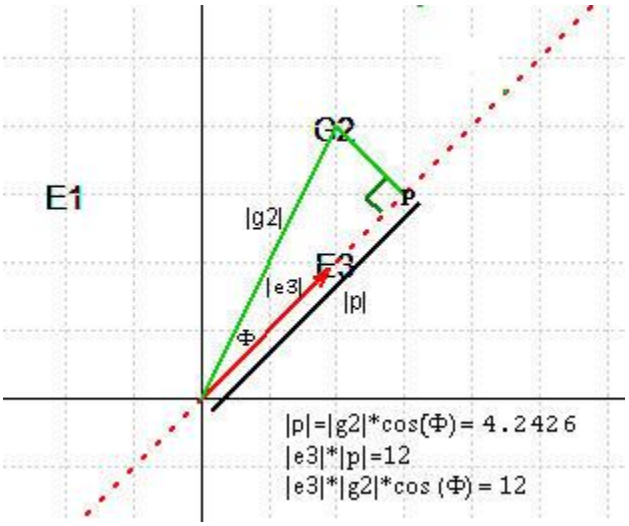
Look at the projection of G2 onto the line formed by environment 3 (E3) in Figure 3-6. The length of the projection from the origin to the point "P" can be calculated ascos(a)* the length of the line from the origin to G2 (|g2|). The length of G2 can just be computed from the Pythagorean theory as $|g2| = \sqrt{(2^2 + 4^2)} = \sqrt{20}$. In this case the angle θ is 18.43 degrees so that cos(18.43) *|g2|=4.2426.

Now if we multiply this length of projection by the length of the line from the origin to point E3, which is just $\sqrt{(2^2 + 2^2)}$, then we will get back the value 12 (after correcting for rounding error). This value 12 was the value in the original data for G2 at E3.

It is this relationship that supports our interpretation of the relationship between the rows (the genotypes) and the columns (the environments) as presented in the biplot

To learn more about the theory and application of biplots, to many disciplines—not just plant breeding, see the wonderful book written by Michael Greenacre:
http://www.multivariatestatistics.org/biplots.html

Figure 3-6. Example of Inner Product Calculation



3.2 Biplots as a visualization of a model

3.2.1 Reducing dimension for graphing

We said in the last section that we had cheated a little. This is because for the example given we had chosen a matrix (Table 3-1) that was exactly the result of the matrix multiplication of the two data tables (matrices) in Table 3-2. In order to graph G and E we needed to express each G and E point with two numbers, an X axis number and a Y axis number. For the simple example chosen in section 3.1, the result (inner product) where the two x values are multiplied together and added to the two y values multiplied together for G1 and E1 gave exactly the value in the main table (Table 3-1) for the intersection of G1 and E1-- (4*-2) +(5*3)= -8+15=7. We will rarely be so lucky when we are working with real plant breeding data. We will not be able to find X and Y values for the G and E points that give us back exactly the values in our initial table using this inner product equation.

Relax, this problem is not fatal. You have dealt with it before. When you make an ANOVA model for the average yield of a crop in a trial, you come up with something like:

Average Yield= grand mean + genotype effect + environment effect

Or perhaps, Average Yield=grand mean + treatment A effect + treatment B effect.

Usually, you are not concerned that these models do not give you back exactly the yield on any one test plot (replication), as long as you are satisfied that you have captured the main patterns in the data with your equation model and that the remaining random effects (**e**) are relatively insignificant.

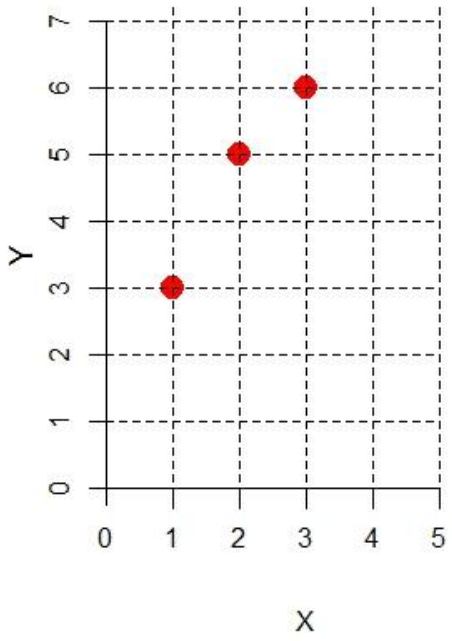
What you did in making a model of this kind, though you probably didn't think of it this way, was to reduce the dimensions of your full dataset, with a yield for each replication, to a model based on fewer dimensions. In creating biplots of the genotype by environment data we will need to do the same thing. We need to reduce the dimensions describing our data to two so that they can be graphed on a two dimensional X-Y graph. Remember in section 1.6 we said that if you graphed data in three dimensions you would have a cloud of points. Suppose you put this cloud of points in a cube and you could twist the cube in different directions to look at the cloud of points in different ways. Now think of putting the cube between the palms of your hands and squishing it flat. The points that were in a cloud now show up as points on a flat sheet (plane). You could get different pictures of the points on the flat sheet depending on the position of the cube between your palms before you fattened it. When you produced a flat picture of points from a cloud of points, in this way you reduced the dimensions of the image of the points from three to two dimensions. By squashing the cube you ended up with all the points that were in the cube being shown on the flat sheet, so your two dimensional picture is actually a projection of the points onto a two dimensional space.

Because we are confined when writing to two dimensions, we will have to discuss this further by looking at a projection of points in two dimensions onto one dimension. Take a look at the following example in Figure 3-7. We have the graphical representation of the data in Table 3-3. When we look at the graph in Figure 3-7 we can see the relationship between the points displayed in two dimensions with a feel for how far away from each other the points are in both the horizontal and vertical directions.

Table 3-3. Data points for reducing dimensions discussion

	X	Y
Point1	1	3
Point2	2	5
Point3	3	6

Figure 3-7. An X-Y plot of the data in Table 3-3



What if we decided for some reason to reduce the number of dimensions for viewing the data? So instead of a flat sheet in two dimensions we would use a line in one dimension. We have a lot of choices on how we might do this, but suppose we look at the points along the X axis. This is done in Figure 3-8(a), and we see the resulting line with the points in Figure 3-8(b)

Figure 3-8. A Projection of points in Table 3-3 onto the X axis

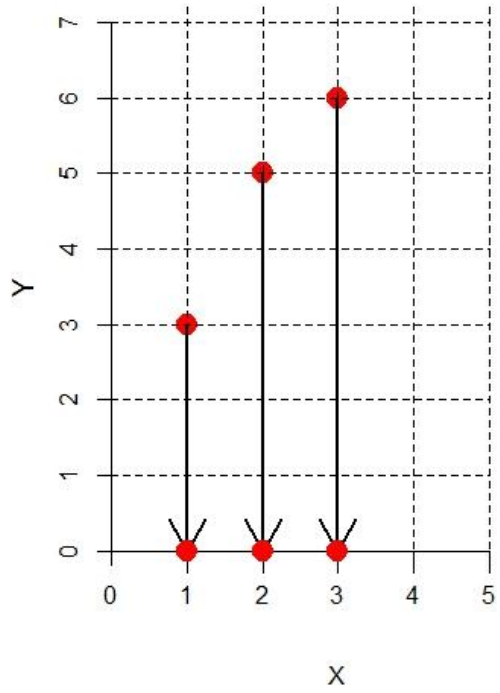
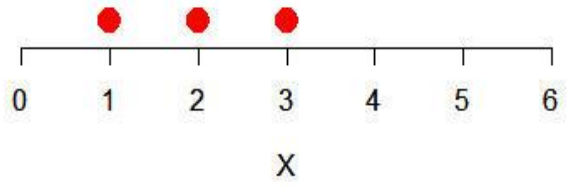


Figure 3-8b. One dimension line graph from projection in Figure 3-8a



This reduction is not really very satisfactory. We still see that the points 1,2,3 are in the correct ranking smallest to largest but we have lost the sense that point 2 is somehow closer to point 3. We have retained some information, but lost other information.

Now let us try using the Y axis for the reduction of the two dimensional graph to a line. The reduction is shown in Figure 3-9(a) and the resulting line graph is given in Figure 3-9(b). The reduced dimension line graph now gives you the feeling for point 2 being closer to point 3, but of course the information is not as complete as if you used the two dimensional X-Y graph. Again, points located by our arrows in the graphs illustrating the reduction of the two dimensional points onto one dimensional line are often called **projections**.

Figure 3-9a. Projections of the data in Table 3-3 onto the Y axis

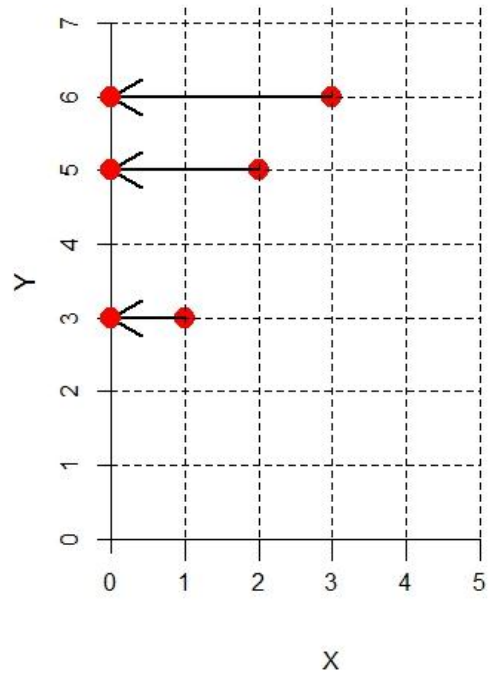
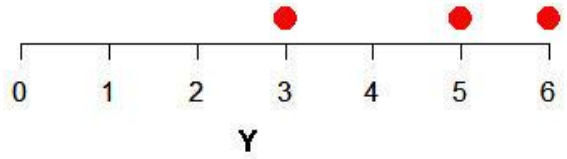


Figure 3-9b. One dimensional line graph resulting from projections in Figure 3-9 a



We are not limited to using the X and the Y axis for our line graph. We could use any line we wanted. Let us try running a line through the first and last points and using this as a new axis for our line graph. Here we will only need one projection of the second point onto the new axis as shown in Figure 3-10(a). The new line graph (Figure 3-9 (b)) is very similar to the one we got from projecting the points onto the Y axis, but this line graph shows us that all the points are a little further apart than we might have thought just looking at the line graphs from the X or Y axes.

Figure 3-10a. Projection of the points in Table 3-3 onto new axis

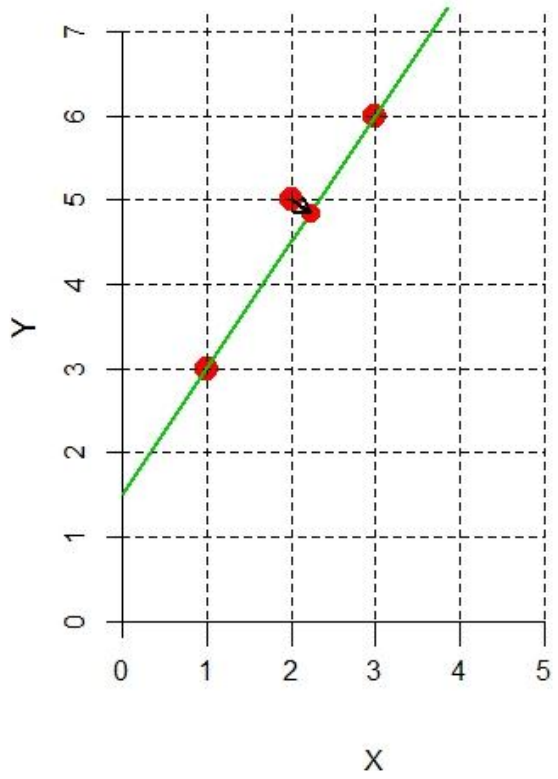
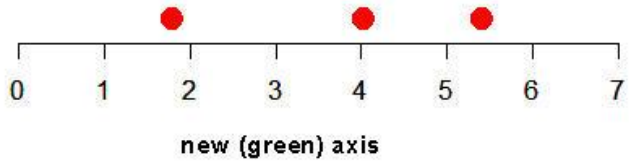


Figure 3-10b. One dimension plot of points in Table 3-3 on new axis



When we are using biplots we need to use a technique to reduce dimensions for data which usually have many more than two dimensions. Think of 30 genotypes grown in 17 locations. The computer software you will use for creating biplots (see Chapter 6) uses a technique called principal component analysis to decide on the best axes to use when projecting the data from higher dimensions onto new axes. When a biplot is produced for data that has more than two dimensions, the computer software will implement principal component analysis, using a technique called singular value decomposition (SVD), and choose the axes (and therefore the X and Y values) with which to graph the genotype and environment points on the biplot. The mathematical explanations you see elsewhere may seem complicated, but it is essentially doing something similar to the example we have just explored. So, again using our earlier analogy, you

can think of the data as a cloud of points. The software "inserts" two axes into this cloud and projects the points onto these axes. These projections give us a two dimensional (planar) view of the cloud of points. Of course we lose information in that there are some distances between points we cannot see in the two dimensional graph just as in our example there were some distances we couldn't see in the line graph. However, the software, through the use of mathematics, should give us the best view possible. That is, the two dimensional plot should retain as much of a sense for the distances between points as is possible. If you want to get a feel for how this dimension reduction is done then look at this animation on the web: http://www.youtube.com/watch?v=BfTMmoDFXyE

We have reduced dimensions and lost information. How much information have we lost? How do decide if the reduced dimension graphs are useful or misleading? We will look at that next.

3.2.2 How good is our visualization?

We can see that we lose some information, an understanding of how the points differ when we reduce dimensions. We can also see that some choices of axes are better than others for retaining the information from the higher dimension display. How can we put a measure on that loss of information?

Well, one measure of difference in a set of numbers is their variance. In the little data set Table 3-3 in the last section the variance of the X values is 1 while the variance, to two decimal places, for the Y values is 2.33 for a total of 3.33.

Now we can ask what proportion of the total variance is displayed in our reduced dimension graph. The points on first line plot Figure 3-8. A(b) have a variance of 1 so the proportion of the total variance of the data displayed by this first line graph is about 1/3.33 or 30%. That makes sense because we can see that we had lost a lot of information when we used the X axis for our reduced dimension graph. In the second case we used the Y axis so our second line graph, Figure 3-9a(b), shows 2.33/3.33 or 70% of the total variation. Finally, using our "new axis" for dimension reduction our third line graph, Figure 3-10a(b) shows 3.31/3.33 or about 99% of the total variation. Thus the new-axis, one-dimension, line graph shows almost all, but not quite all, of the variation of the two dimension plot.

So we can compare the variance in our reduced dimension graph to the total variance in our data and report that as a percentage. Exactly the same process can be used when reducing higher dimension data to two dimensions for biplot construction. The percentage of the total variance can be calculated and displayed to measure the percentage of the total information available in the data that has been captured by the biplot.

Using principal component analysis the first axis will always display the highest proportion of the total variance possible on one axis. The second axis will display the second highest proportion of the total variance after the first axis has already been chosen, the third axis will display the third most, and so forth. Therefore, by default, software to create biplots use the first and second axes, and they should also report the percentages of the total variance displayed by each axis and the percentage of variation displayed by the graph as a whole.

How large a percentage is enough? There is no exact answer to this question. A number of people have suggested ways to approximate an answer to this question but in this introductory text we are not going to review them. Instead you are encouraged to look at these references [Gauch,2007][Yan and Tinker, 2006]

3.3 A row view versus a column view

There is an important adjustment made to calculating the axis1 (pc1) and axis 2 (pc2) coordinates (point values) for the rows and columns of the data used to make the biplot. In plant breeding the rows are often the genotypes while the columns are often the environments or locations. As we saw in section 3.2.1 there is more than one way to reduce the dimensions in the data and thus there is more than one way of calculating the coordinate points for the biplot. In fact, there are many ways, but three are very common (if you want to understand this in more depth look at Box 7). The first way, known as the "row-metric preserving" method, makes sure that distances between the rows (usually genotypes) in the data used to make the biplot is accurately represented by the row points on the biplot. The second method, known as "column-metric preserving", ensures that the distances between the columns (usually environments) in the biplot data is accurately represented by the column points on the biplot. So if our focus is a comparison between genotypes then we will want a row view, or row-metric–preserving biplot, while if we want to concentrate on comparisons among the environments we will want to use a column view or "column-preserving-metric" biplot. A third commonly used method for calculating the biplot coordinates doesn't put emphasis on either the rows or columns and we will call this the "symmetric-view" biplot method. If our focus is on the relationship between genotypes and environments, as in section 2.1, then any of the three methods, or views, will produce a suitable biplot for this objective.

Obviously it is important to know which method has been used to calculate the coordinate points in any biplot you create or interpret. A good biplot software program should indicate the method used.

Box 7. Calculation of genotype and environment coordinates using singular value decomposition

We now understand that in order to make a biplot we need X and Y coordinates for the row variable (genotypes) and X and Y coordinates for the column variable (environments). How do we get these?

Well we start with a table of data like this

P_{11}	P_{12}	P_{13}
P_{21}	P_{22}	P_{23}
P_{31}	P_{32}	P_{33}
P_{41}	P_{42}	P_{43}

But of course it can have any number of rows (n) and any number of columns (m). The numbers represented by P_{11}, P_{12},…P_{43} depend on the model we choose (see section 4.1.1 below). They might be yields in tonnes/ha , they might be deviations from the mean yield of each environment, they might be something else. The software we will use to give us the X and Y coordinates will use a method called "singular value decomposition" (SVD), that will put us on the right path to getting the coordinates we want. SVD can be illustrated as

$$P_{ij} = \sum_{1}^{r} u_{ir}\lambda_r v_{jr}$$

So our value for P_{ij} has been broken into three vales for each of r dimensions. The program gives us the actual values of the u_{ir}, λ_r, and v_{jr}. So to get the value of P_{ij} back again we can just multiply and add up the dimensions

$P_{ij=}(u_{i1}\lambda_1 vj_1) + (u_{i2}\lambda_2 vj_2) + (u_{i3}\lambda_3 v_{j3}) \ldots\ldots (u_{ir}\lambda_r vj_r)$ for r dimensions.

But as we have discussed we have difficulty graphing and interpreting more than two dimensions so we could just take the first two dimensions. Let's say we are working with the value P_{32}

$$\widetilde{P_{32}} = (u_{31}\lambda_1 v_{21}) + (u_{32}\lambda_2 v_{22})$$

where the squiggle over the P_{32} means an approximation of the value of P_{32}

Now how does this get us X and Y values for the genotype 3 and environment 2 ?
This leads to the different "views". It depends what we do with the λ term.

If we use the column metric preserving view, or environmental view we multiply the λ term with the environments so the pairs of coordinates are

Genotype 3 $X=u_{31}$, $Y=u_{32}$

Environment 2 $X=\lambda_1 v_{21}$, $Y=\lambda_2 v_{22}$
If we use the row metric preserving view, or genotype view we multiply the λ term with the

genotypes
Genotype 3 $X= \lambda_1 u_{31}$, $Y=\lambda_2 u_{32}$
Environment 2 $X=v_{21}$, $Y=v_{22}$

If we want a symmetric view then we multiply both genotype and environment terms by the square root of λ.
Genotype 3 $X=\sqrt{(\lambda_1)}\, u_{31}$, $Y=\sqrt{(\lambda_2)} u_{32}$
Environment 2 $X=\sqrt{(\lambda_1)} v_{21}$, $Y=\sqrt{(\lambda_2)} v_{22}$

The beauty of SVD is that it orders the dimensions according to the percentage of total variation of the original data explained. So the first dimension is the dimension that explains, on its own, the largest part of the variation of the P_{ij} values. The second dimension is the dimension that explains the second highest percentage of the variation and so on. In fact the λ_r terms squared are the sums of squares attributable to that dimension so the percentage of the variation explained by the first dimension is $\frac{\lambda_1^2}{\sum_1^r \lambda_r^2}$. In other words the λ term squared over the sum of all the λ squared terms for all the dimensions.

However, if you think there is important information in dimension 3 there is nothing to stop you looking at biplots based on dimension 3 versus dimension 4 or other any other combination. For example:

$$\widetilde{P_{32}} = (u_{33}\lambda_3 v_{23}) + (u_{34}\lambda_4 v_{24})$$

4 Further use of biplots for plant breeding

4.1 Getting the essentials right

Now we understand that it is important to know:

1. What data you are using to make the biplot-- what model you are using.
2. How much of the total variation is represented by the biplot.
3. Which view the biplot is using

Points 2 and 3 have been well discussed in sections 3.2.2 and 3.3 respectively. We want to look now at the first point in more detail.

4.1.1 Which data are you using? / Which model are you using?

The biplot is the visualization of data in a table. Our examples, up to this point, have been tables or matrices, where the rows represent genotypes and the columns are environments or locations. The actual numbers in the table are based on some measurement made on the crop, like yield or days to silking. However, they are usually not the raw data or actual measurements themselves, but values derived from these measurements depending on the model you use.

In section 2.1.1 we saw the basic model:

Model 1

Which we originally wrote in section 2.1.1 as **Response =M+ G + E +GE + e**.

If we write the model as the average or expected response then we can write the model as:

Average Response =M+ G + E +GE

Now in section 1.4 we said we were not interested in the differences between the average yields at different environments so we centered the data on environment by subtracting the average yield at each environment (**M**+**E**) from each of the yield values in a column. You will often see this practice referred to as centering. Thus we created data defined by the model:

Model 2

Average Response – M – E = G +GE

This is the popular **GGE model** (Model 2) that we have been working with so far. The data in the table used in making the biplot represent the effect of genotype and genotype-environment interaction for each genotype by environment combination in the table. Compare Table 1-2 and Table 1-3 to see environment centering of data.

Another popular model is the AMMI model (Model 3) which removes the genotype effect so the data in the biplot table are estimates of the GE, genotype by environment interaction effect.

Model 3

AverageResponse – M – E – G = GE

You can see that we are not limited to these three models. We could leave different elements on the right hand side of the model leading to different numbers in our biplot table. We are not limited to addition and subtraction either. We might choose to multiply (or divide) one or more of our terms by a number. A common adjustment of this type is to divide the values in each column by the standard deviation of the environment represented by that column. You may see this model adjustment referred to as scaling. When we create biplots from plant traits data, traits being phenotypic characteristics like yield, days to anthesis, protein content, taste score, then each column contains information on a different trait. In this case Model 4 b is often used since the measures of the traits can be very different types of measures, like days to anthesis compared to disease resistance score.

Model 4

(Average Response –M – E)/d_j = (G + GE)/d_j(model 4 a)

(Average Response –M – C)/d_j = (G + GE)/d_j(model 4 b) where C is column effect

d_j is the standard deviation for each column in the data table

So, in all cases it is important to ask yourself— What is the model being used? This is the same as asking— What numbers is the biplot trying to represent? As we work through some examples below you will see how to relate the answer to this question to the interpretation of the biplot. Now check out Box 8 and review the essential points to understand when interpreting biplots

Box 8. Questions to ask when creating or interpreting a biplot

(Adapted from Yan, 2006 see www.ggebiplot.com/Biplot_Analysis_of_MET_Data_IITA.pps.)

1. On what model is the biplot based. This is the same as asking how were the numbers in the data table being used to construct the biplot calculated.
2. How well is the biplot representing the data calculated using the chosen model? That is, what proportion of the total variation in the data table is captured in the two dimensions shown in the biplot (see section 3.2.2)
3. Which view is being used? Are we using a row-preserving metric where the distance between the genotype points will represent the difference between the two genotype's performances? Maybe we are using a column-preserving metric view which will allow the best comparison between environments. In some cases a symmetric view may be used which does not promote comparisons between genotypes, or between environments, but does facilitate easy plotting for exploring the relationship between genotypes and environments.
4. Wekai Yan(2006) adds that one should make sure the graphing software used to make the biplot uses the same physical scale on the X and Y axes. That is, the distance on the paper, or screen, between say, 1 and 2, should be the same for both the X and Y axes. This is important for proper interpretation of the biplot.

4.2 Biplots focusing on GE interaction (AMMI biplots)

Plant breeders may want to focus their attention on the genotype by environment interaction in their trials. In this case they may want to use model 3 from the section above. Using our example data, we subtract the genotype and environment means from the data in Table 1-1 to obtain the data shown in Table 4-1 below.

Table 4-1. Model 3 data calculated from Table 1-1

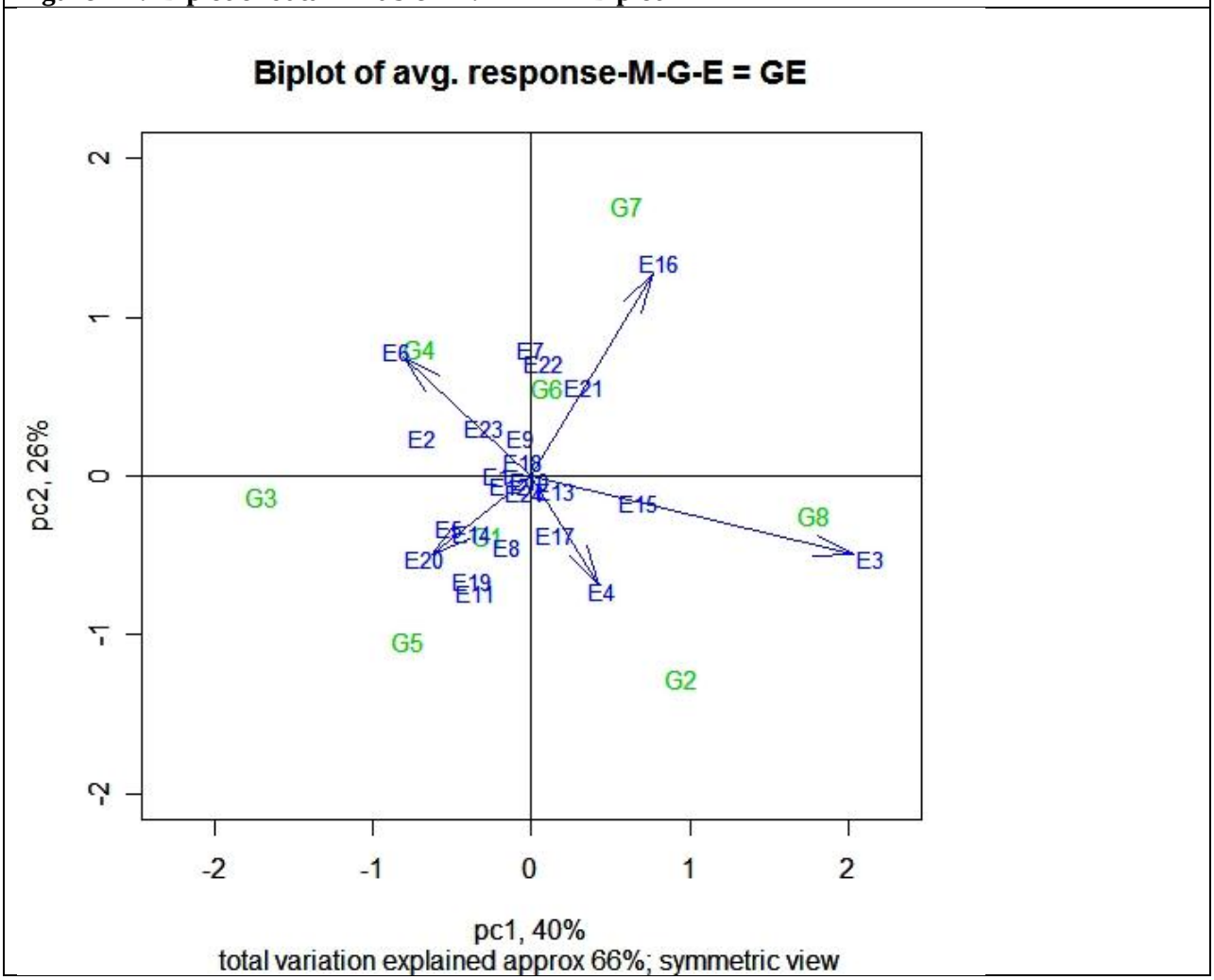
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
G1	0.19	-0.53	-0.95	0.19	-0.54	-0.49	0.30	0.29	0.31	0.61	0.42	0.50
G2	0.11	-1.60	3.08	0.49	-0.47	-1.39	-0.60	0.91	-0.28	-0.79	0.01	0.05
G3	-0.10	0.11	-3.75	-0.55	1.62	0.96	0.88	-0.28	0.87	-0.20	1.22	-0.36
G4	0.51	0.75	-1.71	-1.08	-0.86	2.31	-0.93	0.23	0.14	0.13	-0.87	-0.94
G5	0.28	2.08	-0.56	0.95	1.12	0.21	-1.76	0.48	-1.58	0.05	0.89	1.44
G6	-0.09	0.24	-0.71	-0.82	0.22	-0.93	0.86	0.23	0.75	0.27	-0.17	-0.95
G7	-0.05	0.27	0.97	-0.89	-0.69	1.14	1.87	-1.03	-0.36	-0.63	-1.67	1.38
G8	-0.85	-1.31	3.63	1.71	-0.42	-1.82	-0.62	-0.83	0.15	0.56	0.17	-1.11

	E13	E14	E15	E16	E17	E18	E19	E20	E21	E22	E23	E24
G1	0.08	0.22	-0.17	-0.89	0.77	-0.81	0.82	0.33	-0.09	0.16	-0.49	-0.23
G2	0.76	0.34	0.62	-1.06	0.61	-0.26	0.91	0.39	-1.35	-0.44	-1.37	1.35
G3	-0.41	-0.06	-1.40	-2.13	0.25	0.74	1.03	1.54	0.08	0.29	0.28	-0.62
G4	0.50	0.59	0.10	0.73	-0.69	0.03	-0.63	0.34	-0.76	0.80	0.99	0.33
G5	-0.42	0.93	-0.62	-1.52	-0.52	-0.22	0.37	0.52	-0.48	-2.11	0.59	-0.12
G6	-0.39	-0.02	0.72	1.34	-0.27	0.21	-0.57	-0.78	-0.34	0.52	-0.25	0.92
G7	-0.17	-0.99	-0.77	2.62	-0.66	-0.11	-1.09	-1.32	1.58	0.75	0.10	-0.24
G8	0.05	-1.01	1.53	0.90	0.50	0.42	-0.83	-1.01	1.37	0.03	0.15	-1.39
G8	0.05	-1.01	1.53	0.90	0.50	0.42	-0.83	-1.01	1.37	0.03	0.15	-1.39

In this example we will pretend that we are mostly interested in which genotypes display positive or negative interaction in which environments. Since the interest is in the relationship between genotypes and environments we could use any of the available biplot views but since the symmetric view makes it easy to see both genotype and environments points on the same plot we will use the symmetric view.

The biplot using these specifications is shown in Figure 4-1. We can see that the information needed to review our important points is given. We know that the biplot is based on the genotype by environment interaction model term of our model 3. Approximately 66% of the total variation in Table 1-1 is captured by the first two axes and the symmetric view is used. Arrows to selected environment points have also been added.

Figure 4-1. Biplot of data in Table 4-1: AMMI2 Biplot



This type of biplot in **Figure 4-1** is called an **AMMI2 biplot** in the plant breeding/biplot analysis literature [Gauch, 2007]. Remember that the data displayed in this biplot is the GE interaction only; we are looking at how much more or less the genotype performed in a given environment than was expected based on the main effects of that genotype and environment only. From our model statement we can see that the main effects of genotype and environment have been removed from the data used by this biplot. In fact, the data in Table 4-1 Table 4-1. Model 3 data calculated from Table **1-1** are the same as the residuals from an ANOVA on the data of Table 1-1, with on main effects of genotype and environment in the model. Thus, there is no measure here of average yield, and a large positive interaction does not mean that this genotype is the best performer in that environment. Nonetheless, it may be very useful to identify

positive and negative interactions when trying to understand which genotypic and phenotypic qualities adapt a variety for a particular environment (see section 2.1.3)

Knowing what you do from section 3, you can now interpret this plot. Be aware that you might have to mentally extend arrows and to keep in mind which is the positive direction of the arrows. Remember too that it is not how close a genotype point is to an environment line that is important, but the length of the projection of the genotype point onto the environment line. Looking at E16 it is clear that G7 has a large positive GE interaction with this environment. The genotype G6 also appears to have a positive interaction, while the biplot predicts that genotypes G5, and G1 will have negative GE values in environment E16 because their projections would be onto the negative extension of the E16 arrow. Does this agree with the data in Table 4-1? It appears that it does but, remember, it is possible that the biplot and the data table may disagree because the biplot is only displaying about two-thirds of the variation in Table 4-1. Some relationships between genotypes and environments may be missing from these first two dimensions.

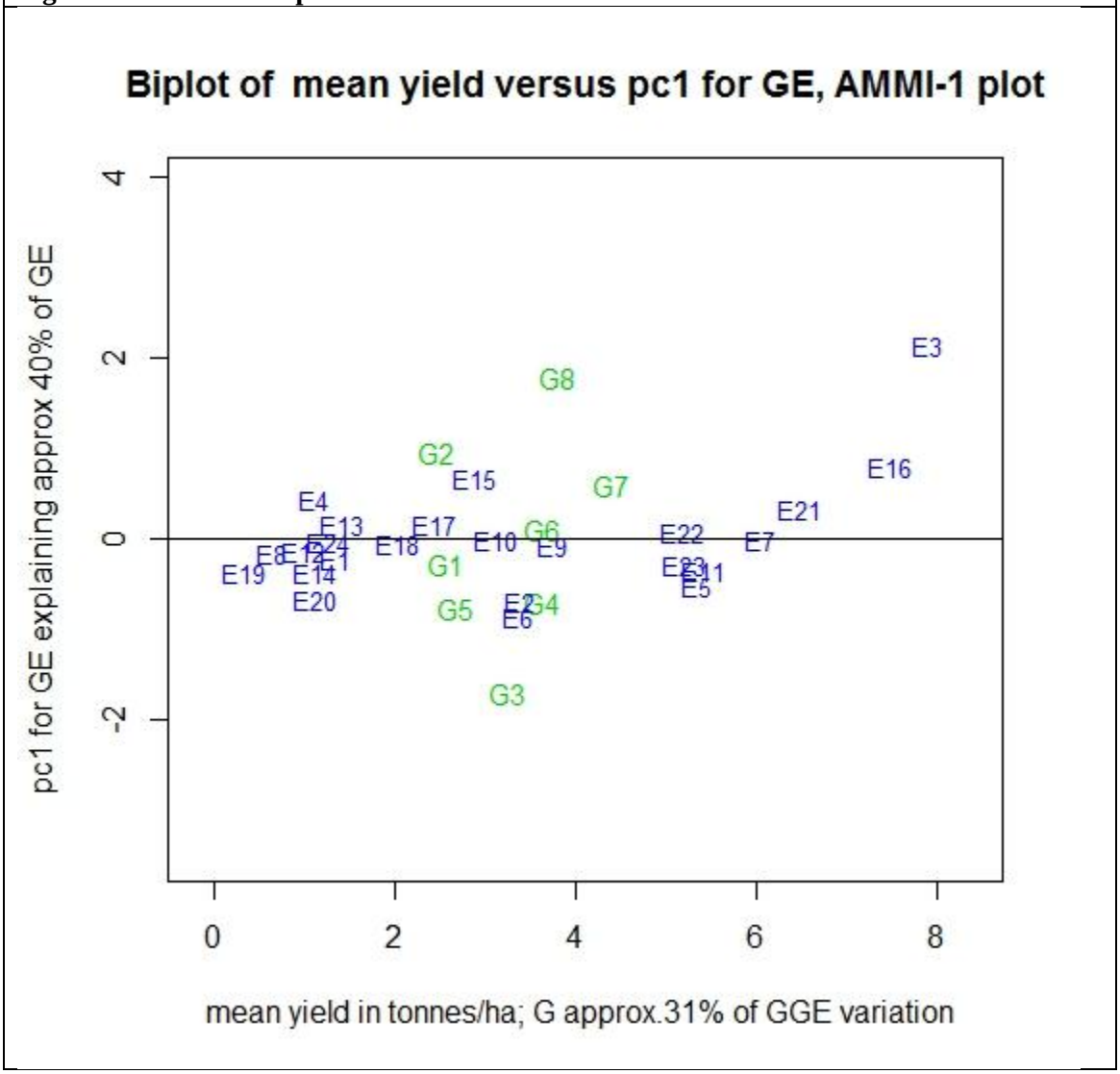
The AMMI2 plot show in Figure 4-1 is less commonly used than the AMMI1 biplot shown in Figure 4-2, but we began with the AMMI2 as it was more like the biplots we had already explored. The AMMI1 biplot allows the mean response (e.g. yield) of each genotype and environment to be displayed along with the first dimension measure of the GE interaction. In this way the AMMI1 biplot gives a direct measure of the yield potential and the stability (low GE interaction) of the genotypes being examined.

On the X-axis of the AMMI1 plot the values of the yield, or other response, are plotted and the values from the first dimension (pc1) values from a biplot analysis of GE values (Model 3) are plotted on the Y-axis. In other words, the values for the genotypes and environments on the X axis of the AMMI2 plot in Figure 4-1are now on the Y axis of the AMMI1 plot in Figure 4-2.

It is easy to see in the AMMI1 plot in Figure 4-2 that environment E3 has the highest mean yield while G7 is the genotype with the highest mean yield. The vertical, Y axis, is showing the best one dimension measure of the GE effect for each genotype. Thus genotypes close to the X axis have a small GE effect while those far from the X axis in either the positive or negative directions have a large GE effect as measured by the first pc1 axis of the biplot analysis. So we can say that this biplot estimates that those genotypes with points close to the X axis are more stable than those further away. Accordingly, G8 and G3 appear more unstable than the other genotypes. Checking back with our AMMI2 biplot of Figure 4-1we can see that G8 had an unexpectedly high yield in E3 while G3 had an unexpectedly low yield in this same environment and a higher than expected yield in environments E6 and E20 among others.

The AMMI1 plot emphasizes the two main criteria of interest within a mega-environment: mean performance and stability. If we regard all the 24 environments in our data set as locations within the same mega-environment then we need to select simultaneously for both attributes. High performance without stability is of questionable value since the performance of released varieties based on this selection may be erratic. For this reason if we regard the pattern seen in Figure 4-2 as repeatable over a number of seasons then we may decide to base our development of a new variety on G6 as it is more stable even though it is lower yielding, on average, than genotypes G7 and G8. However, selection for stability alone is completely useless because all we will be ensuring is that we develop a variety that is consistently bad. The selection of G1, for instance, just because it is relatively stable would be an example of such a nonsensical decision.

Figure 4-2. AMMI1 Biplot for data in Table 4-1



Note that AMMI1 biplot is created differently than the other biplots seen so far. The inner product property no longer holds, and AMMI1 cannot be used to relate environments and genotypes. The AMMI2 plot can be used for the exploration of this relationship for GE. Alternatively, any view of a biplot using model 2 (GGE biplot) can be used to explore the relationship between genotypes and environments for G and GE combined. The ANOVA table of the analysis of performance by genotype and environment can be used to calculate the mean effects of genotype, as a percentage of GGE if desired.

4.3 Assessment of Trial Sites

A column preserving metric (environment) view can be used to group trial sites/locations into mega environments and to make decisions on which locations to continue to use in a breeding program. This type of biplot, as explained in section 3.3, can be used to understand which locations are similar or different with respect to the model response (e.g. G + GE). Figure 4-3 shows such a biplot for the data

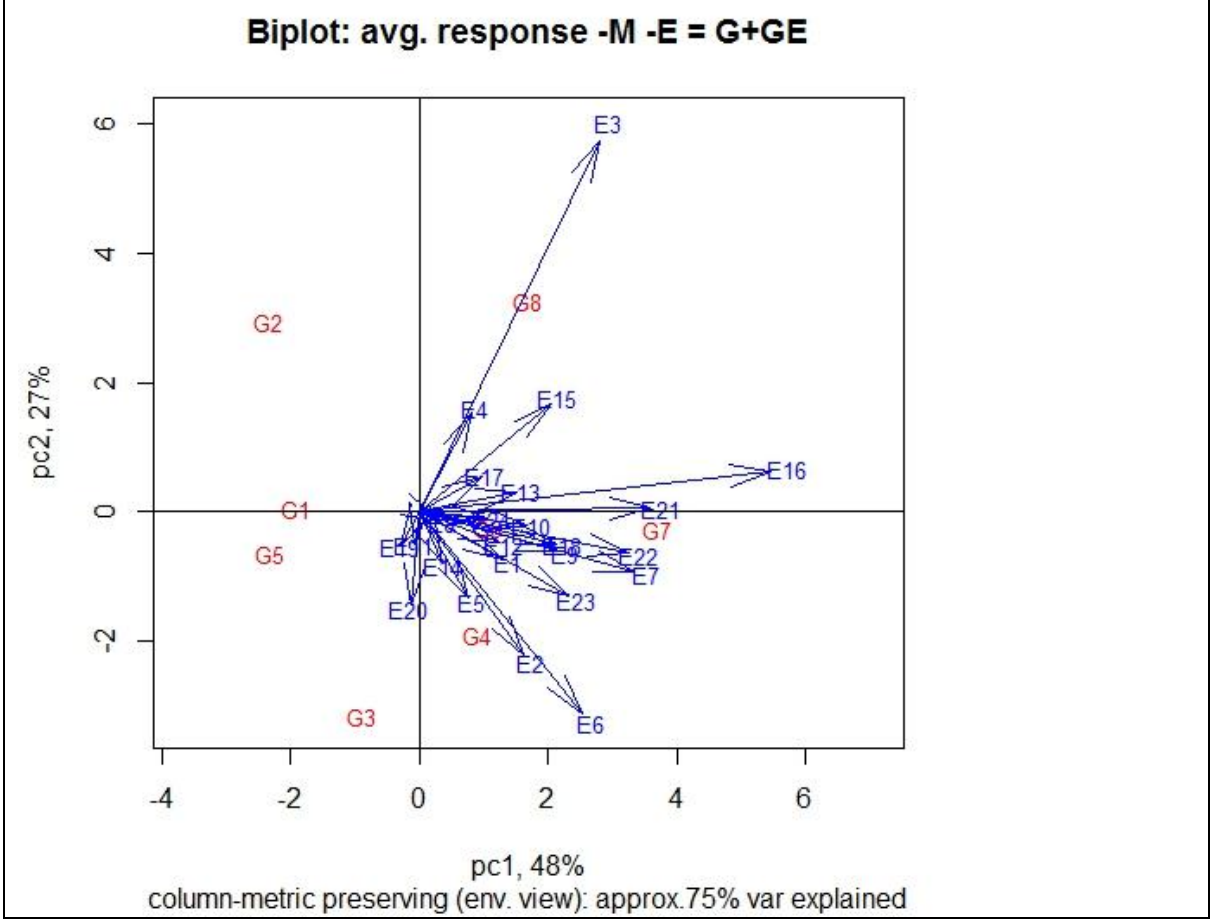
from Table 1-1. There is a small adjustment made to the scaling of the environment coordinates which will make no difference to our discussion. If you are interested, read Box 9 on environmental view adjustment.

Box 9. Adjustments to Column and Row metric preserving biplots

With the environmental view the scale of the values for the genotype and environment points are different. If you plot both points on the same graph the environmental points will be well spread out but the genotype points will tend to be scrunched together near the origin. The same will be true in the reverse for the genotype view. For this reason, the point values for the genotypes have been multiplied by an arbitrary number, in this case about 5.8 in Figure 4-3 which is the range of the environment pc1 values divided by the range of the pc1 values for the genotypes. This division (or multiplication) of the coordinates by any arbitrary number does not change the relationship display or interpretation of the biplot. It is the same thing as multiplying all the values in the data table by an arbitrary number which you might do if you changed the units in which the yield was measured (Smith, 2011). Be aware that many programs producing biplots for genotype and environment views will do the same, and they don't always tell you that they have done so. This is not a problem except if you try and reproduce the plot yourself with direct calculations, and then you will notice that your plots and the automatic software plots differ.

If we use an environment centered model (like Model 2 or Model 3) together with the column-metric preserving-environment view (and we avoid using Model 4 with d=standard deviation or error), the cosine of the angle between two environment vectors (arrows) is proportional to the correlation between those two environments. This means that if the two arrows form an angle of less than 90 degrees, the environments are positively correlated (with respect to the response), and those that form an angle of greater than 90 degrees are negatively correlated. When we say two environments are positively correlated we are saying that the same genotypes tend to do well or badly in the two environments. The distance between two environment points in the environment-view biplot is related to the covariance of the genotype performance in the two environments. Two environment points close together have genotype response patterns that are more similar in sign and size than points further apart. So generally we can view environments that are grouped together on the biplot as having similar genotype performances.

Figure 4-3. Biplot of environment centered data from Table 1-1



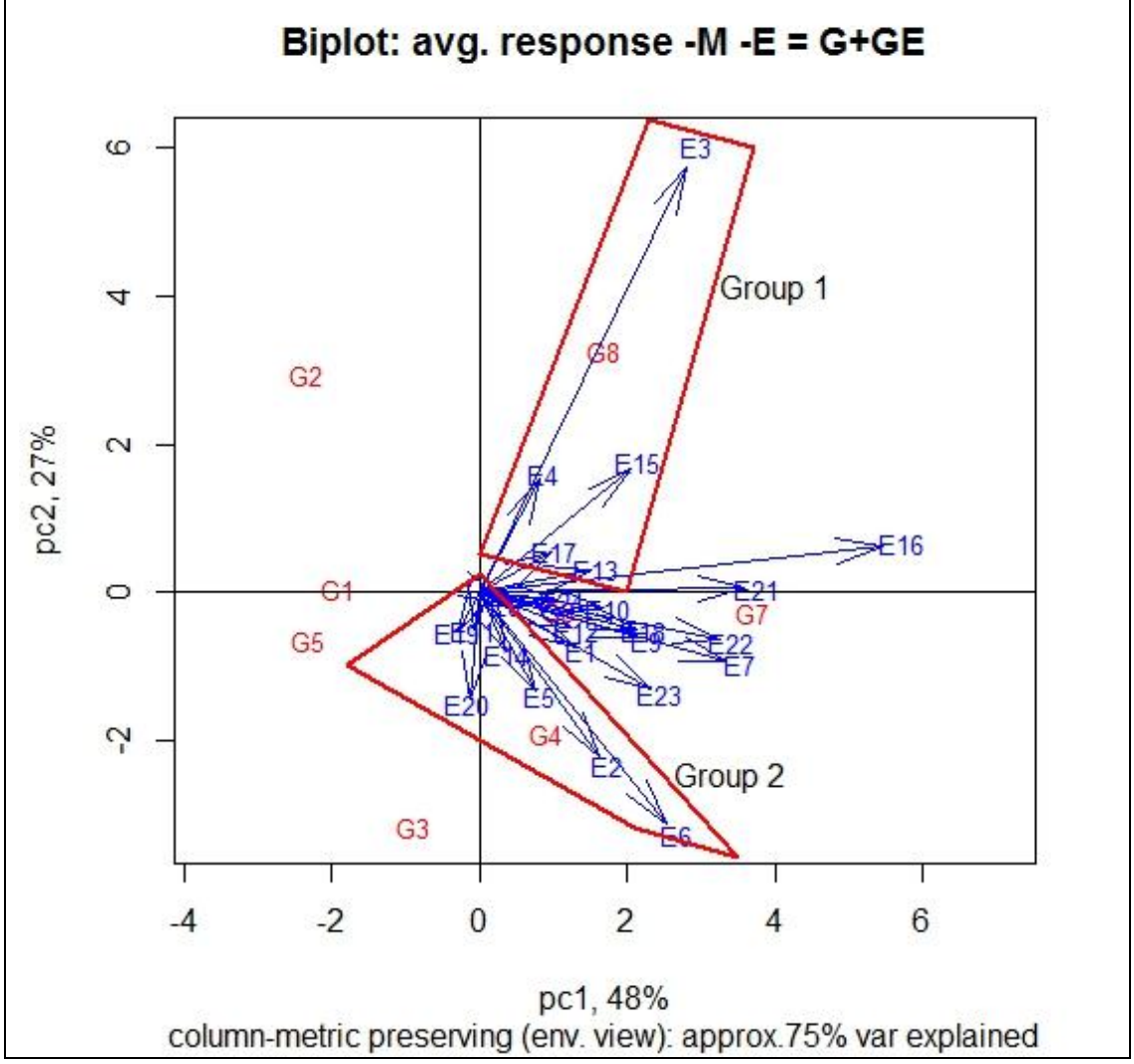
This allows us to partition the environments into groups that are similar. See the groups formed in Figure 4-4.We judge the environments in the upper group one to be different from those in the lower group two, as the angle between vectors of environments in the two groups appears to be larger than 90 degrees. The remaining environments may be assigned to either of the identified groups. It would certainly be of interest to know what characteristics the grouped environments have in common that leads to different genotype performance in the different groups. It might suggest that we should be considering two or more mega-environments (see section2.1.2), or it may just caution us to include representative locations from the different groups in further trials.

Another important characteristic of a trial location is its ability to discriminate between genotypes. That is, we want to see appreciable differences between the yields of the genotypes grown at that location and not have all the genotypes giving more or less the same yield. In other words, we want variability in the yield of the genotypes at a trial location. In the environment-view biplot the variance of the genotype responses is proportional to the length of its vector. A longer vector indicates a location in which there is a larger range of genotype performance.

Thus when we see a number of highly correlated locations like E6, E2, E5 and E14 it suggests that we are getting very similar rankings of genotypes at these locations and may save resources by including only

one of these locations in further trials. Given that the vector for G6 is the longest, we would select G6 as our representative location for this group (all other things, like cost and access, being equal) as it would be the most discriminating.

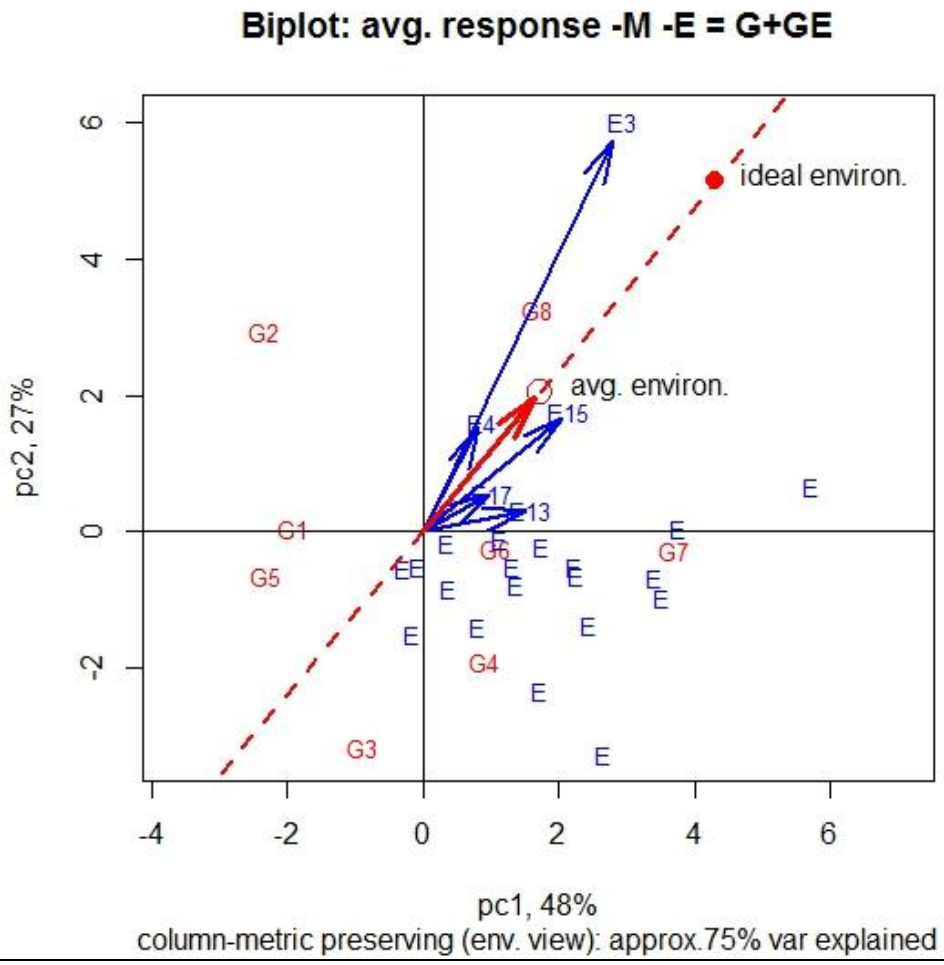
Figure 4-4. Grouping of some of the environments in Figure 4-3



Now suppose for the moment we had decided to treat the locations in group 1 as a mega-environment, and wanted to rate the usefulness of locations, E3, E4, E13, E15 and E17 as trial sites for that mega-environment. We can create an average environment axis (AEA) by drawing a vector from the biplot origin to the point with average coordinate values for the environments in that group 1. This has been done in Figure 4-5 and the open red circle marks the point of average environment coordinates for environments in group 1. You can see that we have extended the axis with dotted red lines. Now we can mark on this axis the point where the distance from the biplot to the point is equal to the length of the longest environment vector (here E3). This point has been marked with a closed red dot in Figure 4-5. This point represents an ideal environment in that it would be both representative of the mega-environment for which it is calculated and it would discriminate between the different genotypes grown

there. Now we are able to look for environment points that are close to this "ideal" environment point. In this case E3 would seem to be the best candidate location. Naturally, we don't have to break the environments up into groups. We can treat all the environments listed on the plot as being part of one mega-environment, and draw an average environmental axis for all the environments in our trial.

Figure 4-5. Biplot of environmentally centered data from Table 1-1 with average environmental axis for environments in group 1 and average environment point and ideal environment point for environments in group 1



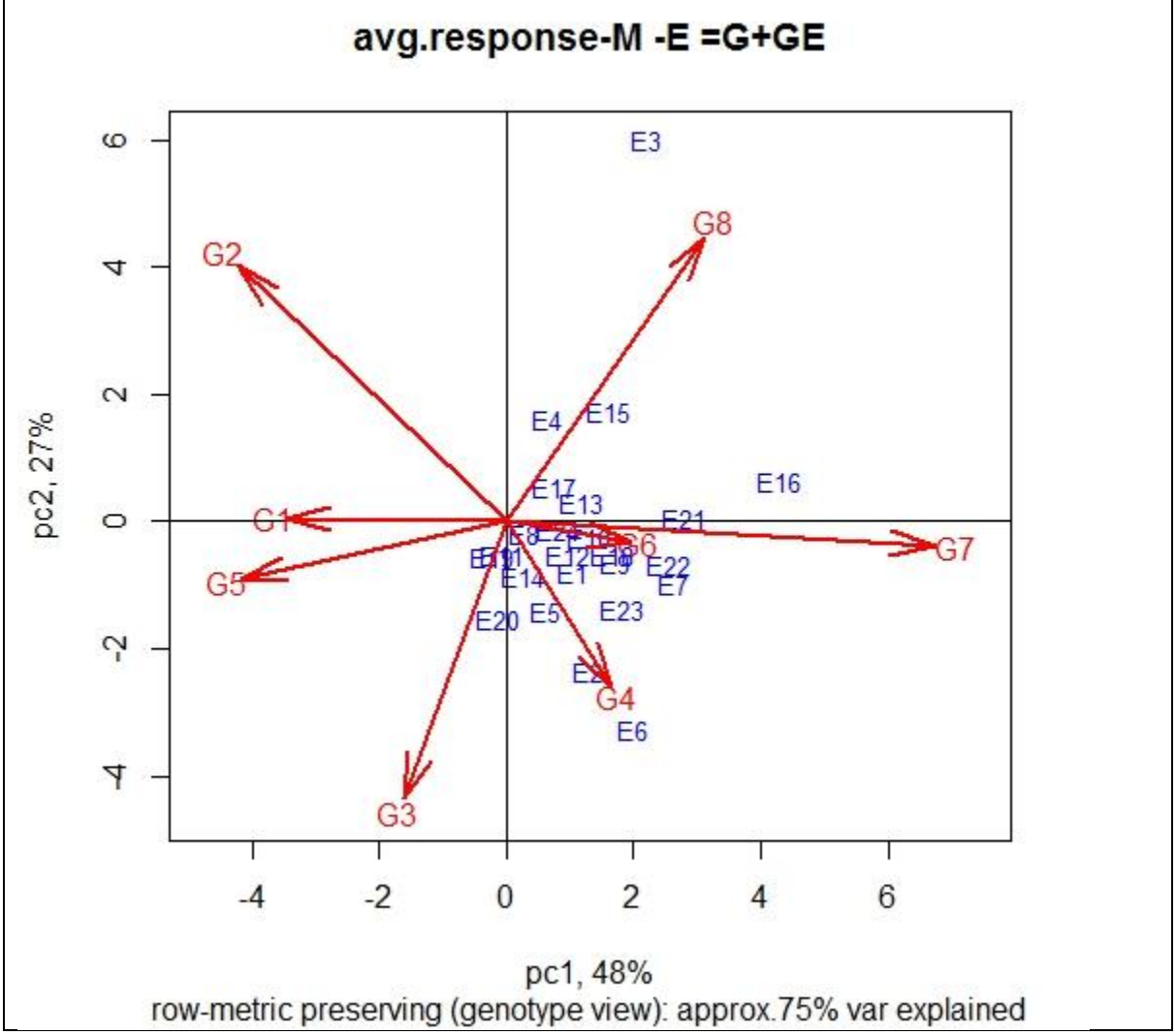
4.4 Comparison of genotypes

Let's switch our view now to the row-metric preserving, or genotype view. This is the view to use if you want to make comparisons between genotypes. Look at Figure 4-6. You might notice that there is a different scale on the X and Y axes compared to the plots above but remember, the numbers on the axes are arbitrary (see Box 9); what matters is the relationship between points that is scale independent.

We use the same rules for comparison of the genotype vectors as we did for the comparison of environment vectors in section 4.3. If the angle between two genotype vectors is less than 90 degrees, then these two genotypes are positively correlated, tending to do well, or badly, in the same environments.

If the angle between the vectors of two genotypes is greater than 90 degrees, then they tend to perform differently over the trial environments. If the angle between the two genotype vectors is 90 degrees then their performance is independent.

Figure 4-6. Biplot of environmentally centered data from Table 1-1 using a row metric preserving, genotype view.

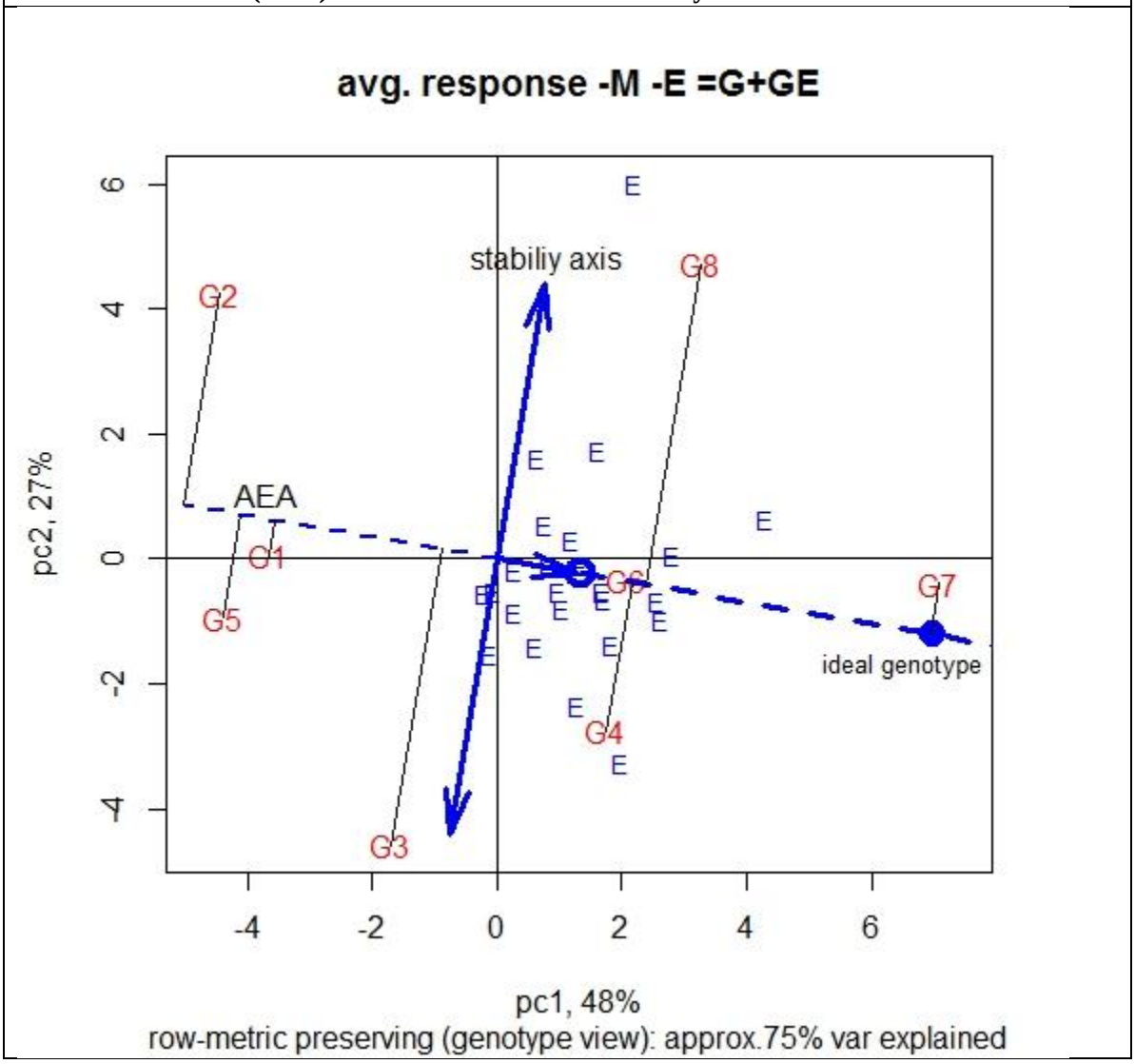


Again we will use the useful idea of an average environment. In this case we will pretend that we are working with one mega-environment only, and create an average environmental axis (AEA) for the entire data set. This A.E.A. is drawn in Figure 4-7, and the open blue circle is at the coordinates of the average environment which defines this AEA. Then we can project lines from our genotype points onto the AEA to visualize how they would rank in performance in the imaginary average environment. We are doing this in just the same way as we ranked performance of genotypes on a single environmental vector in section 3.1.

It turns out that an axis at right angles to the AEA passing through the plot origin can be thought of as a stability axis,(Yan and Kang, 2003). The stability axis is the axis drawn as double headed arrow axis in Figure 4-7. The further away from the AEA in the direction of either arrow a genotype point lies, the less stable it is, according to our model and biplot.

Weikai Yan (2003) proposed the idea of an ideal genotype. What do we want in a genotype? We want high performance combined with good stability. Thus an ideal genotype has high performance (e.g. yield) and high stability (e.g. yields well over a range of environments). A point is placed on the AEA at a distance from the origin equal to the vector of the longest genotype vector in the positive direction of the AEA; marked as a solid blue dot in Figure 4-7. This is our theoretical ideal genotype. It is obvious, according to this model and biplot, that G7 is both high yielding and stable. If for some reason G7 was not available then we would be faced with the choice between G4, G6 and G8 which all have similar mean yields. In this case we would be likely to rate G6 as the best genotype as it is more stable than the other two. We can see that G1 is also very stable but it is low yielding and so not of interest in a breeding program.

Figure 4-7. Biplot of environmentally centered data from Table 1-1 with average environmental axis (AEA) as blue dotted line and stability axis as a solid blue line



4.5 Biplots for traits

Saving the best until last we present examples of biplots for plant or environment traits. Biplots are great for visualizing patterns among very different characteristics measured on the same genotypes or environments; something that is quite difficult to do without the biplot tool. Often plant breeders will record a number of observations on their genotypes, such as plant height, days to flowering, scores given by participating farmers' groups, along with yield. At the same time characteristics of the environment such as precipitation, soil organic matter, disease challenge may also be recorded. Thus the breeder actually has three way data--genotype by environment by traits.

Biplots, of course, can only visualize a two-way, row by column data table; we need to choose which aspect of the three way table we want to look at. We could look at one characteristic in genotype by environment combinations, which is what we have been doing up to this point looking at yield for eight

genotypes grown in 24 environments. Alternatively, we could get the measure of a number of traits for each genotype averaged over all the environments and create a table with genotypes on the rows and traits making up the columns. Similarly we could create a data table with environments on the rows and averaged measures for a number of traits making up the columns. We could also choose to look at a number of traits for all the genotypes in just one environment or for all the environments for just one genotype; the choice is ours depending on our objective.

When creating a data table with different traits, for making up the columns it is common practice to use model 4 (b) (see section 4.1) in which the numbers in each column are divided by the standard deviation of the numbers in that column. This makes a lot of sense because otherwise we are trying to compare apples to oranges when we compare something like plant height, measured on a continuous scale from 30 to 150 cm, to something like farmer acceptance score measured on a discrete 1 to 5 scale. When we use model 4 (b) we center and scale each column so now the numbers in the table are measures of numbers of standard deviations from the mean. For example, a value of 1.5 would mean that this row and column combination had an original number that was 1 and a half standard deviations more than the mean value for that trait. When we do this, of course, we are making the assumption that the measures for the traits are at least roughly normally distributed so that the standard deviation is an acceptable indicator of the spread of the data. You will need to explore the distributions of the trait measures to see if this is a reasonable assumption.

Let us look at an example. Unfortunately there were only complete trait measurements for a subset of our example data that we started with in Table 1-1. We will examine the following traits for eight genotypes averaged over seven environments (Kosgei,2011).

- Yield: grain yield in tonnes per hectare
- Anthesis: days from planting until 50% of the tassels shed pollen
- Height: height (in cm) from the base of a plant to the point of the flag leaf, or insertion of the first tassel branch of the same plant
- Ear-height: height (in cm) from the base of a plant to the insertion of the top ear of the same plant
- Earpos: ratio of ear height to plant height
- EPP: number of ears per plant determined by computing a ratio of the number of ears and the number of plants
- Noplants: number of plants determined by counting the number of plants that survive to complete maturity and are ready for harvesting
- Earasp: this is a score with a scale of 1-5, where 1 is a score for clean, uniform and large cobs with the preferred texture whereas 5 is a score for small non uniform and diseased cobs with an undesirable texture.

The average values for the traits are given in Table 4-2. Earlier examination of the distribution of the trait measures showed the distributions to be roughly bell shaped, except “earasp” which had a positive skew. The biplot for this data, after centering and scaling, is shown in Figure 4-8. We are using a column – metric preserving view, or trait view, to allow us to make comparisons between traits. Not surprisingly plant height and numbers of ears per plant are closely associated with yield. Remember though, that these are averaged values, so what we are saying is those genotypes that have high values, on average, for height and numbers of cobs per plant also have high average yield. We do not have evidence that taller

plants gave higher yields in any one environment. This could be examined separately by looking at a genotype by trait biplot for a single environment.

Table 4-2. Eight traits for eight genotypes averaged over seven environments.

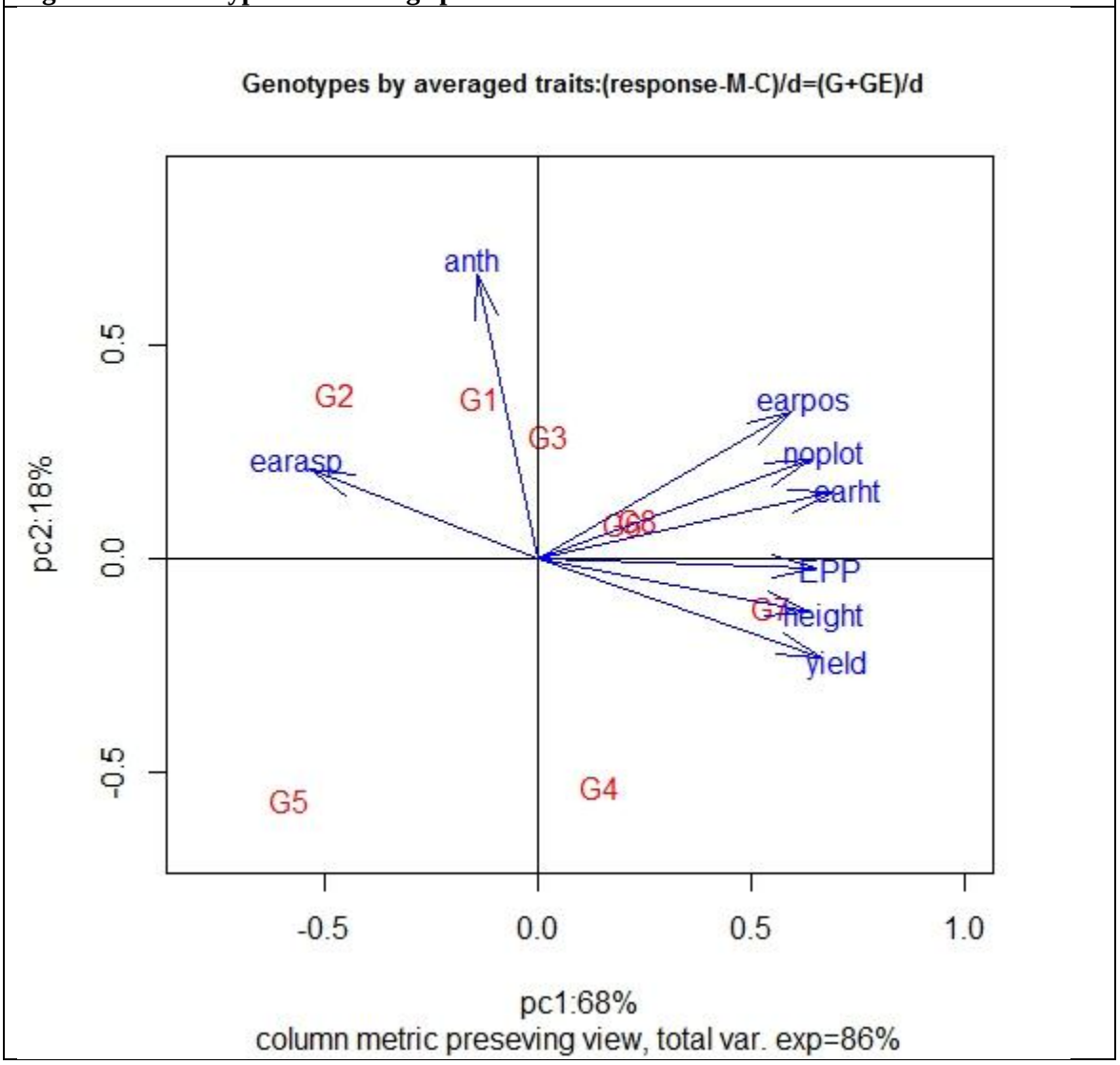
Geno	yield	anth	height	earth	earpos	EPP	noplot	earasp
G1	3.41	31.38	200.38	98.12	0.48	1.00	14.62	2.56
G2	2.82	31.12	188.12	90.62	0.47	0.92	14.12	2.75
G3	4.22	32.25	203.12	99.25	0.48	1.01	14.75	2.19
G4	4.78	26.38	209.38	97.88	0.47	1.06	15.25	2.31
G5	3.52	28.00	196.12	85.25	0.43	0.90	11.62	2.50
G6	4.89	30.12	217.25	106.00	0.48	1.03	15.25	2.50
G7	5.75	28.38	214.88	110.62	0.51	1.05	16.12	1.94
G8	4.67	29.50	217.88	105.50	0.49	1.01	16.25	2.44

The average days to anthesis trait (anth) is slightly negatively correlated with yield, while the average ear aspect score (earasp) is almost completely negatively correlated with yield. Recall that this score is an indicator of disease so the negative correlation is not surprising, and in this small dataset the rankings of average yield and ear aspect are almost exactly opposite. The genotypes G1 and G2 have higher values of "average days to anthesis" and ear aspect and negative values of average yield and all the traits positively associated with average yield. The genotype G5 is low yielding like G1 and G2 but does not have a high "average days to anthesis" score. In agreement with our earlier analyses G7 has highest average yield and the highest positive correlation with all the traits positively associated with average yield. The genotype G4 is interesting in it has a much lower "average days to anthesis" score than the other above-average yielding genotypes.

You must now be convinced of the usefulness of biplots for examining trait data. You can also see that what you need are a number of biplots looking at different aspects of the same data. Imagine the insights, and questions, that might be uncovered looking at genotype and location responses for individual traits by year. Then you could couple those analyses with biplots looking at the relationship between traits. Being comfortable with fairly fast construction of biplots may give you the opportunity to examine data that often remains unused.

But like any tool, biplots have to be used wisely. Please read the next section with care.

Figure 4-8. Genotypes and average plant traits.



5 Important Limitations to Consider

As useful a tool as biplots are in understanding multi environment data, they are not fool proof and should not be used as a standalone analysis tool. First of all, biplots contain no measure of error, no sense of how large the "**e"** is in our model 1 in section 4.1.1., compared to differences between means of genotypes in different environments. You can think of every entry in the data table we are plotting as having some uncertainty, or being a bit 'fuzzy'. That must translate into fuzziness in the biplots. So biplots are a descriptive tool and, as with all descriptive statistical tools, we have to ask ourselves are we looking at a repeatable pattern or a chance arrangement of the data. At present, the only way to answer this question is to go back to other complementary analyses like ANOVA and other linear analyses and get an estimate of the variance of "**e"**. Then using standard post-hoc tests for comparisons of means we could test, for example, if it is likely that G3 would consistently out yield G7 in E5, E14, E11, E19 and E20 as was suggested by the biplot in Figure 2-1. Research continues on methods to add measures of uncertainty to biplots. Hopefully, it will soon be possible to get biplots with visible confidence regions with standard software.

We know too that biplots are giving us a picture of only part of the variation in the data. It is possible that there are one or two significant genotype by environment interactions not captured in the first two dimensions used by the biplot. Of course, it is possible to create a biplot using any two dimensions you wish. You can create a biplot looking at the first and third dimension, the third and fourth dimensions, or fifth and sixth; but you must remember you are looking at successively smaller proportions of the total variation. The higher the dimensions used to create the biplot the more likely the display is one of random noise rather than repeatable patterns. You will be looking specifically for patterns missed in the first two dimensions. One way to check for this is to look at the residuals from the biplot (Gauch and Zoble, 1996). If you remember back to section 3.1 we saw that the values in the data table could be computed from the inner product of the values of the coordinates of the genotype and environment if the data could be perfectly rendered in two dimensions. Thus the reduced dimension biplot's prediction of the values in the data table can be calculated in the same way and subtracted from the actual values to give residual values. These residuals can then be examined, perhaps with the use of an index plot, for values that stand out indicating a genotype-environment combination that is not well described by the biplot using the first two dimensions. Then you can use a biplot with a third, or perhaps higher dimension, to look for the pattern that fits the data that was poorly described using a standard, first two dimension biplot. Now you can even look at three dimensional biplots (three dimensions at once) although they can be tricky to interpret. (see chapter 0.)

6 Moving Forward and Learning More.

Hopefully by this point you are itching to start creating and interpreting some biplots of your own. Although it is not difficult to create your own biplots with any package that will do singular value decomposition and create graphs--the plots in this manual were created directly in R (2011)-- it helps to start off using software specially created for the purpose. Luckily, some of the main developers of biplots for plant breeding have made software and learning materials freely available that will have you plotting your own in no time.

If Model 2, GGE biplots are your main interest then you will want to try out the freely available, demonstration version of the software package GGEbiplot at http://ggebiplot.com/home. This demonstration version is fully functional and very easy to use, but it is limited to showing just the first seven genotypes and first seven environments on the actual biplot. In fact, this software is so easy to use and the website so full of useful learning materials that any student interested in using biplots for plant breeding would be wise to use this site and demonstration software for learning about biplots even if they have to transfer to other packages for their research analysis. You can even try out three dimensional biplots with the demonstration version.

Less easy to use, with a focus on Model 3, GE analysis, the package MATMODEL (http://www.css.cornell.edu/staff/gauch/matmodel.html), though is still very useful and truly open source software with no inbuilt limitations of numbers of genotypes or environments. This package contains a number of model diagnostic options, and the accompanying word file is a wealth of information on AMMI analysis and the model fitting procedure in general.

Both specialized software packages, GGEBiplot and MATMODEL, have examples files and output, and are a good way to begin working with biplots. Both packages also have functions for accompanying statistical analyses.

If you have used the R statistical software before you should try the GGEBiplotGUI package [Bernal, 2011]. Once loaded this package creates a "point-and-click" interface for creating most of the biplots in this manual. You need to have your data in a table in M.S. Excel, version 2003 or earlier, with column names in the first row. Your first column should name the genotypes and the rest of the columns should contain the data for each column, for example, yield data for each genotype in each environment. Then you can just follow the instruction in the dialogue boxes that open when you run "GGEBipot()[WU1]". The really outstanding feature of this package is the ability to look virtually at three dimensions at once. You have to try it just for the fun of it even if you, like me, find them hard to interpret.

Most of the large, general use statistical packages have facilities for creating biplots. Unfortunately, it is outside the resources of the authors of this manual to systematically review their suitability for plant breeding applications, though we will mention a couple of points that may be helpful. CIMMYT (International Maize and Wheat Centre) makes available SAS programs with examples for AMMI and GGE analyses at http://www.cimmyt.org/en/programs-and-units/units-a-labs/crop-research-informatics-laboratory/disciplinary-groups-units/biometrics-and-statistics-unit. The statistical package Genstat ver. 14

is easy to use because it has “point-and-click” functions for both AMMI and GGE analyses and biplots, with the additions mentioned in this manual like drawing the average environmental axis.

Whichever software you use for biplot construction, even if you code your own, you need to ask yourself the questions listed in Box 8 in section 4.1. It helps to use software that answers these questions explicitly.

For a more in depth understanding of biplots for plant breeding research with a slightly more mathematical presentation, we suggest the review paper by Yan and Tinker (2006). For a very accessible explanation of the theory and use of biplots in general, we suggest the book by Greenacre (2010). The book by Gower et al (2011) shows how biplots can be used with many other data types.

7 References

Bernal, E. F. 2011, GGEBiplotGUI R package. http://cran.r-project.org/web/packages/GGEBiplotGUI/index.html

Gauch, H.G. 2007. MATMODEL Version 3.0: Open source software for AMMI and related analyses. Crop and Soil Sciences, Cornell University, Ithaca, NY 14853. accessed at http://www.css.cornell.edu/staff/gauch/matmodel.html

Gauch, H. G. and Zobel. R. W. 1996. AMMI analysis of yield trials. pp85-122, in in M.S. Kang and H.G. Gauch eds. *Genotype-by-Environment Interaction*, CRC Press. Accessed at http://www.crcnetbase.com/doi/abs/10.1201/9781420049374.ch4

Gower, J., Lubbe, S., and le Roux, N. 2011 Understanding biplots. Wiley.

Greenacre, M., 2010, *Biplots in Practice*. Fundación BBVA. pp237. accessed at http://www.multivariatestatistics.org/biplots.html

Kosgei, T., 2011, Combing Abilities, Heterotic Grouping and Stability Analysis of New Maize Hybrids for the Mid Altitude Areas of Kenya. MSc dissertation. University of Nairobi, Nairobi, Kenya

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Smith, Emily, 2011, Graphical Analysis of the Genotype Environment Interaction. MSc dissertation, University of Reading. Reading, U.K.

VSN International ,2011, GenStatfor Windows 14th Edition. VSN International, Hemel Hempstead, UK. Web page: GenStat.co.uk

Yan,W., May, 2006, Biplot analysis of Multi-environment trial data [slide presentation] www.ggebiplot.com/Biplot_Analysis_of_MET_Data_IITA.pps.

Yan, W. and Kang, M.S. 2003. *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC Press: Boca Raton, FL.

Yan, W., and N. A. Tinker. 2006. Biplot analysis of multi-environment trial data: principles and applications. Can. J. Plant Sci. 86: 623–645. Accessed at http://www.ggebiplot.com/Yan%26Tinker2006.pdf