PCA vs. Varimax rotation

The goal of the rotation/transformation in PCA is to maximize the variance of the ‘new’ SNP (eigenSNP), while minimizing the variance around the eigenSNP. Therefore the difference between the variances captured in each eigenSNP is maximized. The constraint, ‘$\Gamma'\Lambda\Gamma$ is diagonal’, on the coefficients of original SNPs and eigenSNPs is a mathematical convenience to make the coefficients unique; however, it can complicate the problem of interpretation. (See the scatter plot (figure 1) of the coefficients (table 2) from the dataset (table 1) where each SNP is represented as a point in the first 2 dimensions of the eigenspace.)

The interpretation of the coefficients is the most straightforward if each SNP is correlated highly on at most one eigenSNP, and if all the coefficient are either large or near zero, with few intermediate values. The SNPs are then split into disjoint sets, each of which is associated with one eigenSNP, perhaps some SNPs are left over.

To achieve this clear pattern of coefficients, we could rotate the axes defined by PCA in any direction without changing the relative locations of the points to each other in every two dimensions; but the actual coordinates of the points would change. The rotated solutions spam in the same geometric space as the original solutions and explain the same amount of variance in the data as the original solution, however the difference of the variances captured in the rotated axes is no longer maximized.

There are several analytical choices of rotation that have been proposed in the past. One of them is the varimax method of orthogonal rotation. The varimax rotation criterion maximizes the sum of the variances of the squared coefficients within each eigenvector, and the rotated axes remain orthogonal.

Figure 2 demonstrates the rotated solution (table 3) after a varimax rotation. After the coordinate axes are rotated clockwise by an angle about 45 degrees, we obtain a clear pattern of SNPs corresponding to rotated eigenSNPs.

In this simple example the overall interpretation is the same whether we rotate the axes or no, but in more complicated situations we could benefit more.

Tables:

	snp1	snp2	snp3	snp4
chr1	1	1	0	1
chr2	1	0	1	1
chr3	0	0	0	0
chr4	0	1	0	1
chr5	1	0	0	0

Table 1. a small dataset of 4 SNPs from 5 chromosomes

Variables	E1	E2
Snp1	0.0978322	0.5690011
Snp2	0.647965	-0.40432
Snp3	0.1198195	0.6968811
Snp4	0.745972	0.164681

Table 2. partial PCA results (unrotated)

Variables	E1	E2
Snp1	0.00012428	0.5773503
Snp2	0.70744623	-0.288827
Snp3	0.00015222	0.7071068
Snp4	0.70716891	0.2885229

Table 3. rotated solution

Figures:

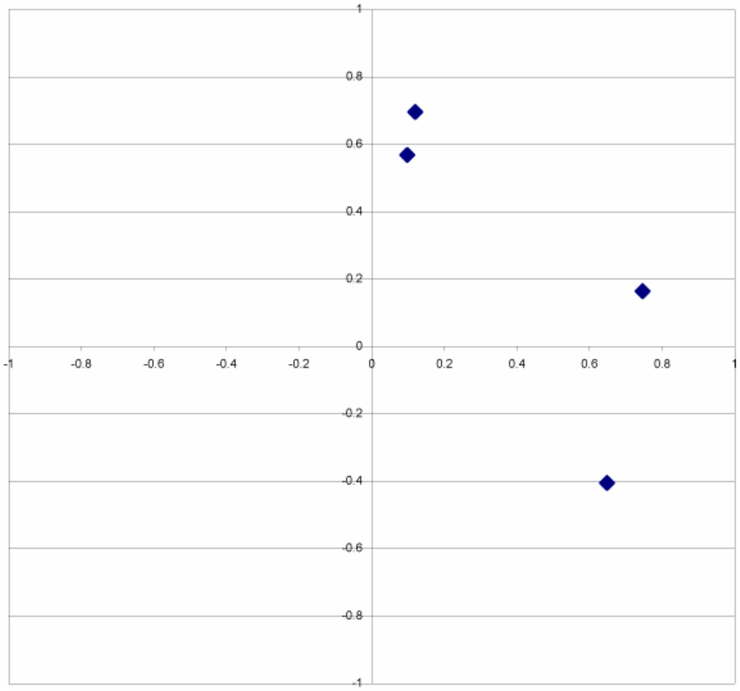

Figure 1. scatter plot of SNPs in the orthogonal space (unrotated)

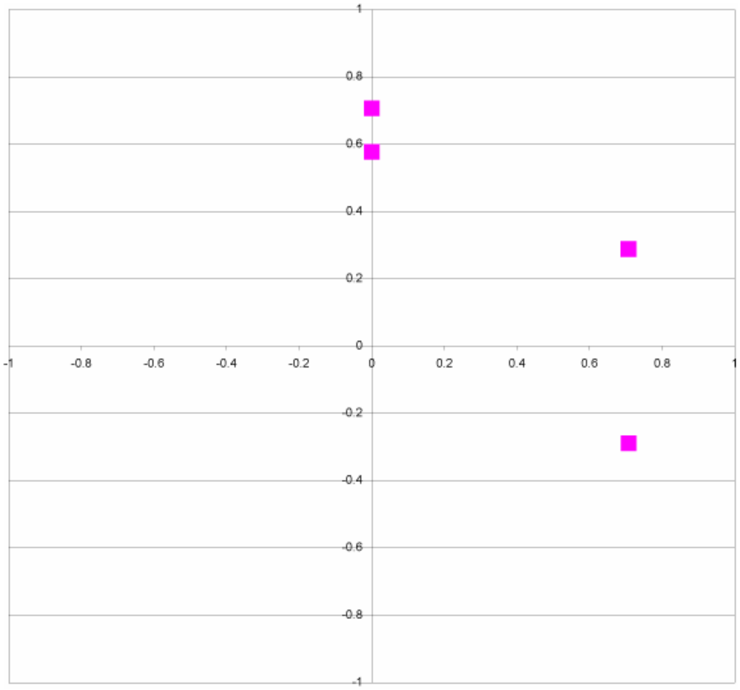

Figure 2. scatter plot of SNPs in the rotated orthogonal space

R source code:

```
data <- c(1,1,0,0,1,1,0,0,1,0,0,1,0,0,0,1,1,0,1,0)
dim(data) <- c(5,4)
snploadings <- loadings(princomp(data, cor=T))
plot(snploadings[,1:2])
rotated <- varimax(snploadings[,1:2])$loadings
plot(rotated)
```

An example of finding htSNPs from a small SNP dataset using the varimax rotation method.

We start with a SNP dataset:

	SNP1	SNP2	SNP3	SNP4	SNP5
Chromosome1	1	1	1	0	1
Chromosome2	1	1	0	0	0
Chromosome3	0	0	0	0	1
Chromosome4	0	0	1	1	0

PCA results, unrotated, are:

	e_1	e_2	e_3	e_4	e_5
SNP1	-0.5532	-0.3854	0	-0.7385	0
SNP2	-0.5532	-0.3854	0	0.6155	0.4082
SNP3	0.2025	-0.5265	-0.7071	0.1231	-0.4082
SNP4	0.5532	-0.3854	0	-0.2132	0.7071
SNP5	-0.2025	0.5265	-0.7071	-0.1231	0.4082

PCA results upon varimax roation (Mardia et al. 1979; Dunteman 1989) are:

	e_1^r	e_2^r	e_3^r	e_4^r	e_5^r
SNP1	0	0	0	-1	0
SNP2	-1	0	0	0	0
SNP3	0	-0.7071	-0.7071	0	0
SNP4	0	0	0	0	1
SNP5	0	0.7071	-0.7071	0	0

We compare the average coefficient for all *k* eigenSNPs (Γ_i) to the one for the rest of (*p*-*k*) eigenSNPs (γ_i) for each SNP; and select the SNP if $\Gamma_i > \gamma_i$, which indicates that this SNP contributes mostly to the *k* eigenSNP (Meng et al. 2003). Suppose *k* = 2, htSNP selections are:

	Γ	γ	htSNP
SNP1	0	0.5	N
SNP2	0.5	0	Y
SNP3	0.3536	0.2357	Y
SNP4	0	0.3333	N
SNP5	0.3536	0.2357	Y

References:

Dunteman GH (1989) Principal components analysis. Sage Publications, Newbury Park

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London ; New York

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115-130