PCA: Reymont Multivariate Data

Jonathan M. Lees
University of North Carolina, Chapel Hill
Department of Geological Sciences
CB #3315, Mitchell Hall
Chapel Hill, NC 27599-3315
email: jonathan.lees@unc.edu
ph: (919) 962-0695

November 10, 2008

1 Reyment Example

The following example is taken from the book *Applied Factor Analysis in the natural sciences* by Richard Reyment and K. G. Joreskog (UNC Library Call Number QA278.5/R49/1993). This book is pretty good, has simple explanations and a long, elementary introduction to matrix algebra.

The opening of the book includes an artificial case aimed at providing a tangible example illustrating the power and use of PCA.

In this example we consider a case of mining for valuable ores. A known deposit exists and we seek a new location where deposits are likely to exist in economic quantities. Where should we drill? How should we go about finding a low risk target. We collect a large number of geochemical analyses distributed throughout the region on a grid of sites. The measurements include chemical composition and observations of rock properties:

- Mg in calcite
- Fe in sphelerite
- Na in Muscovite

- Sulfide
- Crystal size of carbonates
- Spacing of cleavage
- Elongation of ooliths
- Tightness of folds
- Vein material per m^2
- Fractures per m^2

The question is, how should we use these measurements? Why is one variable more important than another? Are there combinations of these observables that best describe/predict the distribution of ore deposits?

To illustrate a possible solution to this problem, we make up a multi-component data set used to find ore deposit accumulations.

```
> Causes1 = c("distribution of paleotemperature", "distribution of deformation",
+     "distribution of permeability")
> Causes2 = c("T", "D", "P")
> GeoProps = c("Mg in calcite", "Fe in sphelerite", "Na in Muscovite",
+     "Sulfide", "Crystal size of carbonates", "Spacing of cleavage",
+     "Elongation of ooliths", "Tightness of folds", "Vein material per $m^2$",
+     "Fractures per $m^2$")
> GeoProps2 = c("Mg", "Fe", "Na", "Sulf", "Size", "Cleav", "Elong",
+     "Folds", "Vein", "Fract")
```

```
> datdir = "/home/lees/Progs/R_stuff/Data_Anal/Reyment"
> proport = matrix(scan(file = paste(sep = "/", datdir, "proport.dat"),
+     what = 0), ncol = 10, byrow = TRUE)
> amount = matrix(scan(file = paste(sep = "/", datdir, "amount.dat"),
+     what = 0), ncol = 3, byrow = TRUE)
> Edim = dim(proport)
> Adim = dim(amount)
> X = amount
> I = rbind(16:20, 15:11, 6:10, 5:1)
> J = I[rev(1:4), ]
> x = seq(from = 10, length = 5, by = 10)
> y = seq(from = 10, length = 4, by = 10)
> EX = matrix(rep(x, length(y)), byrow = TRUE, ncol = length(x))
> WHY = matrix(rev(rep(y, length(x))), byrow = FALSE, ncol = length(x))
```

The data is generated by noting that certain geological observations can be related to specific envirnomental causes. This is assumed for this problem. When we end up searching for the prospect we will not know this information *a priori*. So, for the sake of this synthetic case we set up the relationship between geological observations and a set of three parameters:

- distribution of paleotemperature, T
- distribution of deformation, D
- distribution of permeability, P

This model, or initial information, may be determined by laboratory experiments, or from vast experience. The model is used to project, or confound, the data, i.e. the model is used to connect the known envirnmental parameters to the geological measurements. The relative proportions are listed in Table 1 although they have been transposed for plotting on the page vertically.

Next consider the points on the ground (localities) where prehistoric changes in paleotemperature, deformation and permeabilty vary spatially. These are set artificially as described in Table 2. The localities are plotted and numbered spatially according to Figure 1. Note that the localities start at the lower-right (1) and snake around to the upper-right corner (20).

```
> postscript(file = "ReymLayout.eps", width = 6, height = 6, paper = "special",
+     horizontal = FALSE, onefile = TRUE, print.it = FALSE)
> plot(EX, WHY, type = "n", asp = 1, xlab = "East-West, km", ylab = "North-South, km")
> points(EX, WHY)
> points(EX, WHY, col = rgb(0.8, 0.8, 0.8), pch = 3)
> text(EX, WHY, I, pos = 3, xpd = TRUE, font = 2, cex = 1.2)
> polygon(JMLxA, JMLyA, col = "brown")
> dev.off()
```

	ID	T	D	P
Mg in calcite	Mg	0.95	0.00	0.05
Fe in sphelerite	Fe	0.75	0.10	0.15
Na in Muscovite	Na	0.75	0.20	0.05
Sulfide	Sulf	0.33	0.33	0.34
Crystal size of carbonates	Size	-0.20	0.60	0.60
Spacing of cleavage	Cleav	0.05	0.95	0.00
Elongation of ooliths	Elong	0.20	0.70	0.10
Tightness of folds	Folds	0.10	0.85	0.05
Vein material per m^2	Vein	0.00	0.10	0.90
Fractures per m^2	Fract	0.05	0.25	0.70

Table 1: Proportions of geological observables relative to environmental factors. This information is given *a priori*. It is used to create the synthetic data, so it is unknown to the explorationist.

	T	D	P
1	121.00	21.00	46.00
2	96.00	35.00	42.00
3	78.00	54.00	49.00
4	63.00	51.00	49.00
5	42.00	44.00	44.00
6	39.00	26.00	54.00
7	52.00	36.00	52.00
8	67.00	46.00	54.00
9	90.00	37.00	51.00
10	108.00	27.00	61.00
11	112.00	33.00	59.00
12	91.00	38.00	59.00
13	76.00	39.00	54.00
14	63.00	30.00	51.00
15	43.00	19.00	55.00
16	68.00	16.00	42.00
17	77.00	27.00	41.00
18	93.00	37.00	43.00
19	102.00	47.00	48.00
20	120.00	36.00	46.00

Table 2: Amounts at each locality. These are the environmental factors distributed at each position (left hand column) on the map. This information is used to create the synthetic data but may be unknown to a real life explorationist.

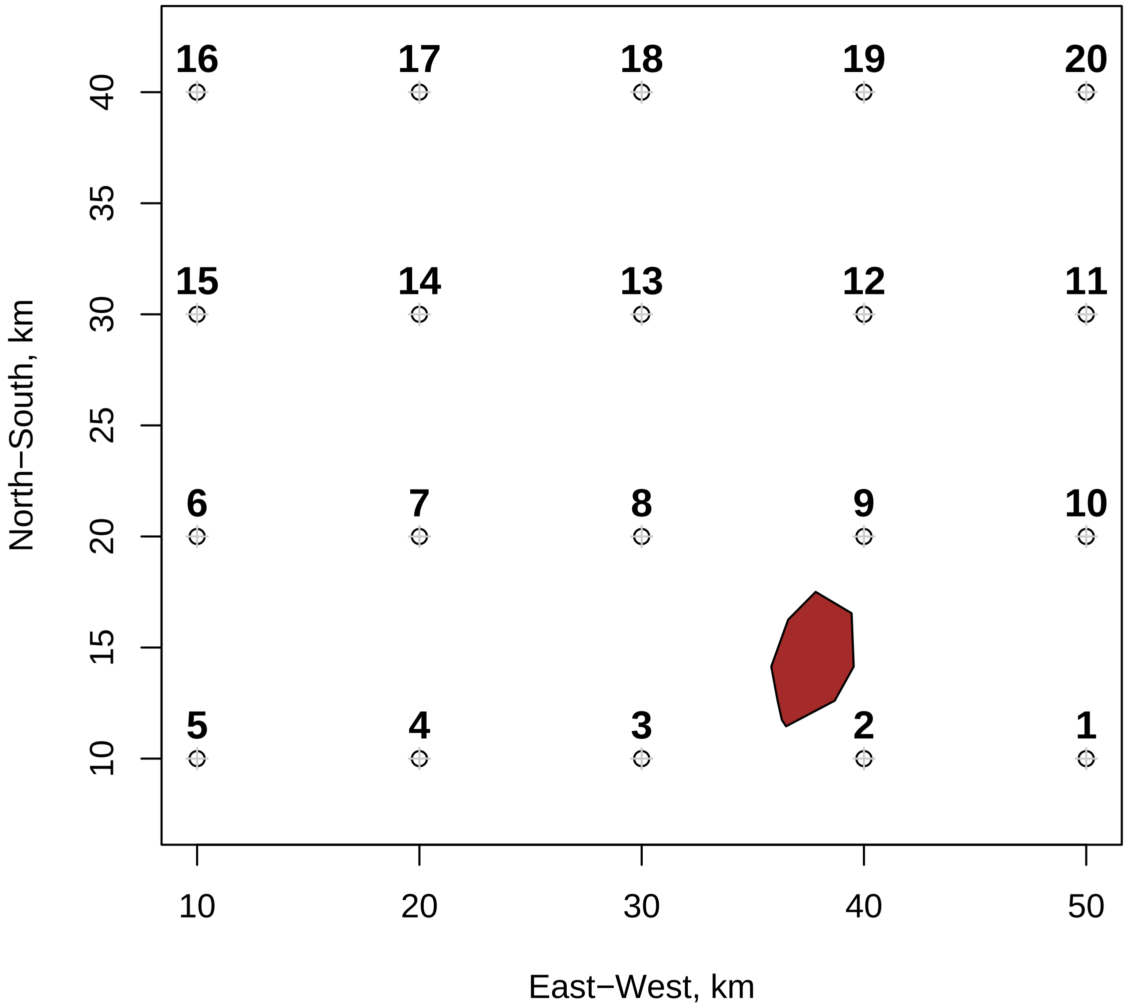

Figure 1: Geographic Layout of localities for Reyment Data

```
> setplotmat <- function(x, y) {
+     EX = matrix(rep(x, length(y)), byrow = TRUE, ncol = length(x))
+     WHY = matrix(rev(rep(y, length(x))), byrow = FALSE, ncol = length(x))
+     return(list(EX = EX, WHY = WHY))
+ }
> cnts = list()
> cnts[[1]] = c(40, 60, 80, 100)
> cnts[[2]] = c(20, 30, 40)
> cnts[[3]] = c(45, 50, 55, 60)
> xo = seq(from = 5, to = 55, by = 0.5)
> yo = seq(from = 5, to = 45, by = 0.5)
```

The three geologic variables (T, D, P) have been assigned to localities and can be contoured (Figure 2). If a successful prospect lies on this map (red zone), we may want to explore other areas where all three variables agree with variables in the known region. A quick visual inspection shows that the area with the dashed blue lines in Figure 4 appears to have (T,D,P) parameters that match the known prospect (red), i.e. roughly where $80 < T < 100$, $35 < D < 45$ and $45 < P < 50$. This can be seen to occur generally within the dashed blue boundary. Of course, the explorer will not have access to this map. Rather, they will have access to compilation of 10 geological measurements obtained at each site. The power of PCA is to show how we can still extract this information without knowing *a priori* information or how the environmental variables are related to the geologic observations.

```
> library(spatial)
> postscript(file = "Reyment1.eps", width = 8, height = 8, paper = "special",
+     horizontal = FALSE, onefile = TRUE, print.it = FALSE)
> plot(EX, WHY, type = "n", asp = 1, xlim = c(5, 55), ylim = c(5,
+     45), xlab = "East-West, km", ylab = "North-South, km")
> points(EX, WHY, col = rgb(0.8, 0.8, 0.8), pch = 3)
> polygon(JMLxA, JMLyA, col = "brown")
> polygon(JMLxB, JMLyB, lty = 2, lwd = 2, border = 4)
> for (j in 1:3) {
+     Z = matrix(X[I, j], ncol = 5, nrow = 4)
+     JSUR = data.frame(x = as.vector(EX), y = as.vector(WHY),
+         z = as.vector(Z))
+     JSUR.kr <- surf.gls(3, expcov, JSUR, d = 200)
+     prsurf <- prmat(JSUR.kr, 0, 60, 0, 50, 100)
+     contour(prsurf, levels = cnts[[j]], xlim = c(0, 60), ylim = c(0,
+         50), lty = j, add = TRUE, col = j, method = "flattest",
+         vfont = c("sans serif", "plain"))
+ }
> legend("bottomleft", legend = Causes2, lty = 1:3, col = 1:3,
+     bg = "white")
> dev.off()
```

To create the data (unknown to the observer) we perform a matrix multiplication to "confound" the environmental proportions ($\boldsymbol{E}$, Table 1) with the locality information ($\boldsymbol{A}$, Table 2). This is presented as a matrix equation

$$\boldsymbol{D} = \boldsymbol{A}\boldsymbol{E} \tag{1.1}$$

Since $\boldsymbol{A}$ is a 20×3 matrix and $\boldsymbol{E}$ is a 3×10 matrix, the resulting "DATA" is a 20×10 matrix representing 10 geological measurements at each of the positions shown in Figure 1. The data is presented in Table 3. In **R** this is achieved by,

```
> prospect = amount %*% proport
```

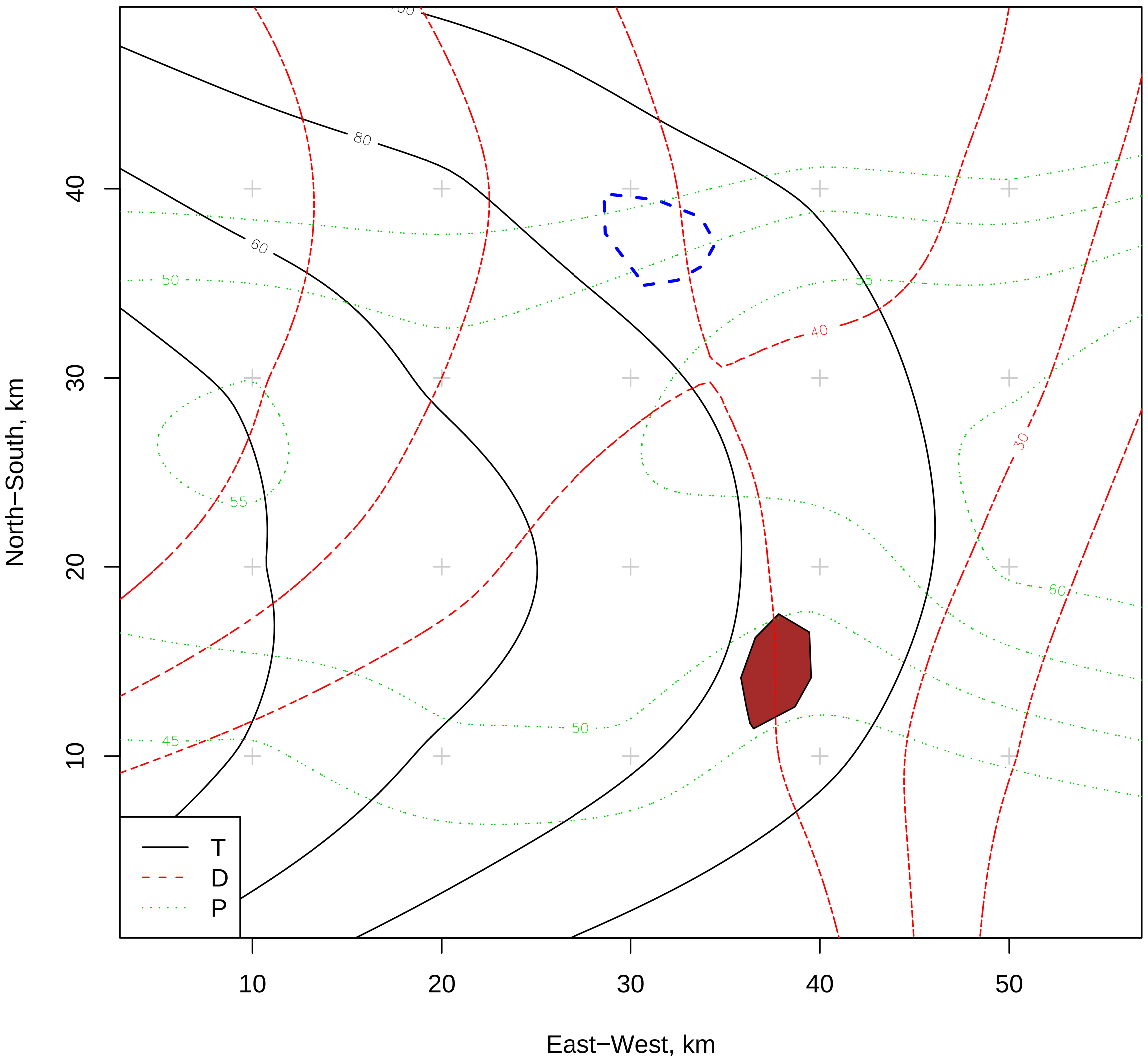

Figure 2: Initial Reyment Contour of Known Environmental Factors. These are used to create the synthetic data. The prospect is hi-lighted in red. A possible (unknown) prospect is shown with a dashed blue line.

	Mg	Fe	Na	Sulf	Size	Cleav	Elong	Folds	Vein	Fract
1	117.25	99.75	97.25	62.50	16.00	26.00	43.50	32.25	43.50	43.50
2	93.30	81.80	81.10	57.51	27.00	38.05	47.90	41.45	41.30	42.95
3	76.55	71.25	71.75	60.22	46.20	55.20	58.30	56.15	49.50	51.70
4	62.30	59.70	59.90	54.28	47.40	51.60	53.20	52.10	49.20	50.20
5	42.10	42.50	42.50	43.34	44.40	43.90	43.60	43.80	44.00	43.90
6	39.75	39.95	37.15	39.81	40.20	26.65	31.40	28.70	51.20	46.25
7	52.00	50.40	48.80	46.72	42.40	36.80	40.80	38.40	50.40	48.00
8	66.35	62.95	62.15	55.65	46.60	47.05	51.00	48.50	53.20	52.65
9	88.05	78.85	77.45	59.25	34.80	39.65	49.00	43.00	49.60	49.45
10	105.65	92.85	89.45	65.29	31.20	31.05	46.60	36.80	57.60	54.85
11	109.35	96.15	93.55	67.91	32.80	36.95	51.40	42.20	56.40	55.15
12	89.40	80.90	78.80	62.63	40.00	40.65	50.70	44.35	56.90	55.35
13	74.90	69.00	67.50	56.31	40.60	40.85	47.90	43.45	52.50	51.35
14	62.40	57.90	55.80	48.03	36.00	31.65	38.70	34.35	48.90	46.35
15	43.60	42.40	38.80	39.16	35.80	20.20	27.40	23.20	51.40	45.40
16	66.70	58.90	56.30	42.00	21.20	18.60	29.00	22.50	39.40	36.80
17	75.20	66.60	65.20	48.26	25.40	29.50	38.40	32.70	39.60	39.30
18	90.50	79.90	79.30	57.52	29.40	39.80	48.80	42.90	42.40	44.00
19	99.30	88.40	88.30	65.49	36.60	49.75	58.10	52.55	47.90	50.45
20	116.30	100.50	99.50	67.12	25.20	40.20	53.80	44.90	45.00	47.20

Table 3: Reyment Data. Each Column represents a specific geologic observable measured at locations indicated by the rows.

1.1 Data Analysis: Inversion

Now that we have an synthetic data set we can treat the data as if it were real data, and attempt to extract information that will guide us on where to prospect.

The next set of calculations scales the data matrix, produces the variance-covariance matrix and extracts the singular values and eigenvectors. We proceed as in R-factor analysis.

```
> X = prospect
> N = dim(X)
> Z = scale(X)
> r = var(Z)
> S = svd(r)
> D = diag(S$d)
> E = cbind(S$d, cumsum(S$d), 100 * cumsum(S$d))
> EE = cbind(S$d, 100 * S$d/sum(S$d), cumsum(100 * S$d/sum(S$d)))
```

Inspecting the singular value of the SVD gives an idea on what might be a useful rotation and projection scheme. If we plot the Singular values as a spectrum we can visually estimate how large the dimensionality of the projection should be. Figure 3 shows that three singular values dominate the spectrum, suggesting that three dimensions should capture most of the information contained in the data.

Indeed, as shown in Table 4 of the singular values, their percentages and cumulative sums this conclusion is bourne out.

```
> postscript(file = "reyEspectrum.eps", width = 6, height = 6,
+     paper = "special", horizontal = FALSE, onefile = TRUE, print.it = FALSE)
> plot(1:length(S$d), S$d, xlab = "singular value index", ylab = "Singular Value",
+     type = "h", lwd = 2, col = "blue", axes = FALSE)
> axis(1, at = 1:length(S$d))
> axis(2, at = 1:length(S$d))
> box()
> ex1 = rep(4, times = 3)
> why1 = rep(4, times = 3)
> ex2 = 1:3
> why2 = S$d[1:3]
> vx1 = ex2 - ex1
> vy1 = why2 - why1
> len1 = sqrt(vx1^2 + vy1^2)
> Ax = ex1 + 0.95 * vx1
> Ay = why1 + 0.95 * vy1
> arrows(rep(4, times = 3), rep(4, times = 3), Ax, Ay, length = 0.1)
> text(4, 4, "Large Singular Values", pos = 4)
> dev.off()
```

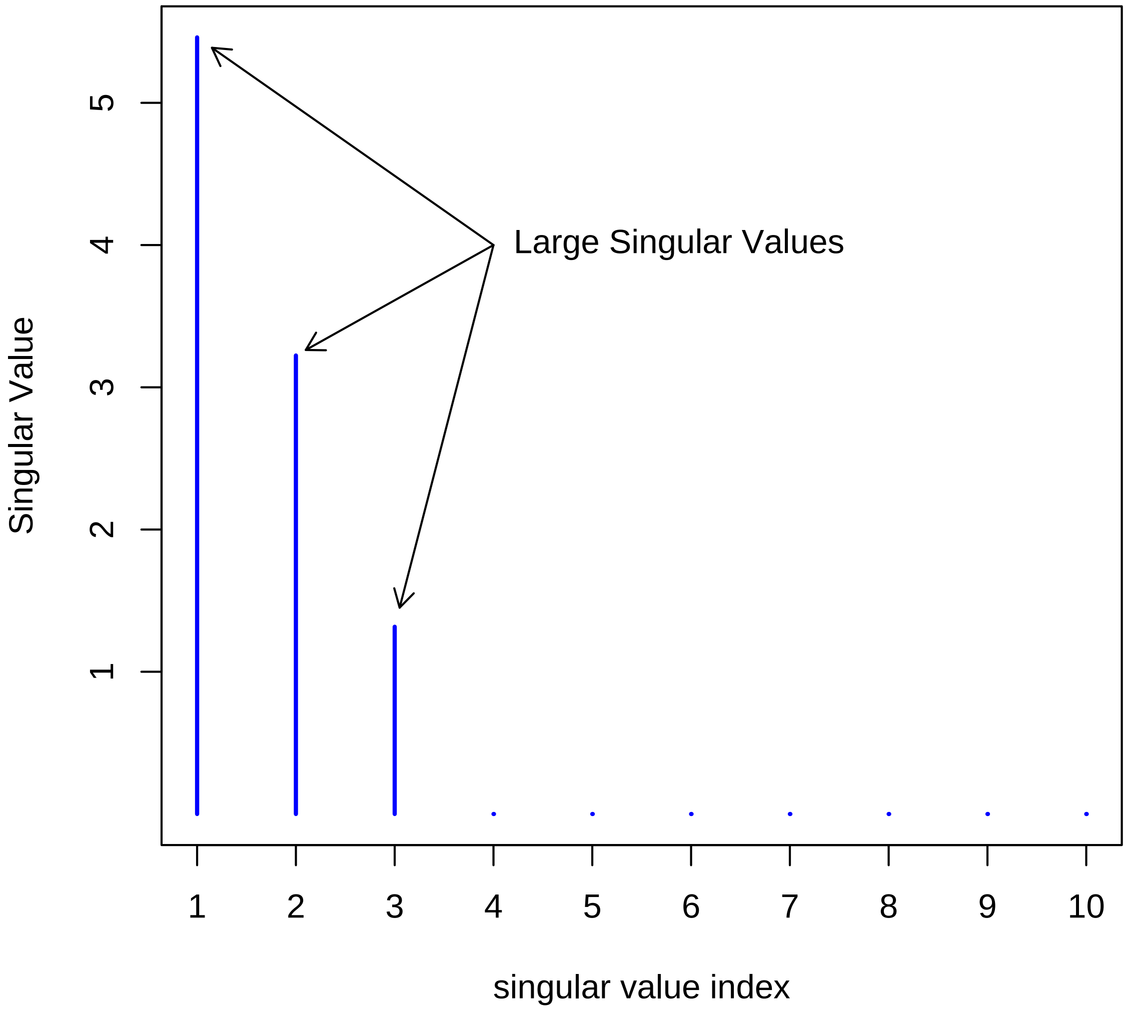

Figure 3: Spectrum of Singular Values for Reyment data

	Value	Percent	CumSum
1	5.46	54.60	54.60
2	3.22	32.24	86.84
3	1.32	13.16	100.00
4	0.00	0.00	100.00
5	0.00	0.00	100.00
6	0.00	0.00	100.00
7	0.00	0.00	100.00
8	0.00	0.00	100.00
9	0.00	0.00	100.00
10	0.00	0.00	100.00

Table 4: Singular values, percentages and cumulative sums.

```
> U = S$u
> U[, 1] = -U[, 1]
```

The **R** program `varimax` is used to adjust the new coordinates slightly for optimal projection. Finally, the loadings are extracted and used to plot contours of the largest three eigenvectors. For plotting, the values must be positioned in their repsective locations illustrated in Figure 1.

```
> U = U[, 1:3]
> D = D[, 1:3]
> lding = U %*% sqrt(diag(diag(D)))
> vm = varimax(lding, normalize = TRUE, eps = 1e-05)
> vfactor = vm$loadings[, 1:3]
> f = Z %*% vfactor %*% solve(t(vfactor) %*% vfactor)
```

```
> postscript(file = "reymentANAL1.eps", width = 6, height = 6,
+     paper = "special", horizontal = FALSE, onefile = TRUE, print.it = FALSE)
> plot(EX, WHY, type = "n", asp = TRUE, xlim = c(5, 55), ylim = c(5,
+     45), xlab = "East-West, km", ylab = "North-South, km")
> points(EX, WHY, col = rgb(0.8, 0.8, 0.8), pch = 3)
> polygon(JMLxA, JMLyA, col = "brown")
> polygon(JMLxB, JMLyB, lty = 2, lwd = 2, border = 4)
> BOB = list()
> for (j in 1:3) {
+     Z = matrix(f[I, j], ncol = 5, nrow = 4)
+     JSUR = data.frame(x = as.vector(EX), y = as.vector(WHY),
+         z = as.vector(Z))
+     JSUR.kr <- surf.gls(3, expcov, JSUR, d = 200)
+     prsurf <- prmat(JSUR.kr, 0, 60, 0, 50, 100)
+     BOB[[j]] = prsurf
+     contour(prsurf, xlim = c(0, 60), ylim = c(0, 50), lty = j,
+         add = TRUE, col = j)
+ }
> polygon(JMLxB, JMLyB, lty = 2, lwd = 2, border = 4)
> legend("bottomleft", legend = c("Factor 1", "Factor 2", "Factor 3"),
+     lty = 1:3, col = 1:3, bg = "white")
> dev.off()
```

The last step is find locations of the three principal components that match values of the known location of ore deposits. Since the eigenvectors are mixtures of the observable parameters, they may or may not relate to the underlying environemtal factors. That is, it may not be easy to see the connection, or the connection may be unknown. The decomposition, however, is not concerned with the underlying geologic reasons for the correlation we are about to reveal. (Although investigation and study of the relationship of the variables may have important implications for geological research.) For now we are mainly concerned with locating geographic locations for high probabilty prospects.

For each singular vector, we cycle through and find the known high score values - i.e. those values that coincide with the known ore body. We then find all locations in the three contour maps that match all three conditions. The points where these conditions are met are hi-lighted on Figure 5, where points are target plotted as small circles. Note the original ore body is clearly marked - the red body is covered with "hits". Further inspection reveals a second target area, the unknown prospect pops out as a potential area mining success.latex Reyment.tex

```
> Rx = range(JMLxA)
> Ry = range(JMLyA)
> j = 1
> my = list()
> for (j in 1:3) {
+     SJ = setplotmat(BOB[[j]]$x, BOB[[j]]$y)
+     flagx = BOB[[j]]$x > Rx[1] & BOB[[j]]$x < Rx[2]
+     flagy = BOB[[j]]$y > Ry[1] & BOB[[j]]$y < Ry[2]
+     BOB[[j]]$x[flagx]
+     BOB[[j]]$y[flagy]
+     VARBLE = BOB[[j]]$z[flagx, flagy]
+     my[[j]] = range(VARBLE)
+     print(my[j])
+ }
> KAM = list()
> for (j in 1:3) {
+     zim = BOB[[j]]$z
+     KAM[[j]] = zim > my[[j]][1] & zim < my[[j]][2]
+ }
> BIGKAM = KAM[[1]] & KAM[[2]] & KAM[[3]]
> BIGKAM = t(BIGKAM)
> BIGKAM = BIGKAM[101:1, ]
> W = which(BIGKAM)
> targ = setplotmat(BOB[[j]]$x, BOB[[j]]$y)
```

```
> postscript(file = "ReyFigFinal.eps", width = 6, height = 6, paper = "special",
+     horizontal = FALSE, onefile = TRUE, print.it = FALSE)
> plot(EX, WHY, type = "n", asp = 1, xlim = c(5, 55), ylim = c(5,
+     45), xlab = "East-West, km", ylab = "North-South, km")
```

```
> points(EX, WHY, pch = 3, col = grey(0.8))
> polygon(JMLxA, JMLyA, col = "brown")
> BOB = list()
> for (j in 1:3) {
+     Z = matrix(f[I, j], ncol = 5, nrow = 4)
+     JSUR = data.frame(x = as.vector(EX), y = as.vector(WHY),
+         z = as.vector(Z))
+     JSUR.kr <- surf.gls(3, expcov, JSUR, d = 200)
+     prsurf <- prmat(JSUR.kr, 0, 60, 0, 50, 100)
+     BOB[[j]] = prsurf
+     contour(prsurf, xlim = c(0, 60), ylim = c(0, 50), lty = j,
+         add = TRUE, col = j)
+ }
> polygon(JMLxB, JMLyB, lty = 2, lwd = 2, border = 4)
> points(targ$EX[W], targ$WHY[W], pch = 1)
> legend("bottomleft", legend = c("Factor 1", "Factor 2", "Factor 3"),
+     lty = 1:3, col = 1:3, bg = "white")
> dev.off()
```

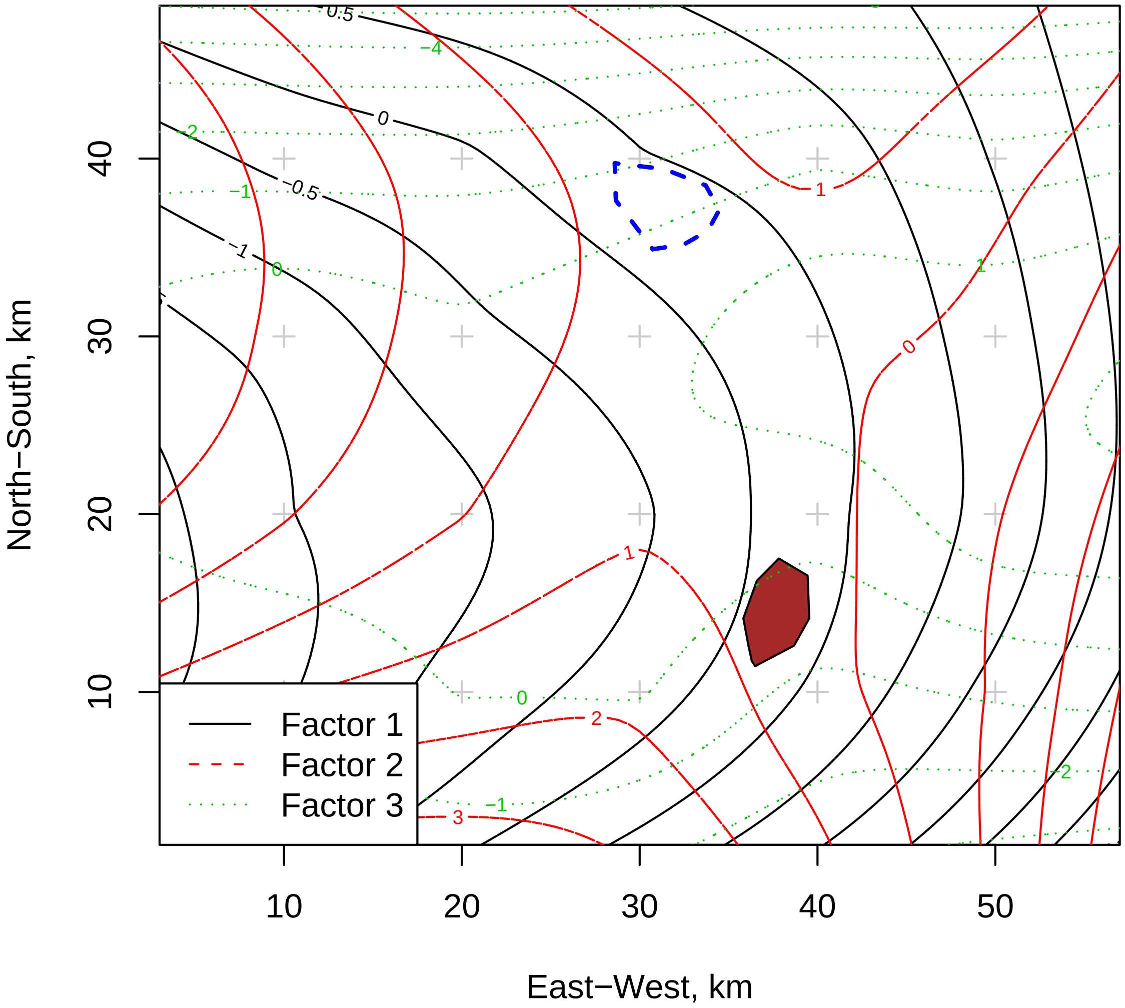

Figure 4: Reyment Contours of three largest eignevectors. These contours come from PCA analysis of the data and represent combinations of the geological measurements made at each of the site locations. The contour levels may or may not be interpretable or meaningful. The important point is to note the pattern of where the three contour fields match the known ore body values. The original ore body is plotted in red. Comparing similar contour values, a possible (unknown) prospect is shown with a dashed blue line.

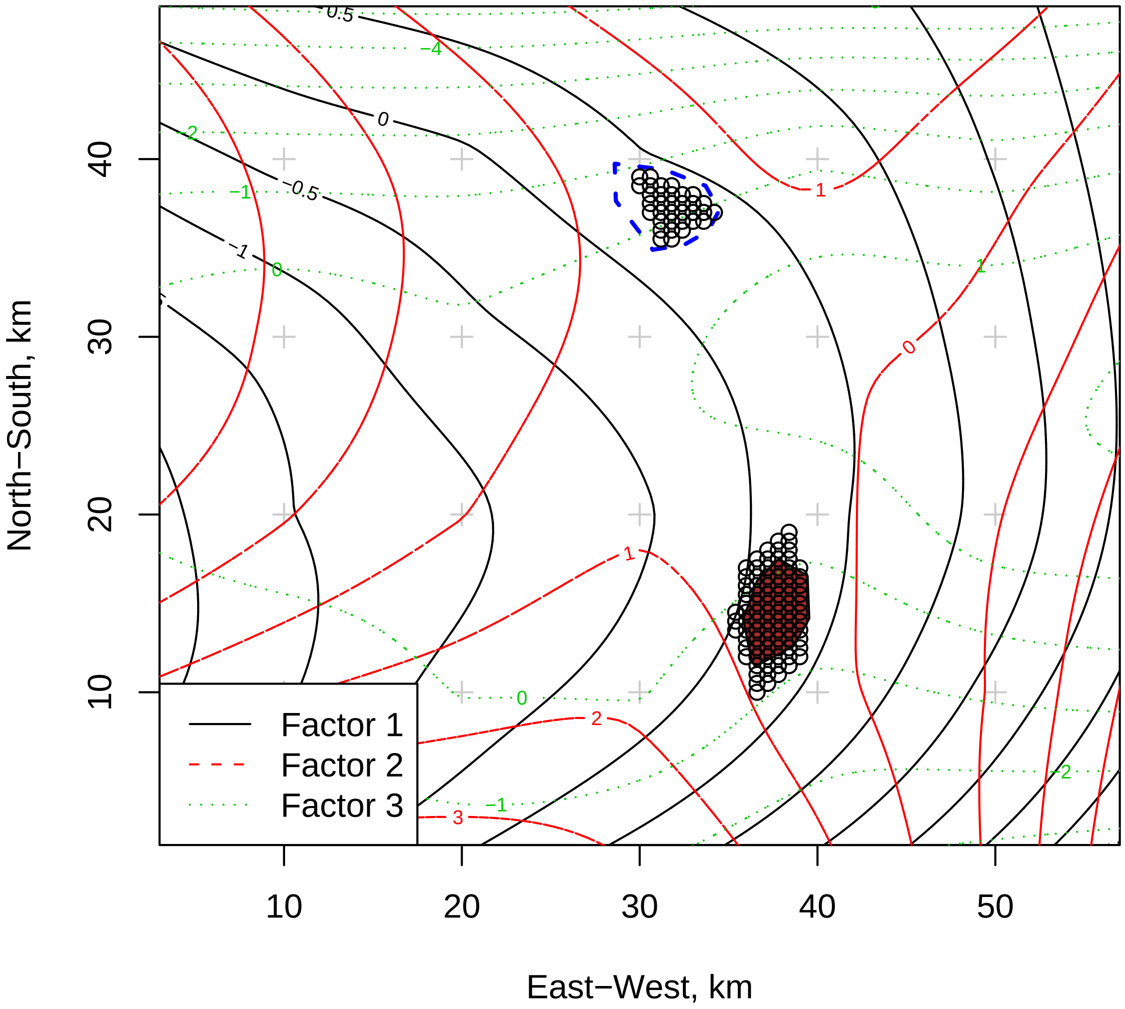

Figure 5: Reyment Ore body target found quantitatively. This is the same as Figure 4 but the location of where the values agree is generated automatically in **R** . Each plotted point represents a location where all three eigenvector values fell within a specified target range.