Directional and Circular Data

- The analysis of directional data is relevant, for instance, in investigating
    - movement pattern
    - climatological records,
    - geological fracturing etc.
- A good introduction to circular statistics can be found in
    - Upton and Fingleton (1989). *Spatial Data Analysis by Example. Categorical and Directional Data. Volume 2.* New York: John Wiley & Sons.
    - See also **DAVISSPATIALANALYSIS.PDF** pp 316-330.
- The R package **`circular`** covers many methods of analyzing and describing directional data.
    - This package and additional functions are discussed in Pewsey, Neuhaeuser and Ruxton (2013). *Circular Statistics in R*. Oxford Press
- The GIS software TNTMips (www.microimages.com) implemented some statistics for directional data.

Problems and Rose Diagram

- Problems with ***cyclic*** data, where the ***direction*** matters, are:
    - not having natural origin
    - having a wrapped scale which prohibits calculating average directions:
        - $(1^\circ + 2^\circ)/2 = 1.5^\circ \Rightarrow$ okay value
        - $(359^\circ + 1^\circ)/2 = 180^\circ \neq 0^\circ \Rightarrow$ insensible value

- Therefore, the standard histogram is replaced by a circular histogram:

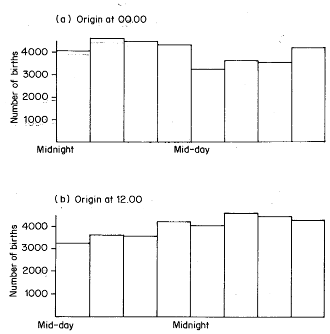
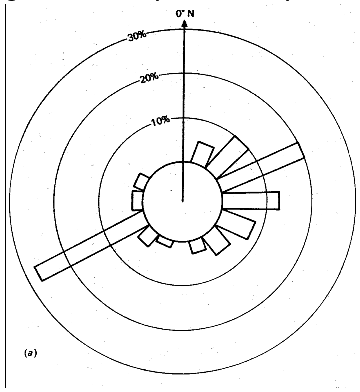


Figure 9.2 Alternative linear histograms for the birth data (Table 9.7)



(a)

Directional Mean and Variance

- To calculate the ***directional mean*** one adds the directional vectors (assume to have unit length) and determine the direction of the resultant vector.

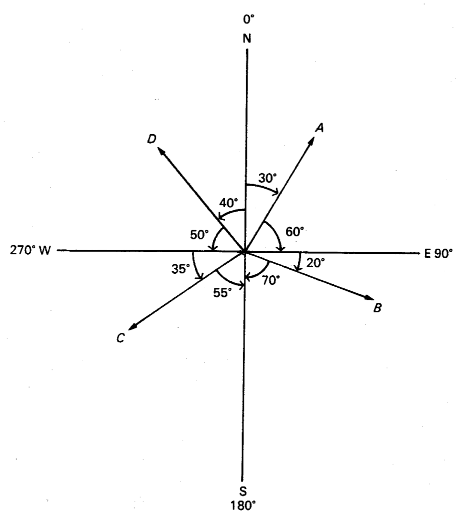


FIGURE 3-16. Measurement of direction.

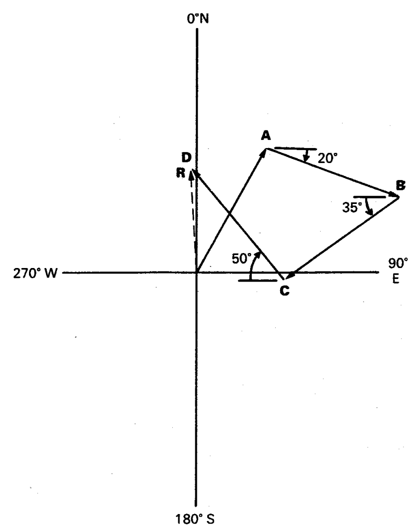
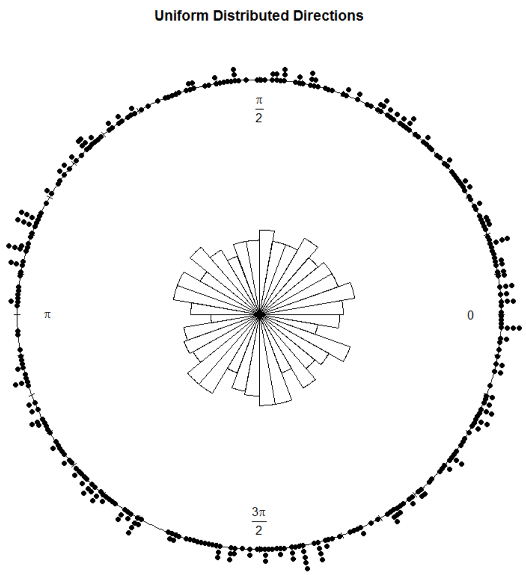


FIGURE 3-19. Graphical solution for finding average direction.

- Since the individual vectors have unit lengths
  - their aggregated endpoints can be express by sin- and cos-expressions:
$$\bar{X} = \frac{\sum_{i=1}^{n} \cos\theta_i}{n} \; and \; \bar{Y} = \frac{\sum_{i=1}^{n} \sin\theta_i}{n}$$
  - The $\arctan$-function can be used to derive the average direction
$$\bar{\theta} = \begin{cases} \arctan\left(\bar{X}/\bar{Y}\right) & \text{if } \bar{Y} > 0, \bar{X} > 0 \\ \arctan\left(\bar{X}/\bar{Y}\right) + 180° & \text{if } \bar{Y} < 0 \\ \arctan\left(\bar{X}/\bar{Y}\right) + 360° & \text{if } \bar{Y} > 0, \bar{X} < 0 \end{cases}$$
- ***Axial*** data with ***equivalent directions*** along a line segment have the property that $\theta \Leftrightarrow \theta + 180°$, the transformation making both angles truly equivalent is
$$\theta^* = \begin{cases} \theta & \text{if } \theta \leq 180° \\ \theta - 180° & \text{if } 180° < \theta \leq 360° \end{cases}$$
with $\theta^* \in (0° \ldots 180°)$.
For example, $\theta = 180° + 30° = 210° \Leftrightarrow \theta^* = 210° - 180° = 30°$
- The length of the ***standardized resultant*** $\bar{R} = \sqrt{\bar{X}^2 + \bar{Y}^2}$ is the length of the directional mean vector.
- The standardized resultant $\bar{R}$ can be used to test whether all directions are equally prevalent (i.e., uniform distribution $\theta_i \sim U[0, 2\cdot\pi]$) or cluster around a particular direction:
  - If all vectors point into the same direction then the length of $\bar{R} = 1$
  - If all vectors are uniformly distributed (isotropic) with $\theta_i = U[0°, 360°]$ then the length of $\bar{R} = 0$
  - A measure of ***circular variance*** is $s = 1 - \bar{R}$
- See code `Circular.r` for examples:




```
> summary(x)
          n        Mean         Rho
400.00000000 -0.52181515  0.04044952
```

Mean direction: -0.52 radiants
Standardized resultant vector length: 0.04

```
> rayleigh.test(x)              # test for uniformity

       Rayleigh Test of Uniformity
       General Unimodal Alternative

Test Statistic:  0.0404
P-value:  0.5197
```

Von Mises Distribution

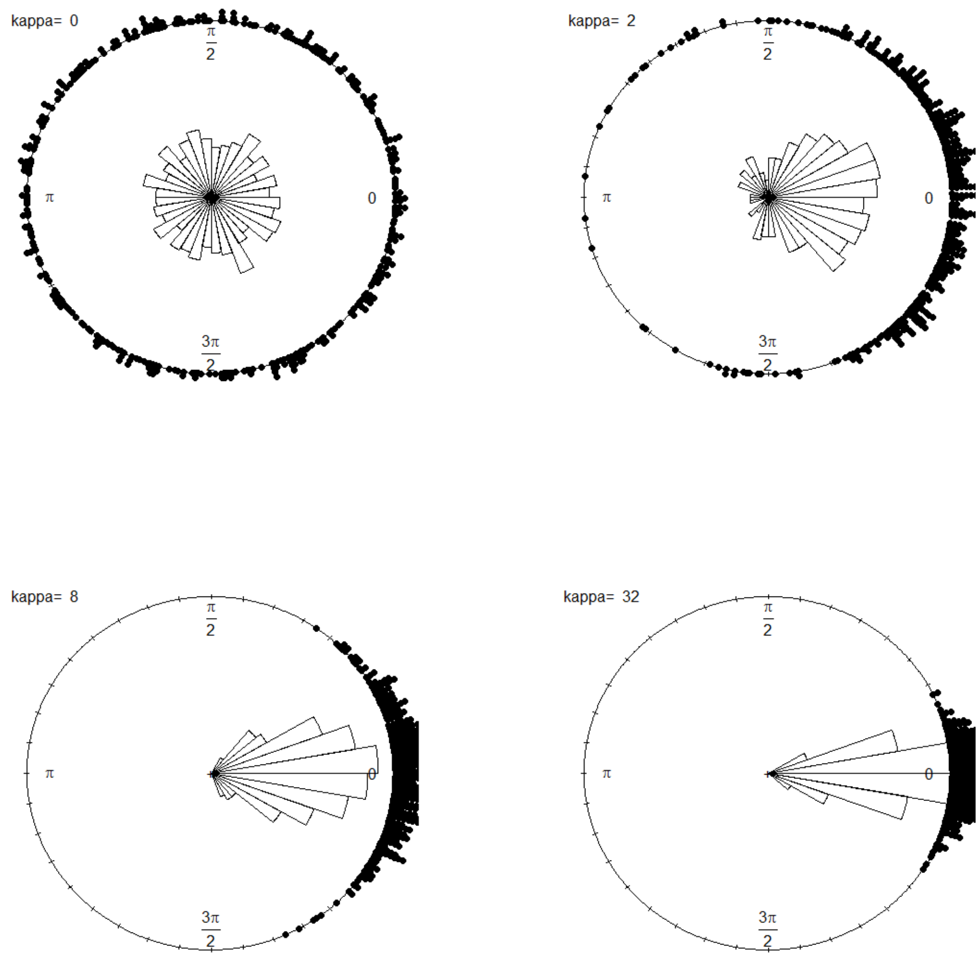
- An alternative model, which deviates from the isotropic distribution assumption, is the ***von Mises distribution***. It favors a particular direction at varying strength:

  $f(\theta|\mu,\kappa) = \frac{1}{2\cdot\pi\cdot I_0(\kappa)} \cdot \exp(\kappa \cdot \cos(\theta - \mu))$ with $0 \leq \theta \leq 2\cdot\pi$, $0 \leq \mu \leq 2\cdot\pi$, and $\kappa \geq 0$

  where $I_0(\kappa) = \sum_{r=0}^{\infty} \frac{(\kappa/2)^{2\cdot r}}{r!\cdot\Gamma(r+1)}$ is the Bessel function of order zero.

  - The normalization constant $I_0(\kappa) = \int_0^{2\cdot\pi} \frac{1}{2\cdot\pi} \cdot \exp(\kappa \cdot \cos(\theta - \mu)) \cdot d\theta$ guarantees that the *von Mises* distribution integrates to one.
  - The parameter μ is the expected direction and κ is the spread around the expected direction.
  - The parameter κ controls the concentration of the directions, with $\kappa = 0 \Rightarrow \theta = U[0,2\cdot\pi]$

- Simulated distributions in dependence of the concentration parameter κ:



To plot directional data see the R-script **`PewsyChapt02.R`**.

Correlations Using Circular Data (not test relevant)

- Classes of associations:
  - Cylindrical data: The association between circular data and a metric variable. E.g., in a wind farm the relationships between the wind direction and velocity of the wind.
    ⇨ Johnson-Wehrly-Mardia Correlation Coefficient
  - Toroidal data: The association between two circular variables. E.g., wind directions at two points in time; flight direction of migrating birds and the prevailing wind direction.
    ⇨ Jammalamadaka-Sarma Correlation Coefficient

Like daily temperature

- See R-script **`PewsyChapt08Correlations.R`**

Has methods for unknown distribution

Measures of Spatial Spread of Points

Spread Ellipsoids

Green area are standard deviation from the central mean

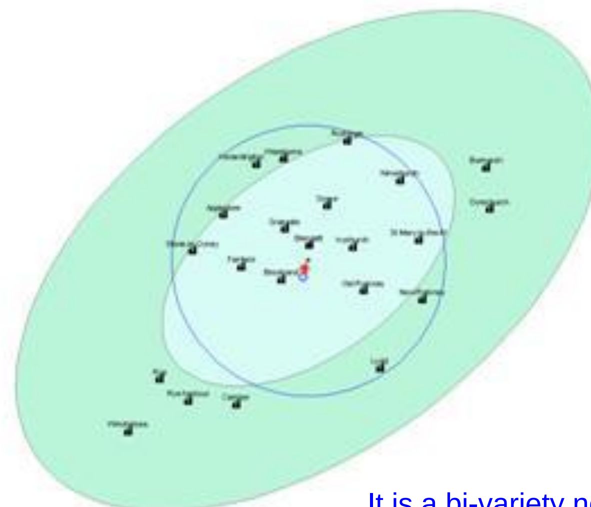
- The ***standard distance*** for a set of *n* points (X_i, Y_i) around the mean center $(\bar{X}, \bar{Y})$ is calculated by

$$SD = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n} + \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n}} = \sqrt{\tfrac{1}{n} \cdot \sum_{i=1}^{n} d_i^2} \; .$$

This expression ignores any directional orientation of the point cloud.

- To accommodate the orientation one first needs to identify the angle

$$\theta = \arctan\left(\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \sqrt{C}}{2 \cdot \sum_{i=1}^{n}\left(X_i - \bar{X}\right) \cdot \left(Y_i - \bar{Y}\right)}\right)$$



It is a bi-variety normal distribution
Because the slope is not flat

$$\text{with } C = \left[\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 - \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 \right]^2 + 4 \cdot \sum_{i=1}^{n} \left[\left(X_i - \bar{X} \right) \cdot \left(Y_i - \bar{Y} \right) \right]^2$$

and the calculate the spread along each major axis

$$SD_X = \sqrt{\frac{2 \cdot \sum_{i=1}^{n} \left[\cos\theta \cdot \left(X_i - \bar{X} \right) - \sin\theta \cdot \left(Y_i - \bar{Y} \right) \right]^2}{n-2}} \text{ and } SD_Y = \sqrt{\frac{2 \cdot \sum_{i=1}^{n} \left[\sin\theta \cdot \left(X_i - \bar{X} \right) - \cos\theta \cdot \left(Y_i - \bar{Y} \right) \right]^2}{n-2}}$$

Normal Distribution Mixtures

- The distribution of points with multiple centers and ellipsoidal standard distances can be modeled by a mixture of multi-normal distributions:

  $\begin{bmatrix} x_i \\ y_i \end{bmatrix} \sim \sum_{p=1}^{P} \pi_p \cdot N\left(\begin{bmatrix} \mu_{x,p} \\ \mu_{y,p} \end{bmatrix}, \begin{bmatrix} \sigma_{x,p}^2 & \sigma_{xy,p} \\ \sigma_{xy,p} & \sigma_{y,p}^2 \end{bmatrix} \right)$ with $\pi_p > 0$ and $\sum_{p=1}^{P} \pi_p = 1$

- Unknowns are [a] the number of clusters P, [b] the proportion of points per cluster π_p, [c] the P cluster centers $\begin{bmatrix} \mu_{x,p} \\ \mu_{y,p} \end{bmatrix}$ and [d] their P ellipsoidal spreads $\begin{bmatrix} \sigma_{x,p}^2 & \sigma_{xy,p} \\ \sigma_{xy,p} & \sigma_{y,p}^2 \end{bmatrix}$.

- The covariance matrices of the different classes can be constrained using different restrictions:
  - Volume: each cluster has approximately the number of observations
  - Shape: each cluster has approximately the same variance so that the distribution is spherical
  - Orientation: each cluster is forced to be axis-aligned

TABLE 22.1: Parameterizations of the covariance matrix

Model	Family	Volume	Shape	Orientation	Identifier
1	Spherical	Equal	Equal	NA	EII
2	Spherical	Variable	Equal	NA	VII
3	Diagonal	Equal	Equal	Axes	EEI
4	Diagonal	Variable	Equal	Axes	VEI
5	Diagonal	Equal	Variable	Axes	EVI
6	Diagonal	Variable	Variable	Axes	VVI
7	General	Equal	Equal	Equal	EEE
8	General	Equal	Variable	Equal	EVE
9	General	Variable	Equal	Equal	VEE
10	General	Variable	Variable	Equal	VVE
11	General	Equal	Equal	Variable	EEV
12	General	Variable	Equal	Variable	VEV
13	General	Equal	Variable	Variable	EVV
14	General	Variable	Variable	Variable	VVV

- The different specifications of the parameters of the cluster covariance matrices allows the Gaussian mixture model to capture different clustering configurations.

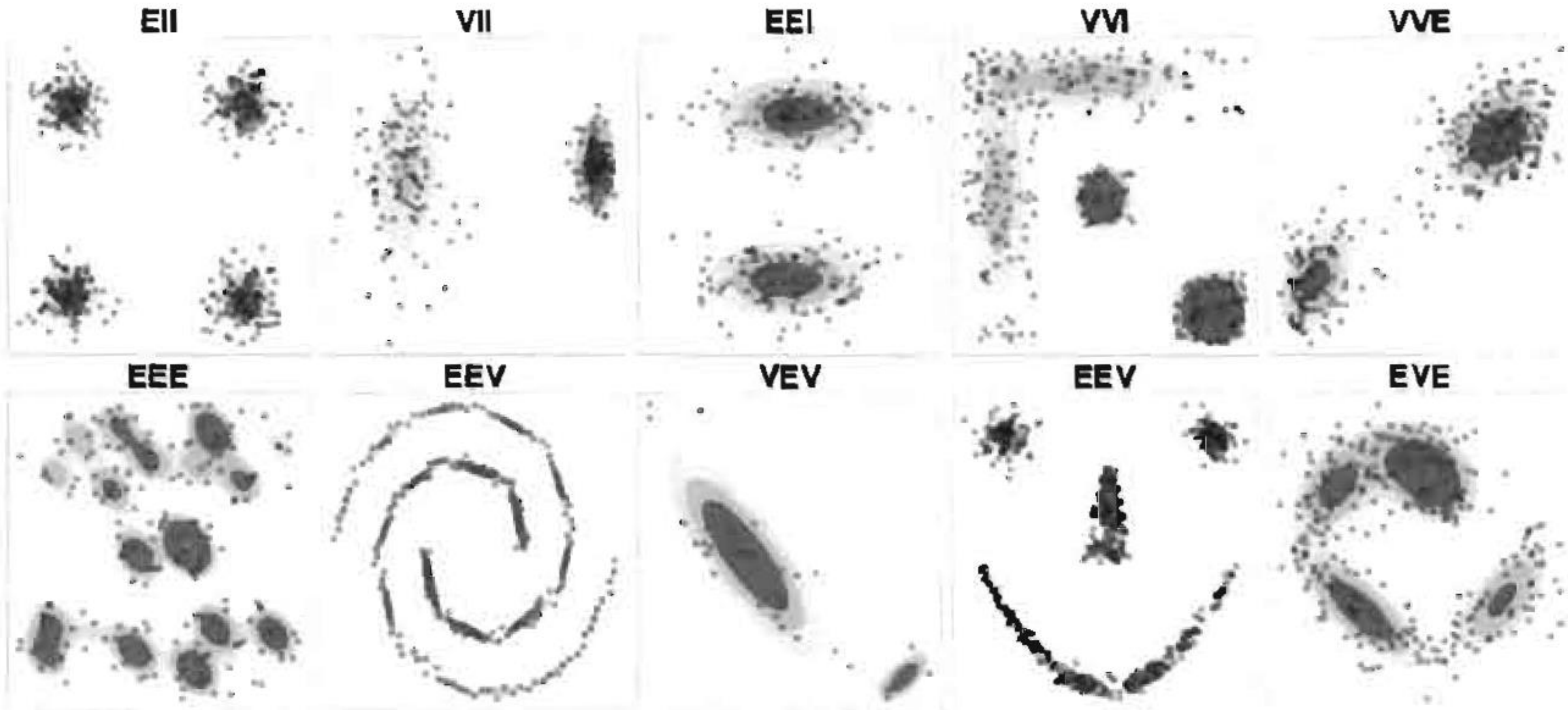
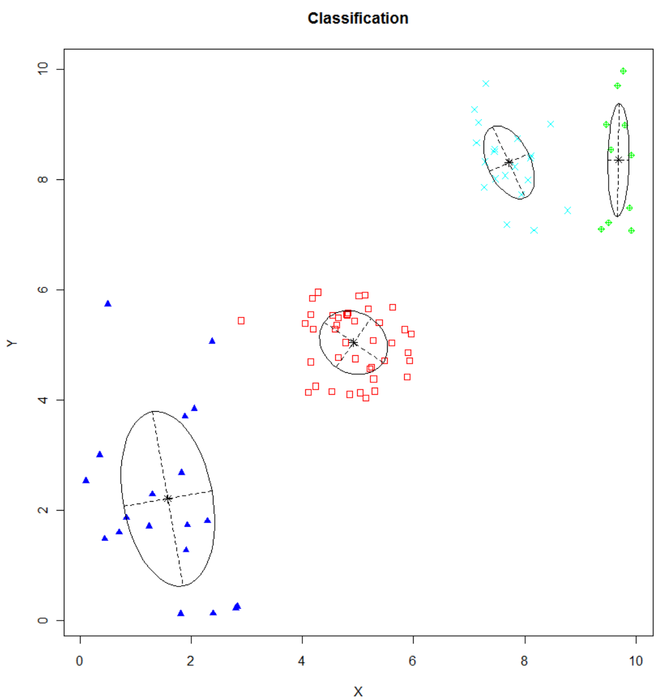


FIGURE 22.3: Graphical representation of how different covariance models allow GMMs to capture different cluster structures.

- See the sample script **`findClusters.r`**:




```
> p.Mclust$parameters

$pro  consist rate
[1] 0.2144090 0.4522577 0.1109484 0.2223849

$mean
      [,1]     [,2]     [,3]     [,4]
X 1.573950 4.919038 9.688369 7.715275
Y 2.208468 5.042728 8.355194 8.312360

$variance$sigma
, , 1
           X          Y
X  0.7124123 -0.3320267
Y -0.3320267  2.5172177   variance

, , 2
            X           Y
X  0.37565944 -0.05669909   co-variance
Y -0.05669909  0.33411790

, , 3
           X          Y
X 0.03605865 0.01643843
Y 0.01643843 1.05226865

, , 4
           X          Y
X  0.2100294 -0.1332837
Y -0.1332837  0.4367538
```

Regression: Metric Variable on Circular Variable

- Review the document **`PoorManFourier.pdf`**.
- See R-script **`SinCosRegression.R`**

Measuring Shape by Fourier Circles (not test relevant)

- See also See also **`DavisSpatialAnalysis.pdf`** pp 359-366.
- Shape analysis is done in morphometrics aiming at
  - statistically classifying outlines of different objects

- statistically describing variations of shapes of similar objects
- monitoring and explaining the change of shapes over time

- Davis provides a motivational explanation on the technique using the Fourier transformation of radius variations.
- An in-depth discussion on more robust Fourier methods for shapes is given by
    - Bonhomme, V, S Picq, C Gaucherel, and J Claude (2014). Nomocs: Outline Analysis Using R. *Journal of Statistical Software*, **56**:??-??, which introduces the R packet **Monocs**.
- The underlying idea is to translate the shape into a sequence of radii around an interior reference point.

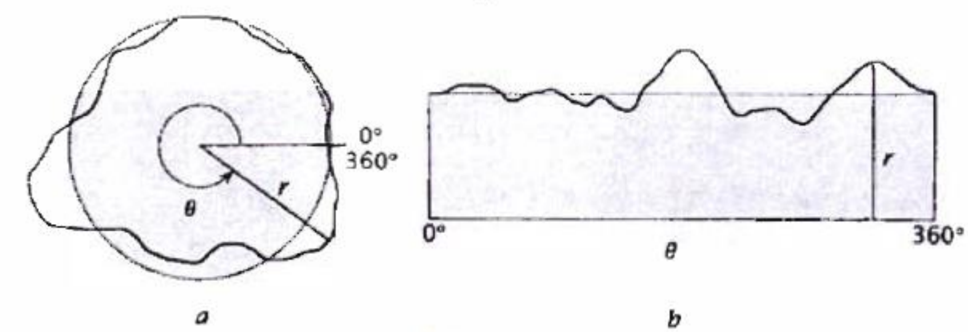


Figure 5–56. Equivalency between polar and Cartesian representations. (*a*) Grain outline shown in polar notation. (*b*) Polar coordinates plotted as r versus θ.

- The amplitude at a given radius dependents on the selected central origin.
- Drawbacks:
    - The amplitude at a given radius dependents on the selected central origin.
    - A radius must be measured in fixed angular increments, which may require interpolation among the existing shape points.
- The graph of the radii in the polar coordinates plot can be decomposed into waves of varying [a] wave length, [b] amplitudes and [c] phase shifts.
    - The phase shift and amplitudes can be estimate for waves of different length starting with one complete wave (the circle) in the interval.

- If a wave at a given length does not contribute to the polar coordinates graph its amplitude will be close to zero.
- Using the inverse Fourier transformed of calibrated amplitude and phase shift at given wave length the contributing circle can be recreated.
- Combining recreated waves with a relevant amplitude efficiently approximates the given shape with just a few Fourier coefficients rather than n data points.

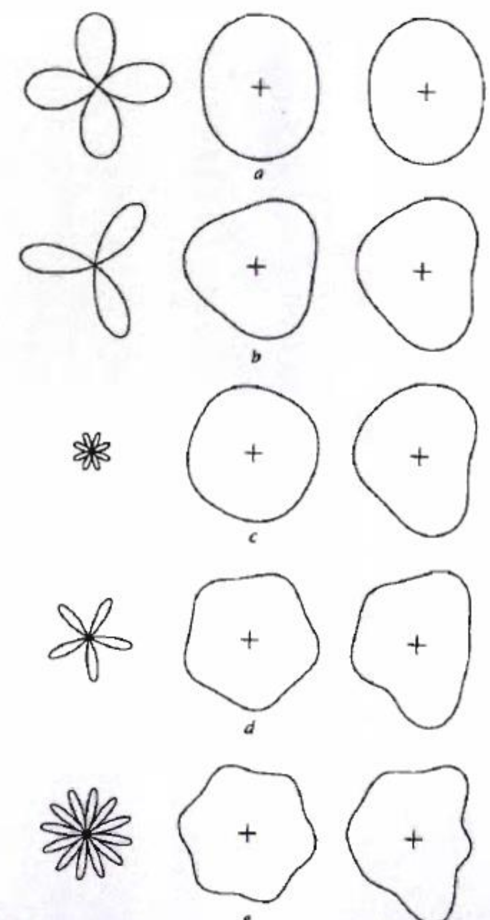


Figure 5–57. Reconstruction of the outline of a sand grain by polar Fourier series. (a) Plot of second harmonic (left), second harmonic plus circle corresponding to mean radius or zeroth harmonic (center), cumulative sum of harmonics (right). (b) Third harmonic. (c) Fourth harmonic. (d) Fifth harmonic. (e) Sixth harmonic.

- The power spectrum displays the standardized amplitude at given wave lengths. Waves with a low power could be dropped from recreating the underlying shape.

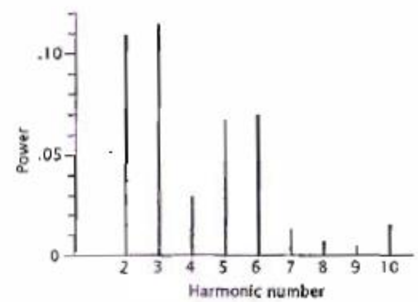


Figure 5–59. Power spectrum of digitized sand grain shown in Figure 5–56.

- This approach does not work for double-valued radii that may arise from concave shapes.
- The **Momocs** library does not model the radii but performs a discrete Fourier transformation on the set of x and y coordinates of the underlying shape points. It therefore is more robust because it does not dependent on [a] an arbitrary reference point, [b] equal spaced coordinates, and [c] the shape being convex.
- See the R script **FOURIERSHAPE.R** for an application of the elliptical discrete Fourier analysis.