Is the kidney data Gaussian
or
how can we model
the kidney data distribution ?

Inge Koch

University of Newcastle

J. S. Marron

Department of Statistics
University of North Carolina

for high dimensions d = 264

and low sample n = 36

do PCA to reduce data to 7 dimensions

analyse data in PC directions:

data could be normal - QQ-plots

some outliers ?

sphere or whiten PC data:

normalise with covariance matrix

this uncorrelates data

and makes all directions same length

Independent Component Analysis – ICA

decomposes data into independent

non-Gaussian directions

first direction: most non-Gaussian

has large skewness and kurtosis

second direction: 'more' Gaussian etc

ICA does not give unique directions

does not converge if ICA cannot find

non-Gaussian directions

compare first IC direction with

data simulated from Gaussian

calculate p-value

ICA finds 4 outliers

outliers not in a cluster

try: removing these outliers

ICA finds 3 new ones ...

ICA gives clear indication that data non-Gaussian

try: transform data -- use sphered data

power transformations of distance to origin

re-sphere and compare to Gaussian

power transform with

a = 0.55

works best

how Gaussian is this transformed data?

do ICA on it and compare to original data transformation has moved data

closer to Gaussian

but transformed data clearly non-Gaussian

how can we simulate this data?

calculate EDF of real data

want simulated data which envelops EDF

simulated Gaussian data does **not** cover

EDF of data

apply 'un-sphering' and power transform

to simulated Gaussian

compare to EDF of data

best for

b = 2

with b = 1/a

averages of such data very close to real data,

but power transformed simulated data

has large variability

so better to take averages of such data

can adequately model kidney data by

average of un-sphered and

power transformed Gaussians