Profiling Doctors' Performance

Min Zhang

May 1, 2002

OUTLINE

I. Motivation

II. Principal Component Analysis

III. Parametric Model

   III.1 Two-Part Model

   III.2 Bayesian Analysis

   III.3 Performance Index

   III.4 Preliminary Results

IV. Discussion

I. Motivation

- Goal

Evaluate the doctors' performance by using cost data of patients within their doctors.

- Issues

1. How to statistically model the cost data?
2. What are concerned in medical care system about doctors' performance?
3. How to rank and identify doctors?

II. Principal Component Analysis

- Several modeling approaches not fit the data well...

- Exploratory Data Analysis

1. Is there any pattern existing in the data?
2. Is it possible to classify the doctors statistically?
3. Are there any outliers?

II. Principal Component Analysis

- Description of Raw Data

1. There are 58686 patients clustered in 252 doctors.
2. Different number of patients are observed in each doctor: the range is from 1 to 1501.
3. 29 categorical costs are observed for each patient.

For example, antibacterial, cardiovascular, dermatological agents, analgesics etc.

II. Principal Component Analysis (Continued)

- Let Z_{ijk} be k-th categorical cost of patient j of doctor i.

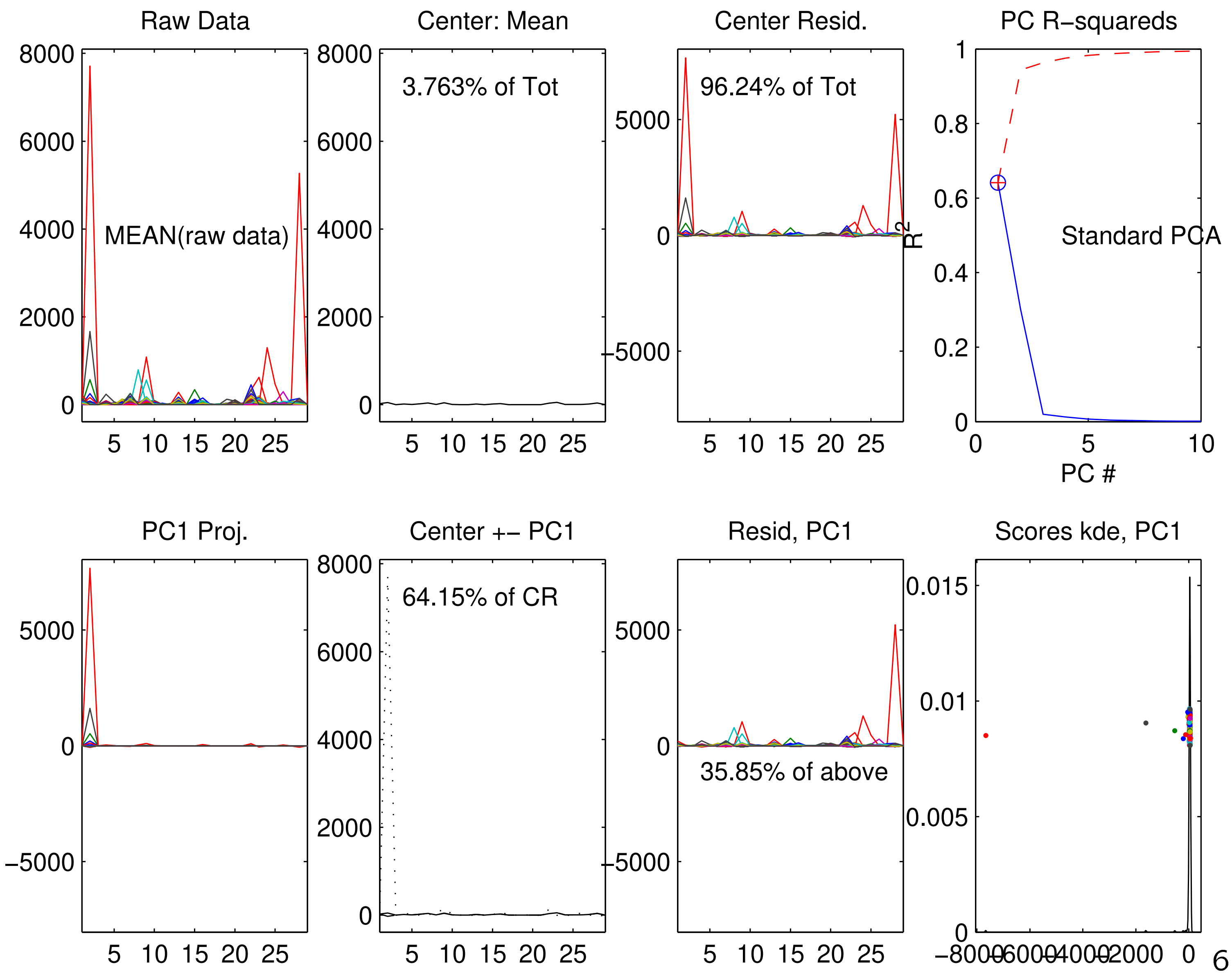
$$\bar{Z}_{i\cdot k} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ijk}$$

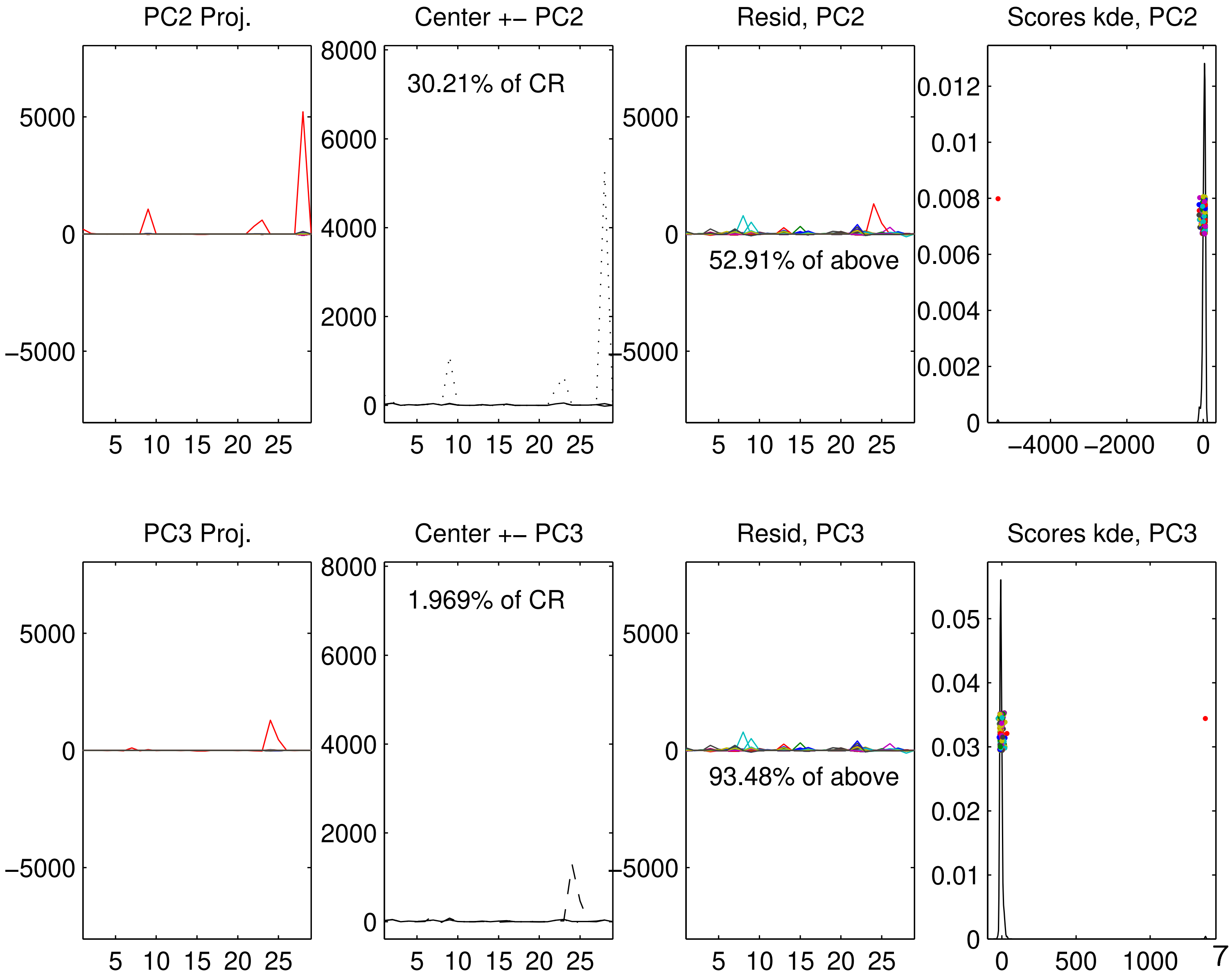
$$\hat{V}_{i\cdot k} = \left(\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Z_{ijk} - \bar{Z}_{i\cdot k})^2 \right)^{1/2}$$

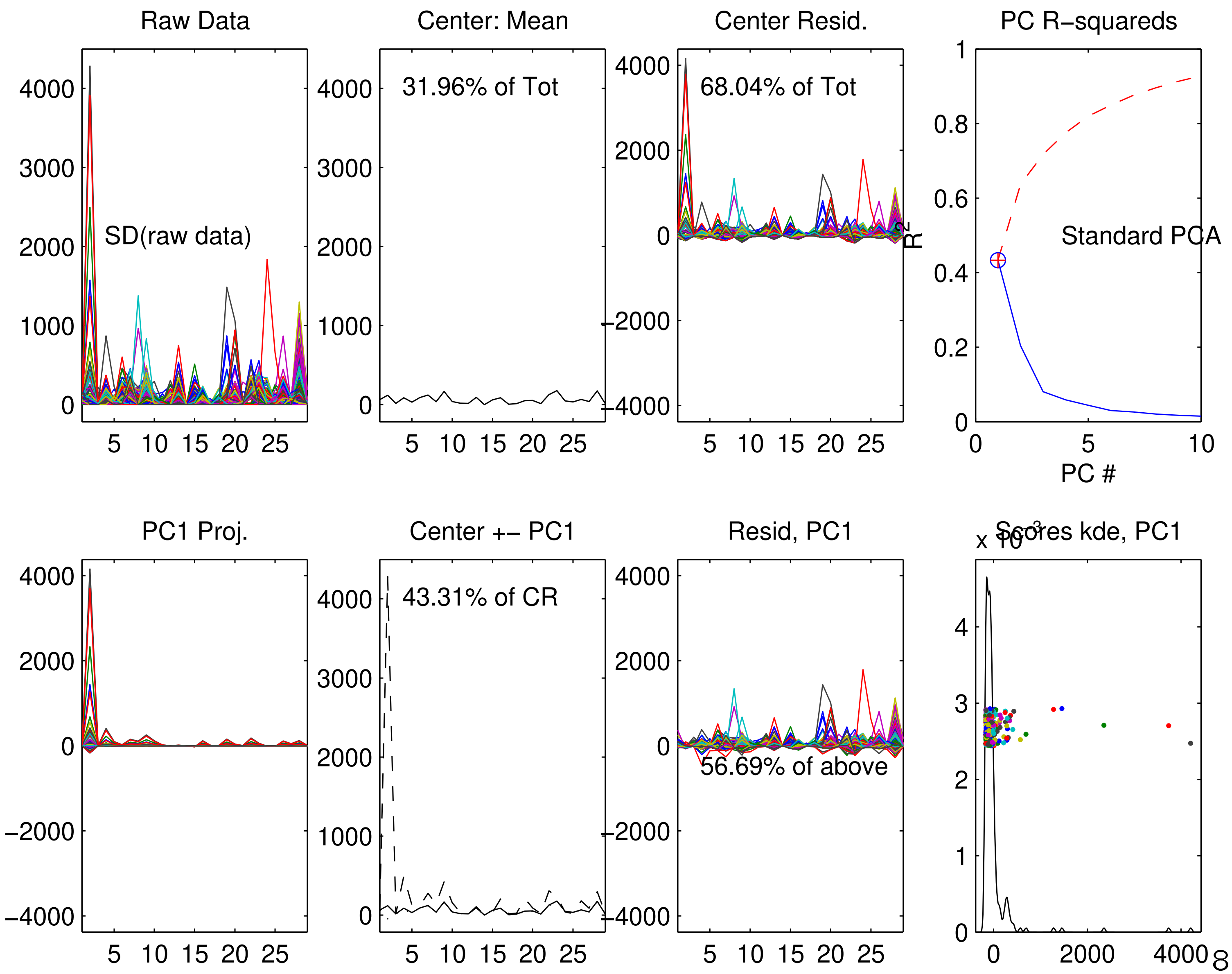
- Data Summarization Matrices:

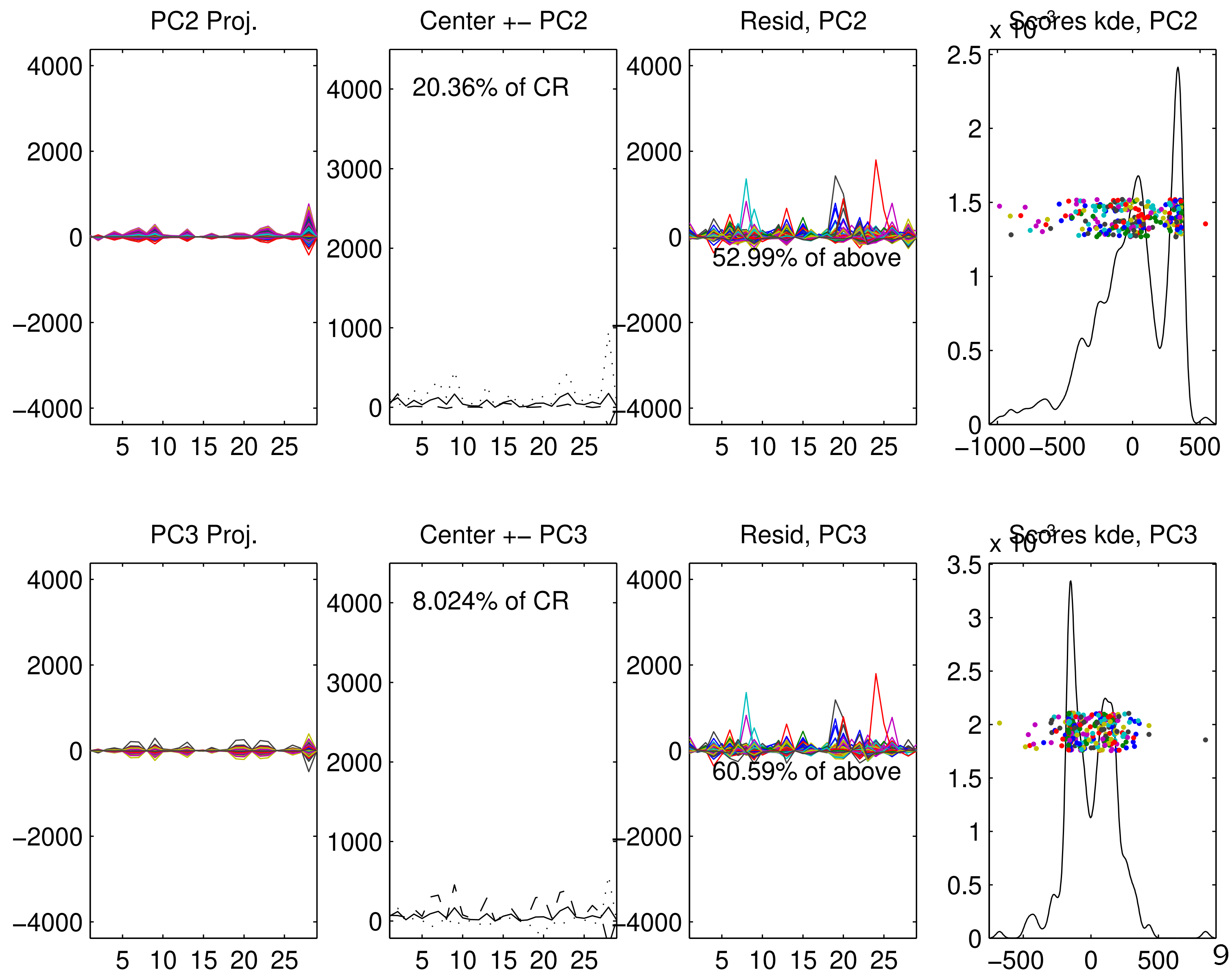
<u>Mean Matrix</u> <u>SD Matrix</u>

$$\begin{pmatrix} \bar{Z}_{1\cdot 1} & \cdots & \bar{Z}_{1\cdot K} \\ \bar{Z}_{2\cdot 1} & \cdots & \bar{Z}_{2\cdot K} \\ \vdots & \vdots & \vdots \\ \bar{Z}_{M\cdot 1} & \cdots & \bar{Z}_{M\cdot K} \end{pmatrix}_{M\times K} \quad \begin{pmatrix} \hat{V}_{1\cdot 1} & \cdots & \hat{V}_{1\cdot K} \\ \hat{V}_{2\cdot 1} & \cdots & \hat{V}_{2\cdot K} \\ \vdots & \vdots & \vdots \\ \hat{V}_{M\cdot 1} & \cdots & \hat{V}_{M\cdot K} \end{pmatrix}_{M\times K}$$


Raw Data
MEAN(raw data)
Center: Mean
3.763% of Tot
Center Resid.
96.24% of Tot
PC R−squareds
Standard PCA
R^2
PC #
PC1 Proj.
Center +− PC1
64.15% of CR
Resid, PC1
35.85% of above
Scores kde, PC1


PC2 Proj.
Center +− PC2
30.21% of CR
Resid, PC2
52.91% of above
Scores kde, PC2
PC3 Proj.
Center +− PC3
1.969% of CR
Resid, PC3
93.48% of above
Scores kde, PC3


Raw Data
SD(raw data)
Center: Mean
31.96% of Tot
Center Resid.
68.04% of Tot
PC R−squareds
Standard PCA
R^2
PC #
PC1 Proj.
Center +− PC1
43.31% of CR
Resid, PC1
56.69% of above
Scores kde, PC1
x 10^{-3}

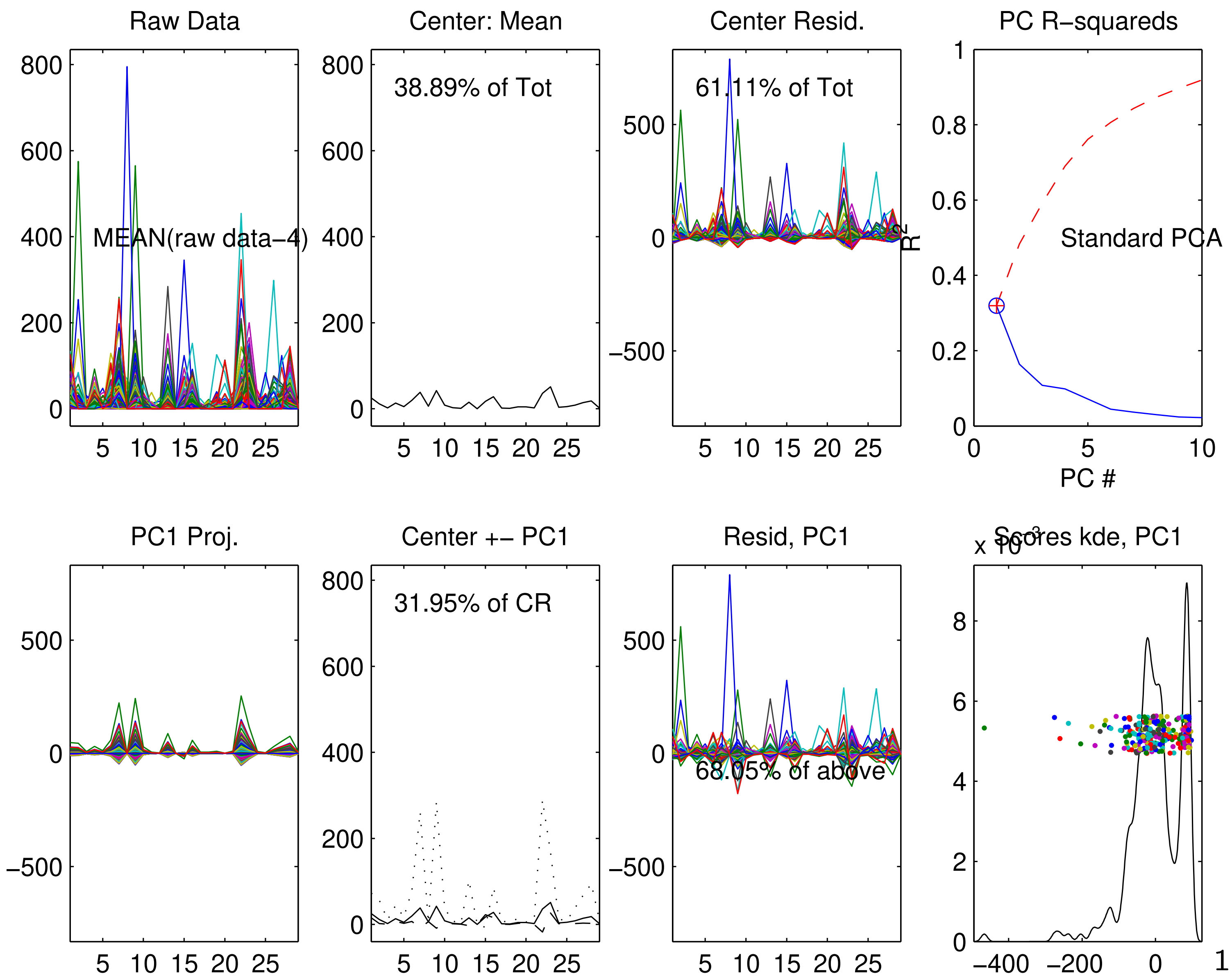

PC2 Proj.
Center +− PC2
20.36% of CR
Resid, PC2
52.99% of above
Scores kde, PC2
x 10^-3
PC3 Proj.
Center +− PC3
8.024% of CR
Resid, PC3
60.59% of above
Scores kde, PC3
x 10^-3

II. Principal Component Analysis (Continued)

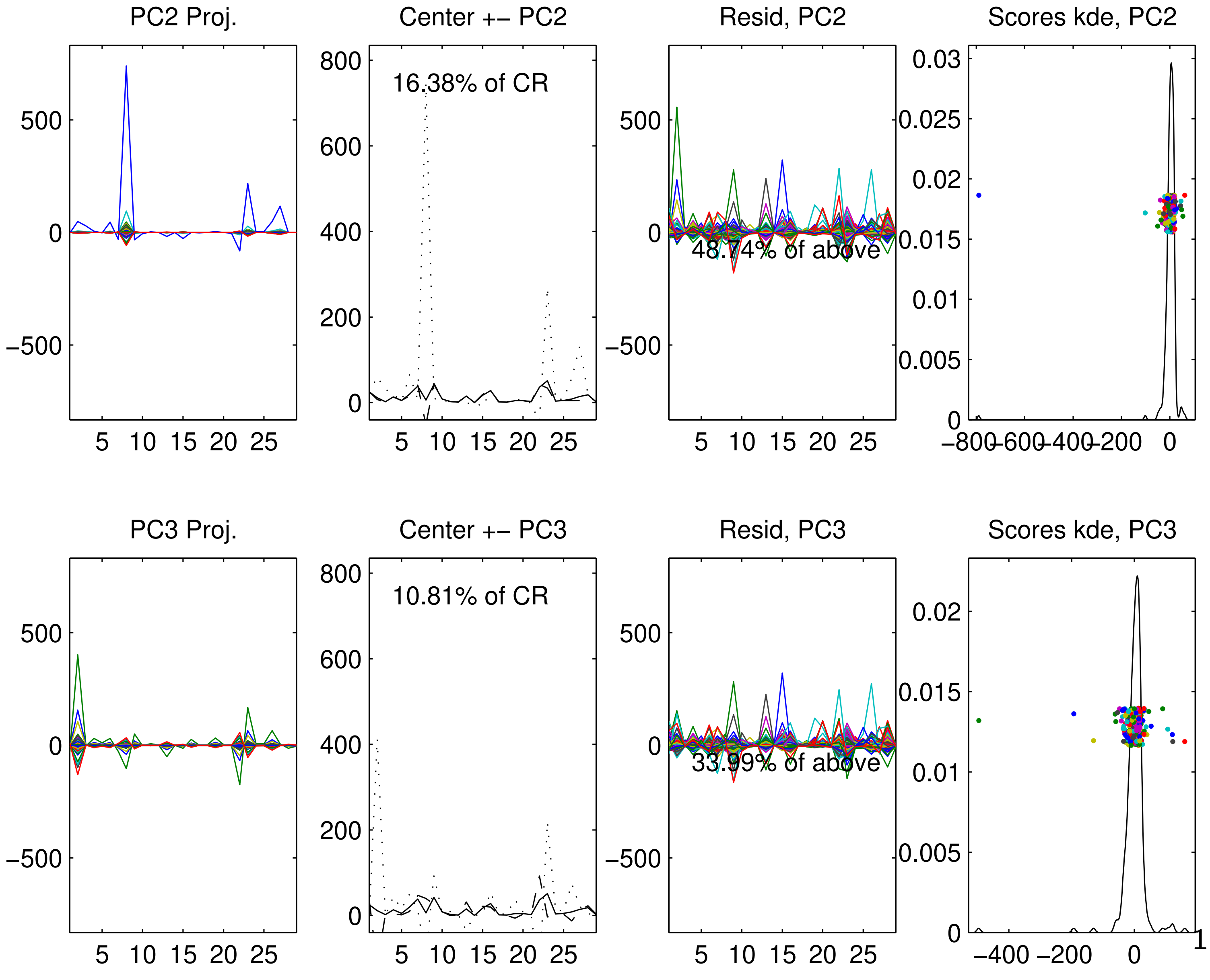
• Principal components of MEAN and SD are highly driven by outliers.

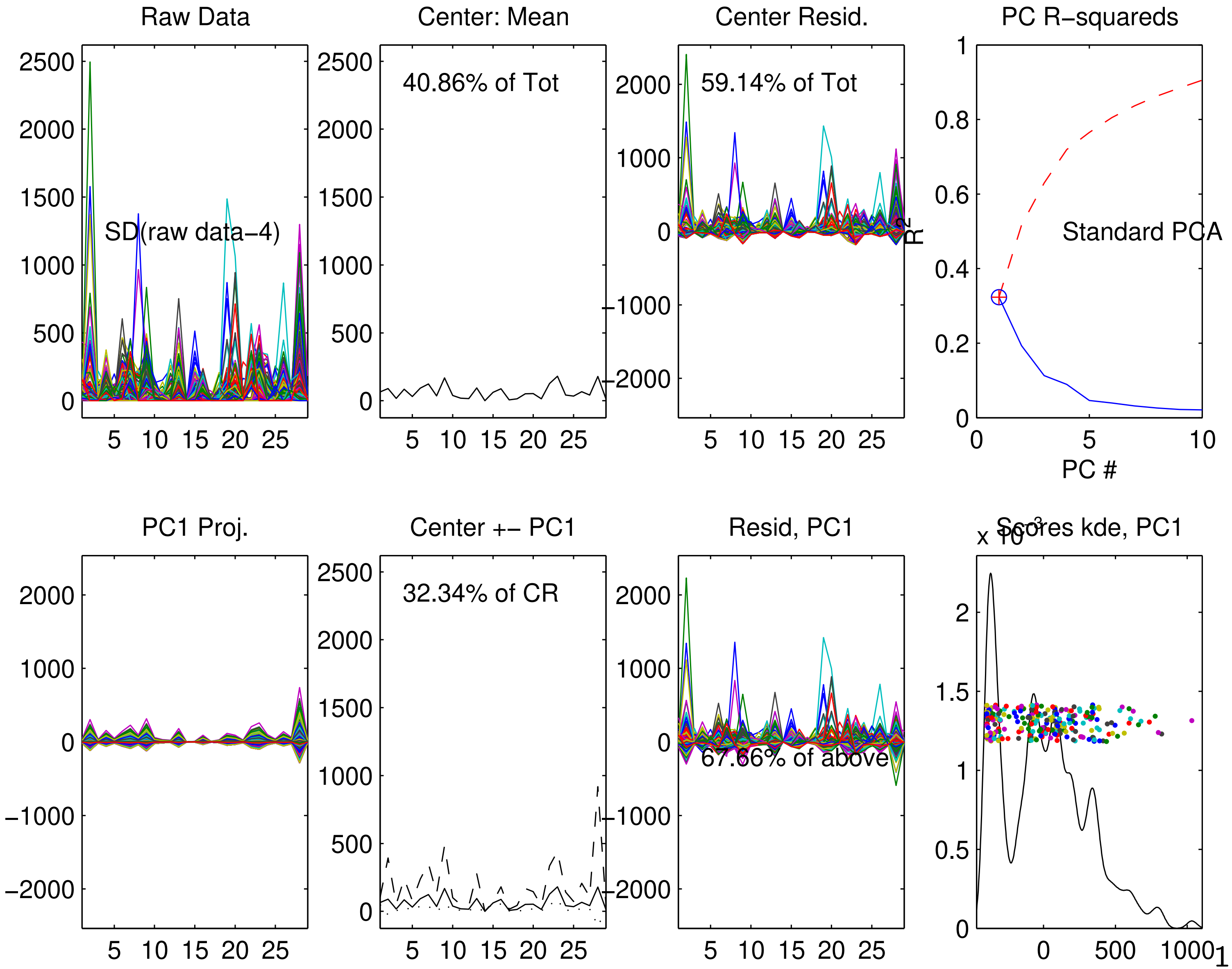
• The outliers are doctors with very few patients.
(1) 1^{st} doctor with 6 patients where 5 patients spent \$8258 to \$10938 on category 2.
(2) 2^{nd} doctor with 1 patient who spent over \$5200 on category 28.
(3)Similar situations for the other 2 outliers.

• Two common approaches to dealing with outliers:
Approach I: exclude the outliers from analysis.

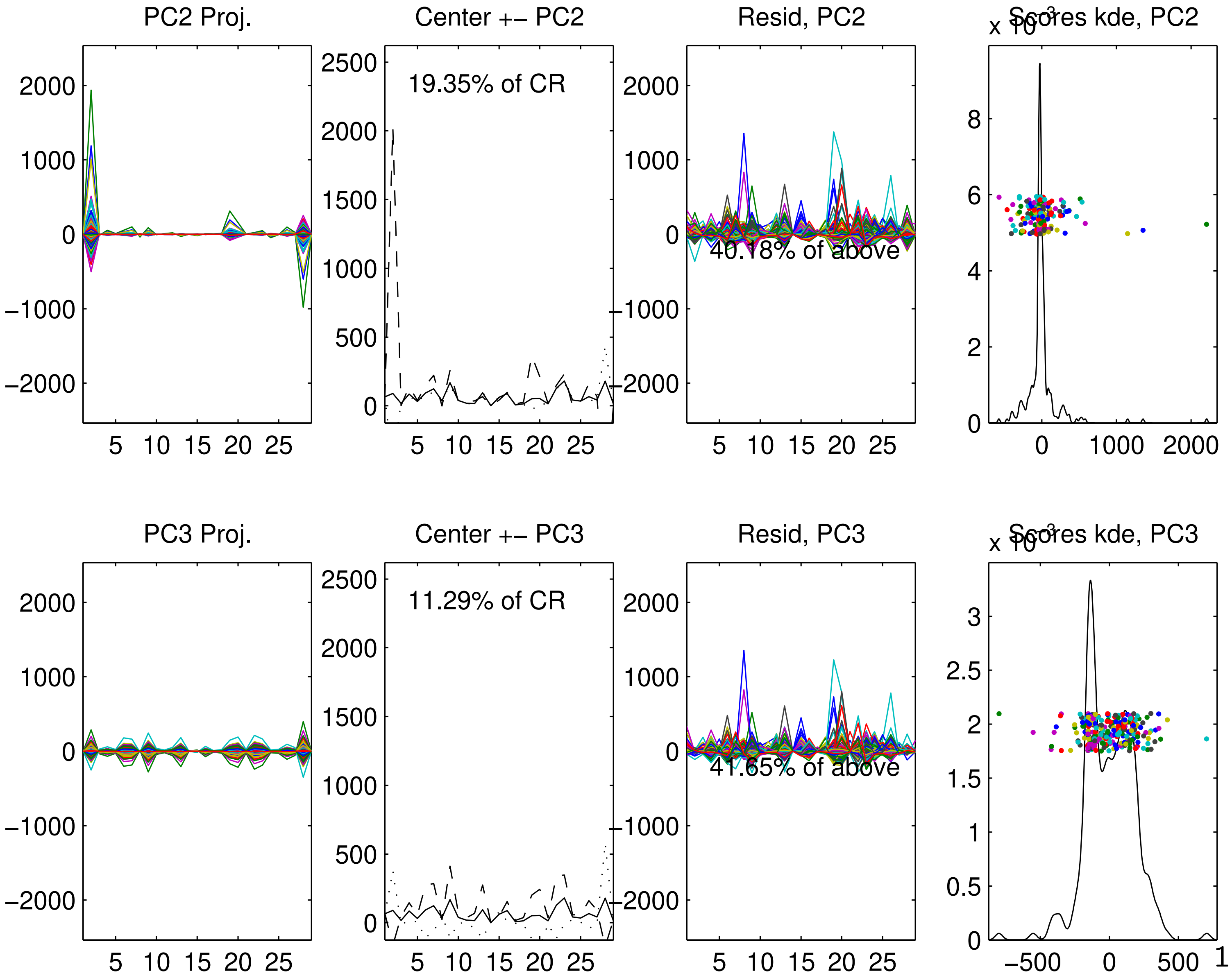


Raw Data
MEAN(raw data−4)
Center: Mean
38.89% of Tot
Center Resid.
61.11% of Tot
PC R−squareds
Standard PCA
R^2
PC #
PC1 Proj.
Center +− PC1
31.95% of CR
Resid, PC1
68.05% of above
Scores kde, PC1
x 10^{-3}


PC2 Proj.
Center +− PC2
16.38% of CR
Resid, PC2
48.74% of above
Scores kde, PC2
PC3 Proj.
Center +− PC3
10.81% of CR
Resid, PC3
33.99% of above
Scores kde, PC3



Raw Data
SD(raw data−4)
Center: Mean
40.86% of Tot
Center Resid.
59.14% of Tot
PC R−squareds
Standard PCA
R2
PC #
PC1 Proj.
Center +− PC1
32.34% of CR
Resid, PC1
67.66% of above
Scores kde, PC1
x 10-3

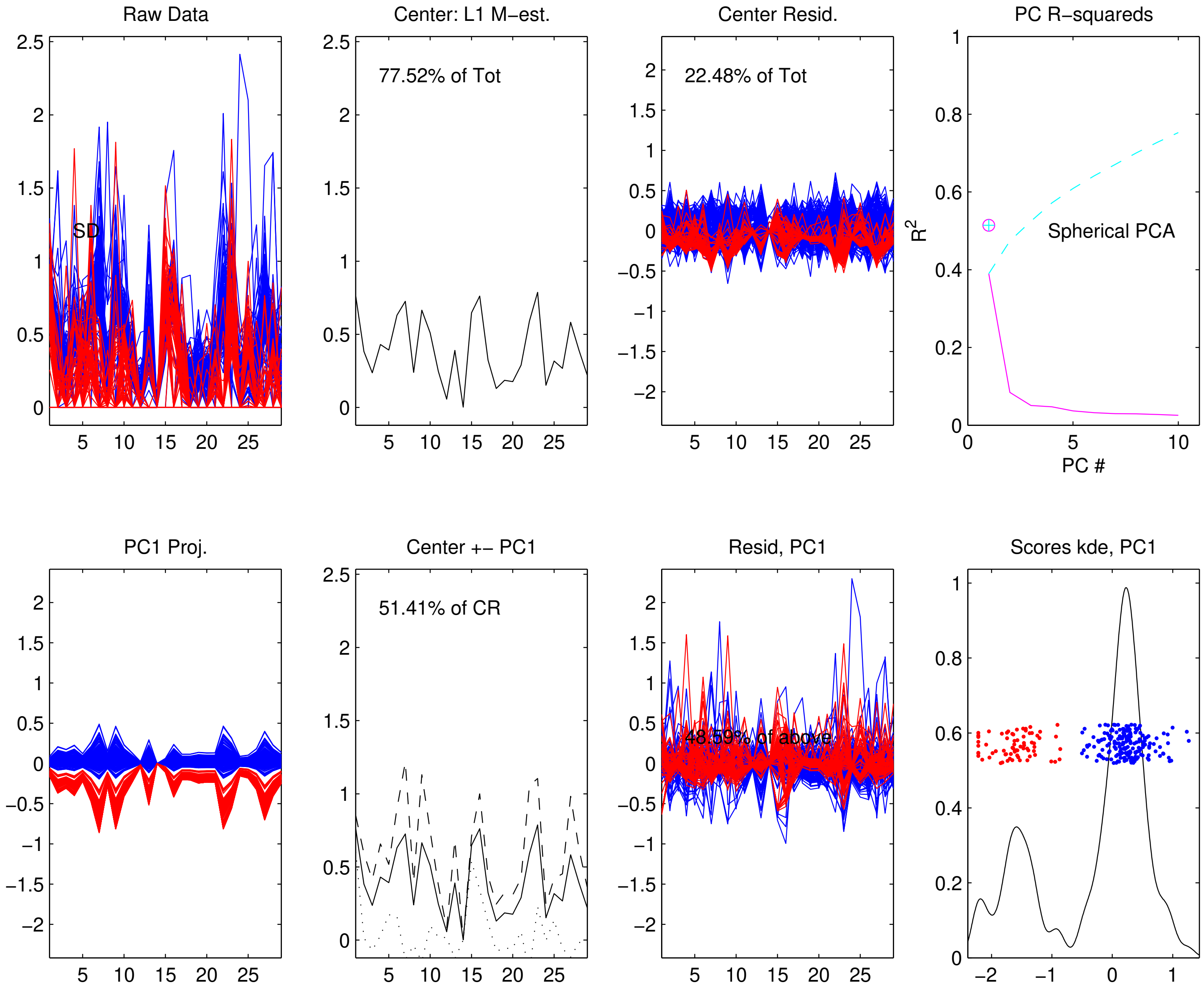

PC2 Proj.
Center +− PC2
19.35% of CR
Resid, PC2
40.18% of above
Scores kde, PC2
PC3 Proj.
Center +− PC3
11.29% of CR
Resid, PC3
41.65% of above
Scores kde, PC3

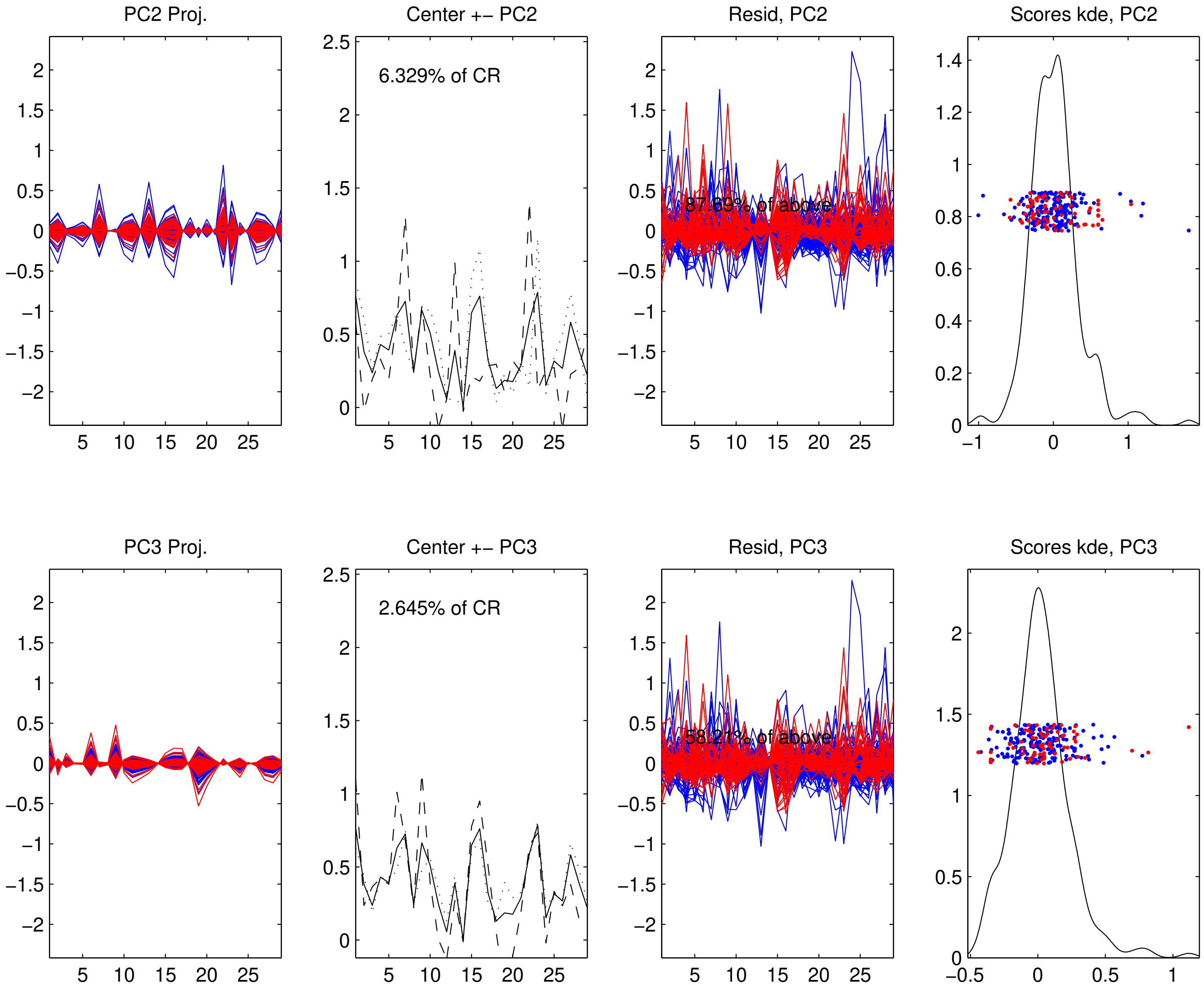
II. Principal Component Analysis (Continued)

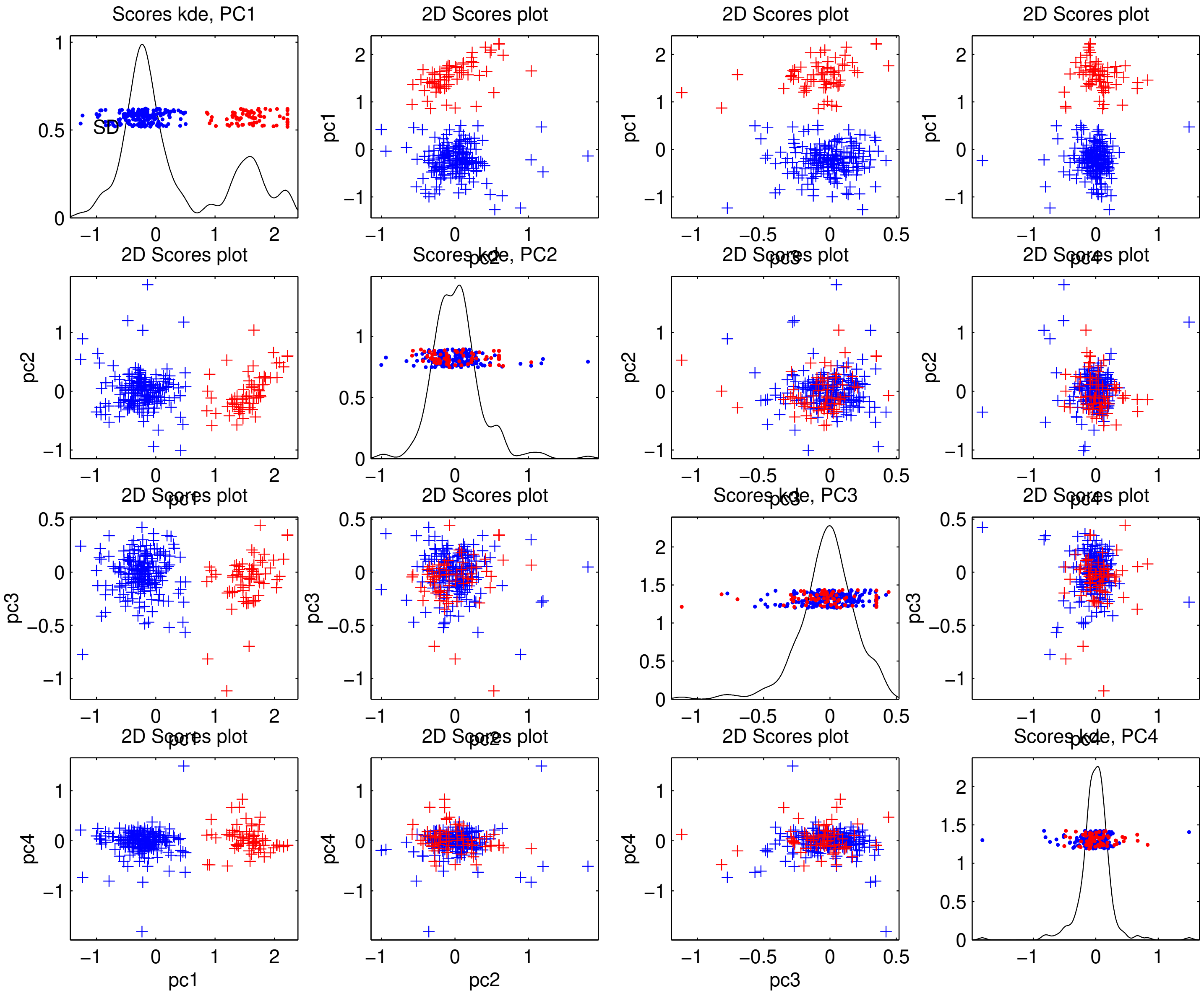
- New outliers show up after old outliers are deleted.

- Try approach II to deal with outliers:
Robust Statistical method: Spherical PCA.

- Take log-transformation of the raw data and get new data summarization matrices: MEAN matrix and SD matrix.
ie. $Z^*_{ijk} = log_{10}(Z_{ijk} + 1)$

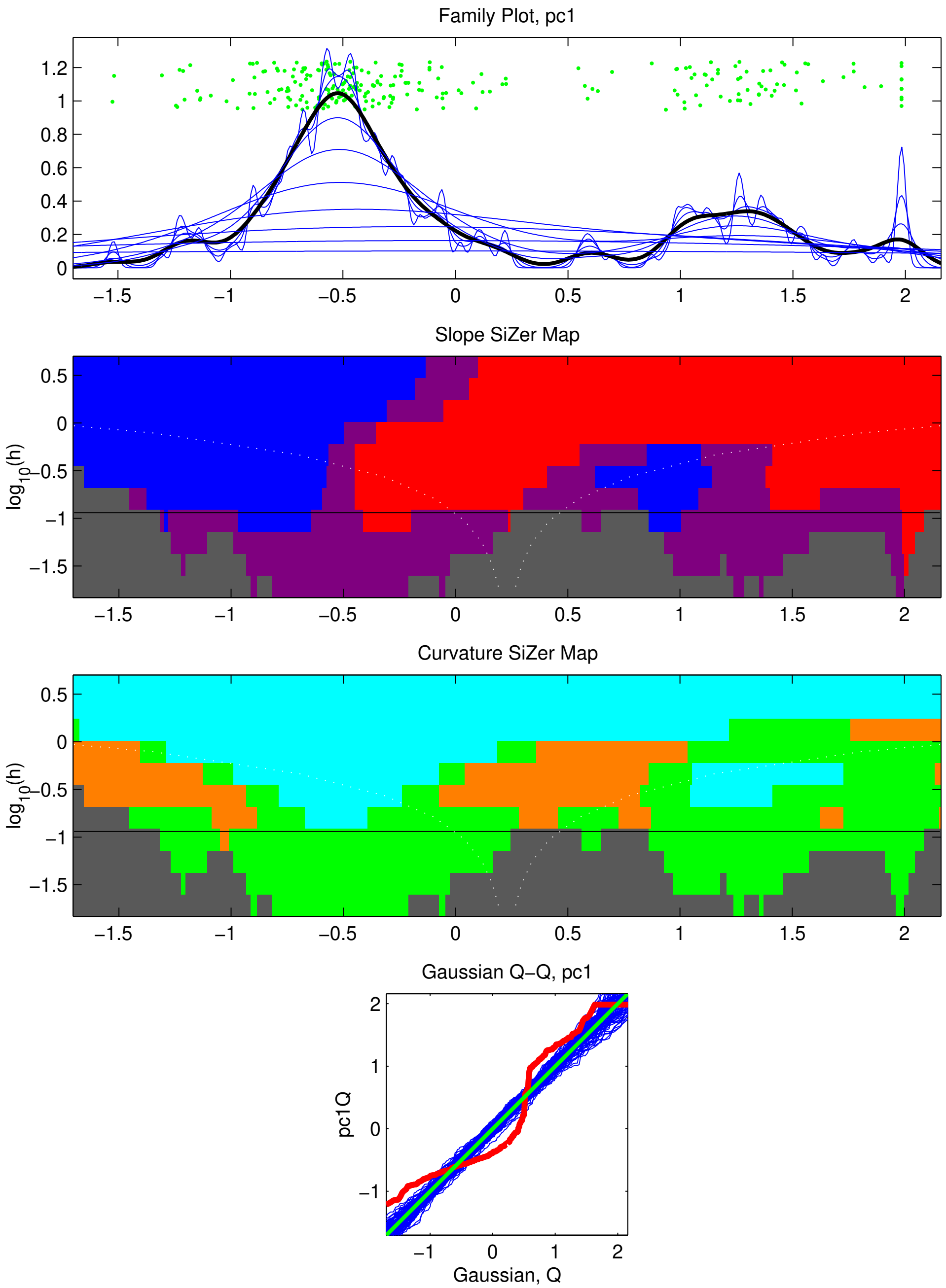

Raw Data
SD
Center: L1 M−est.
77.52% of Tot
Center Resid.
22.48% of Tot
PC R−squareds
Spherical PCA
R^2
PC #
PC1 Proj.
Center +− PC1
51.41% of CR
Resid, PC1
48.59% of above
Scores kde, PC1


PC2 Proj.
Center +– PC2
6.329% of CR
Resid, PC2
87.69% of above
Scores kde, PC2
PC3 Proj.
Center +– PC3
2.645% of CR
Resid, PC3
58.21% of above
Scores kde, PC3


Scores kde, PC1
2D Scores plot
2D Scores plot
2D Scores plot
SD
pc1
pc1
pc1
2D Scores plot
Scores kde, PC2
2D Scores plot
2D Scores plot
pc2
pc2
pc2
2D Scores plot
2D Scores plot
Scores kde, PC3
2D Scores plot
pc3
pc3
pc3
2D Scores plot
2D Scores plot
2D Scores plot
Scores kde, PC4
pc4
pc4
pc4
pc1
pc2
pc3

II. Principal Component Analysis (Continued)

- It seems that there are two clusters in population.
- 2-d plot gives a more obvious visualization.
- Statistical Question: Are there really two groups?


Family Plot, pc1
Slope SiZer Map
log10(h)
Curvature SiZer Map
log10(h)
Gaussian Q−Q, pc1
pc1Q
Gaussian, Q

II. Principal Component Analysis (Continued)

- Result

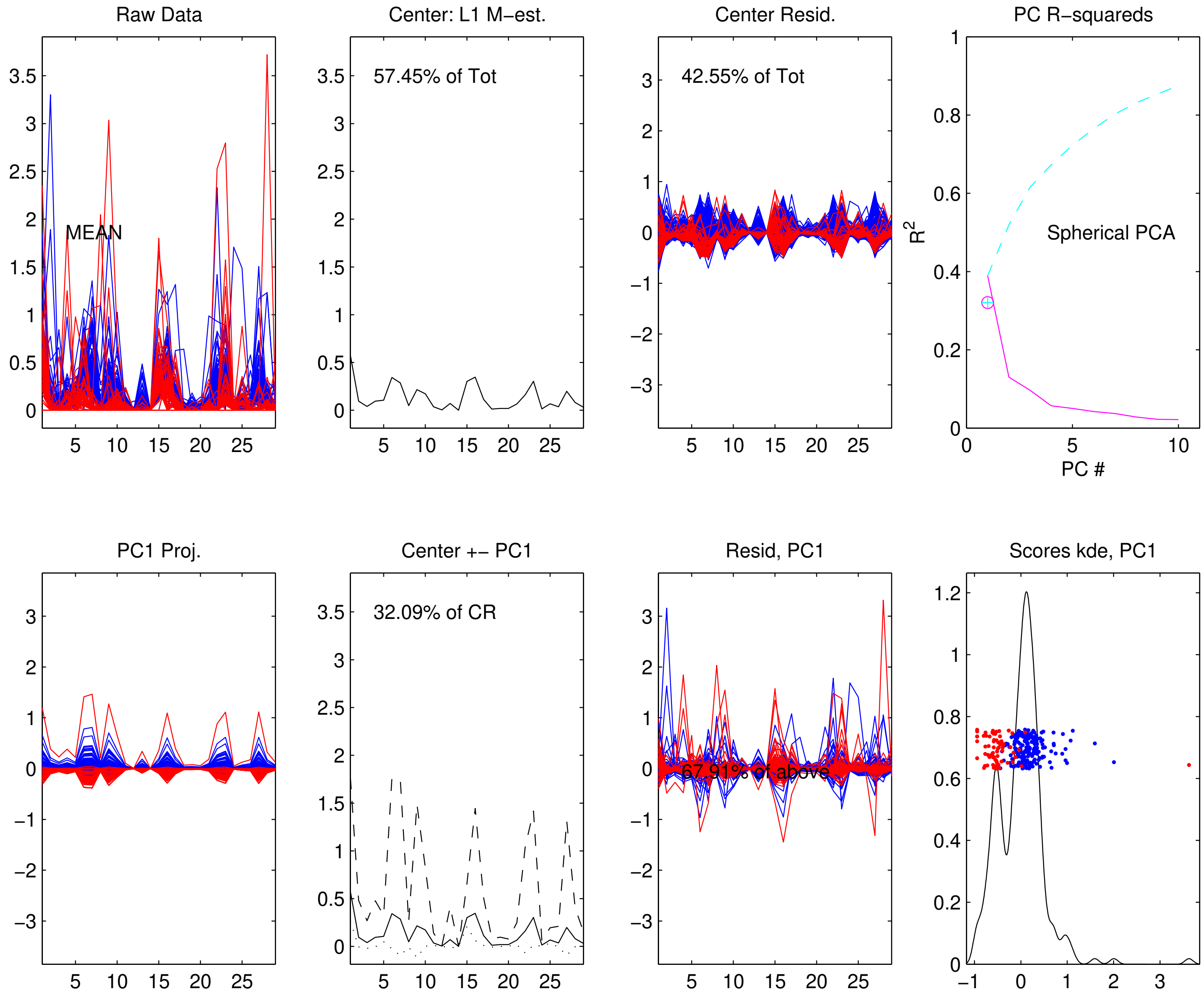
SiZer analysis show the two groups are significantly different.

- Next question

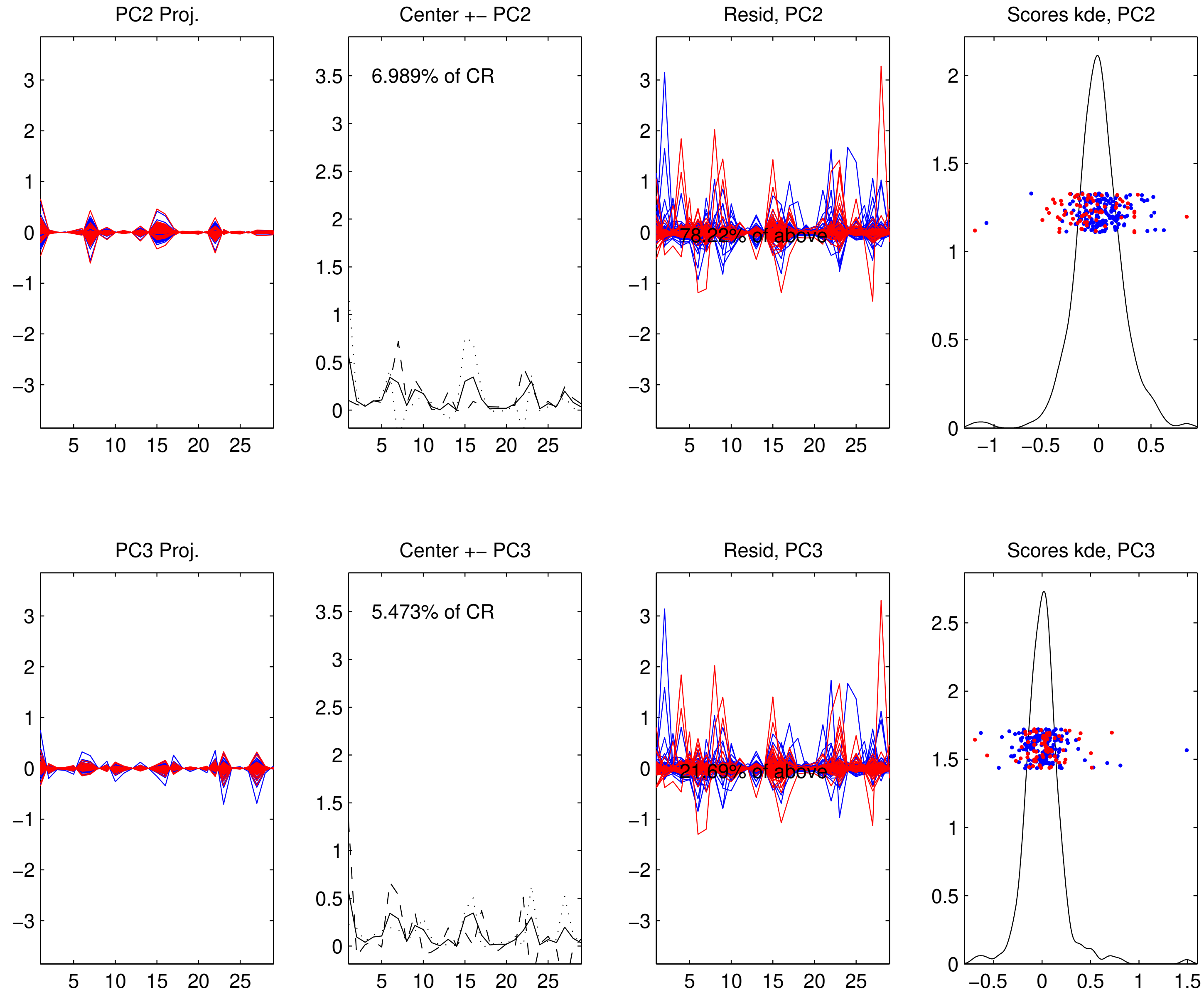
How the two clusters distributed in MEAN matrix analysis?

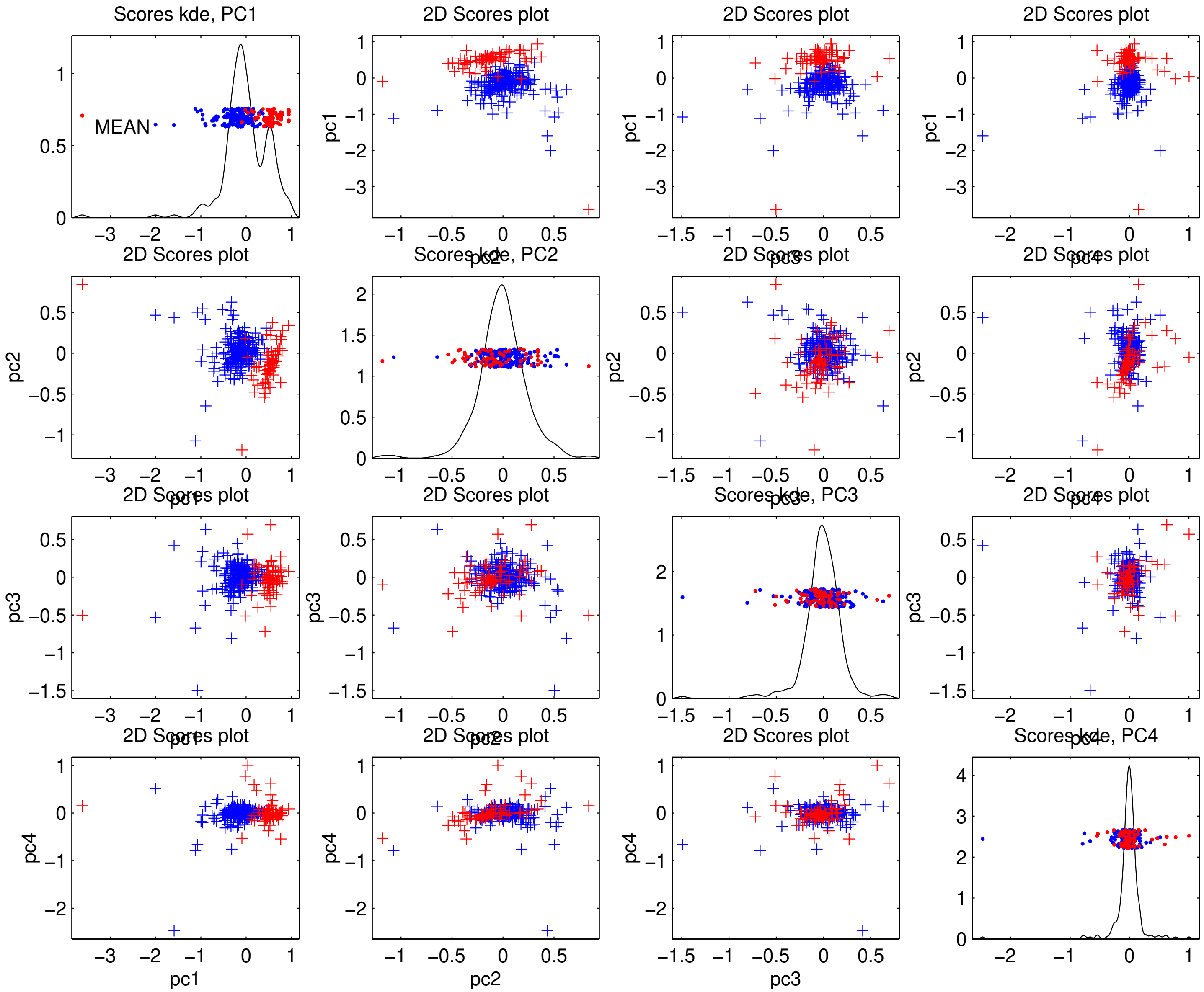
- Idea

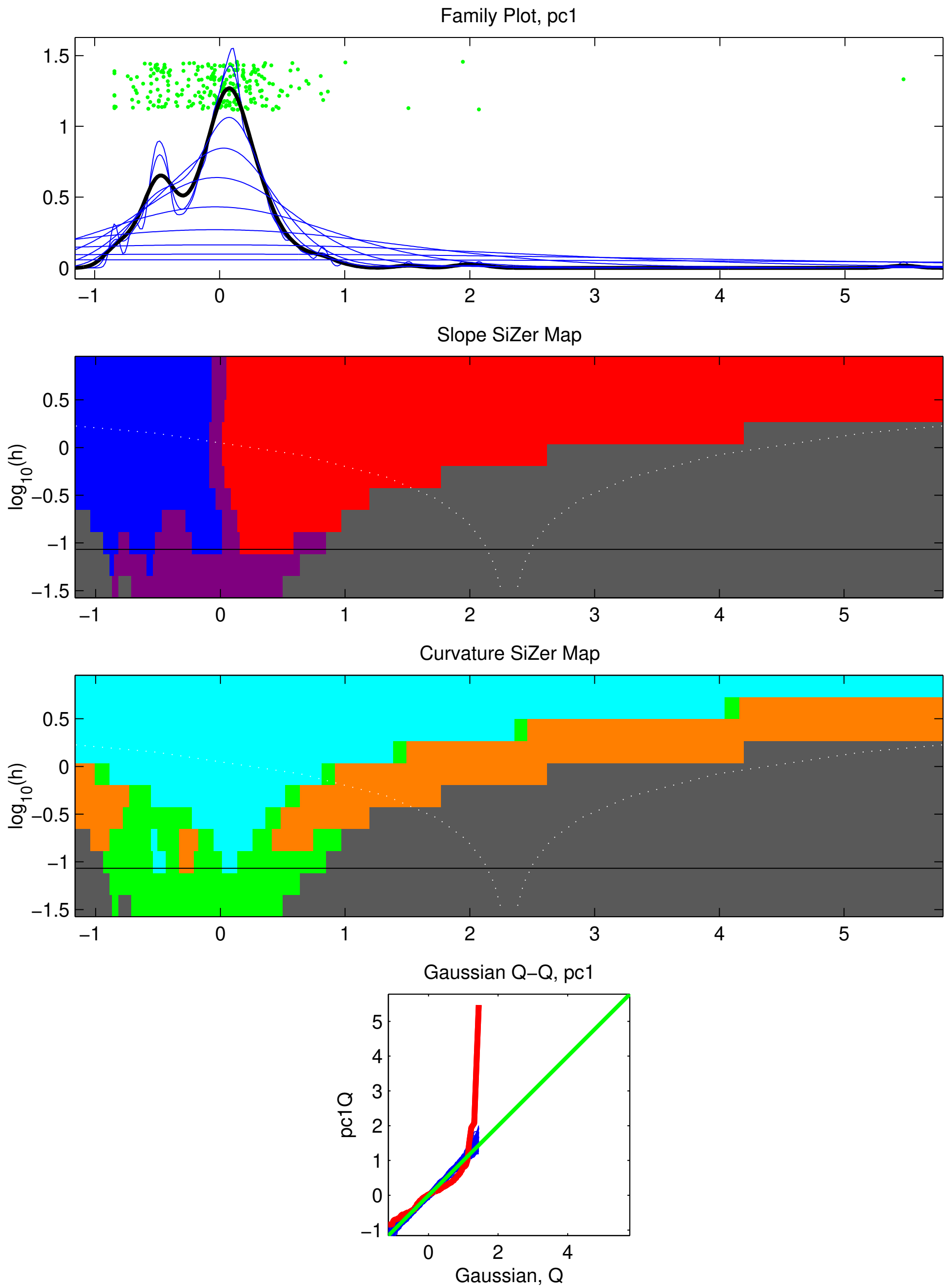
Using the same color matrix in MEAN data analysis.



Raw Data
MEAN
Center: L1 M−est.
57.45% of Tot
Center Resid.
42.55% of Tot
PC R−squareds
Spherical PCA
R^2
PC #
PC1 Proj.
Center +− PC1
32.09% of CR
Resid, PC1
67.91% of above
Scores kde, PC1


PC2 Proj.
Center +− PC2
6.989% of CR
Resid, PC2
78.22% of above
Scores kde, PC2
PC3 Proj.
Center +− PC3
5.473% of CR
Resid, PC3
21.69% of above
Scores kde, PC3


Scores kde, PC1
MEAN
2D Scores plot
Scores kde, PC2
Scores kde, PC3
Scores kde, PC4
pc1
pc2
pc3
pc4

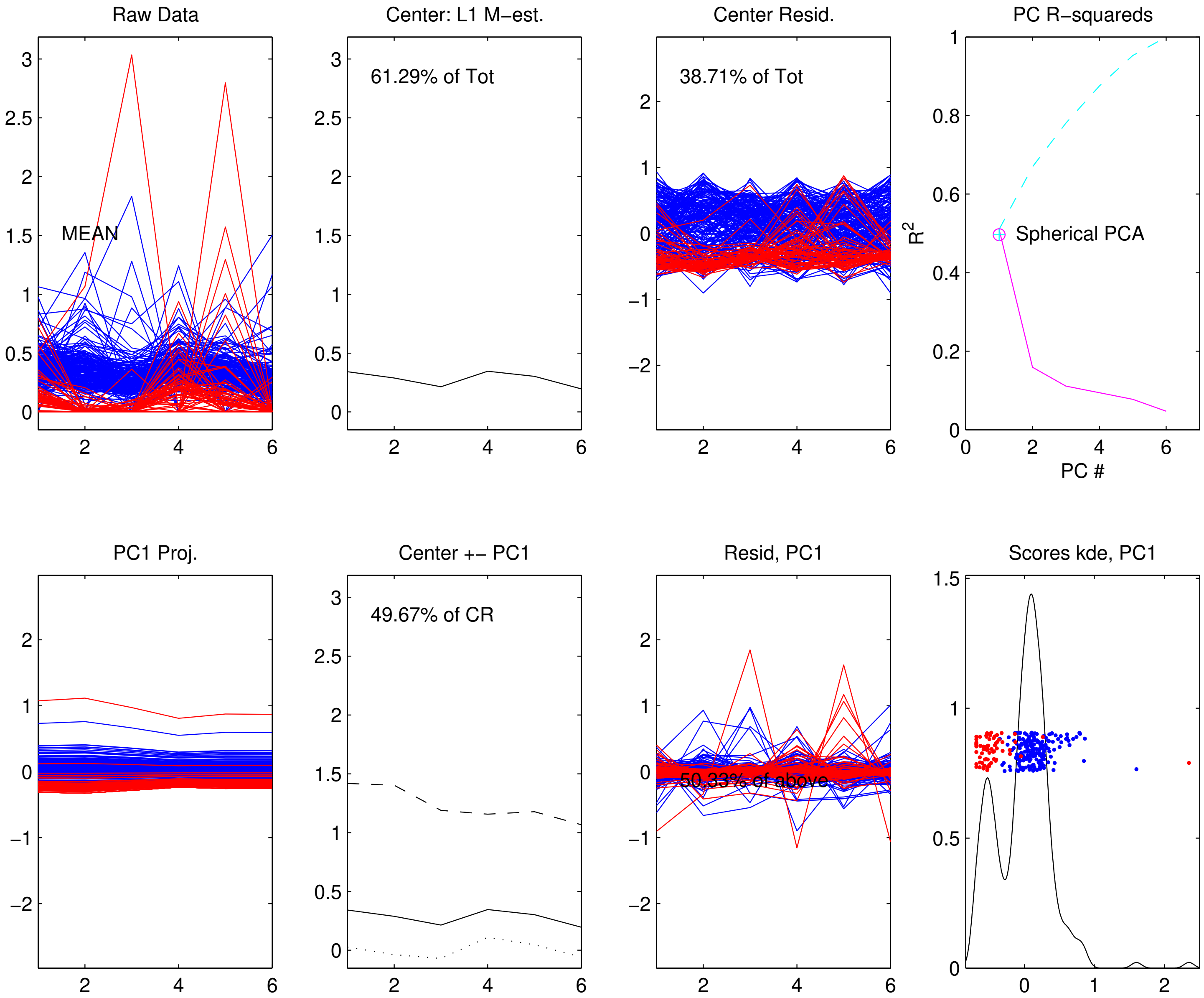

Family Plot, pc1
Slope SiZer Map
log10(h)
Curvature SiZer Map
log10(h)
Gaussian Q–Q, pc1
pc1Q
Gaussian, Q

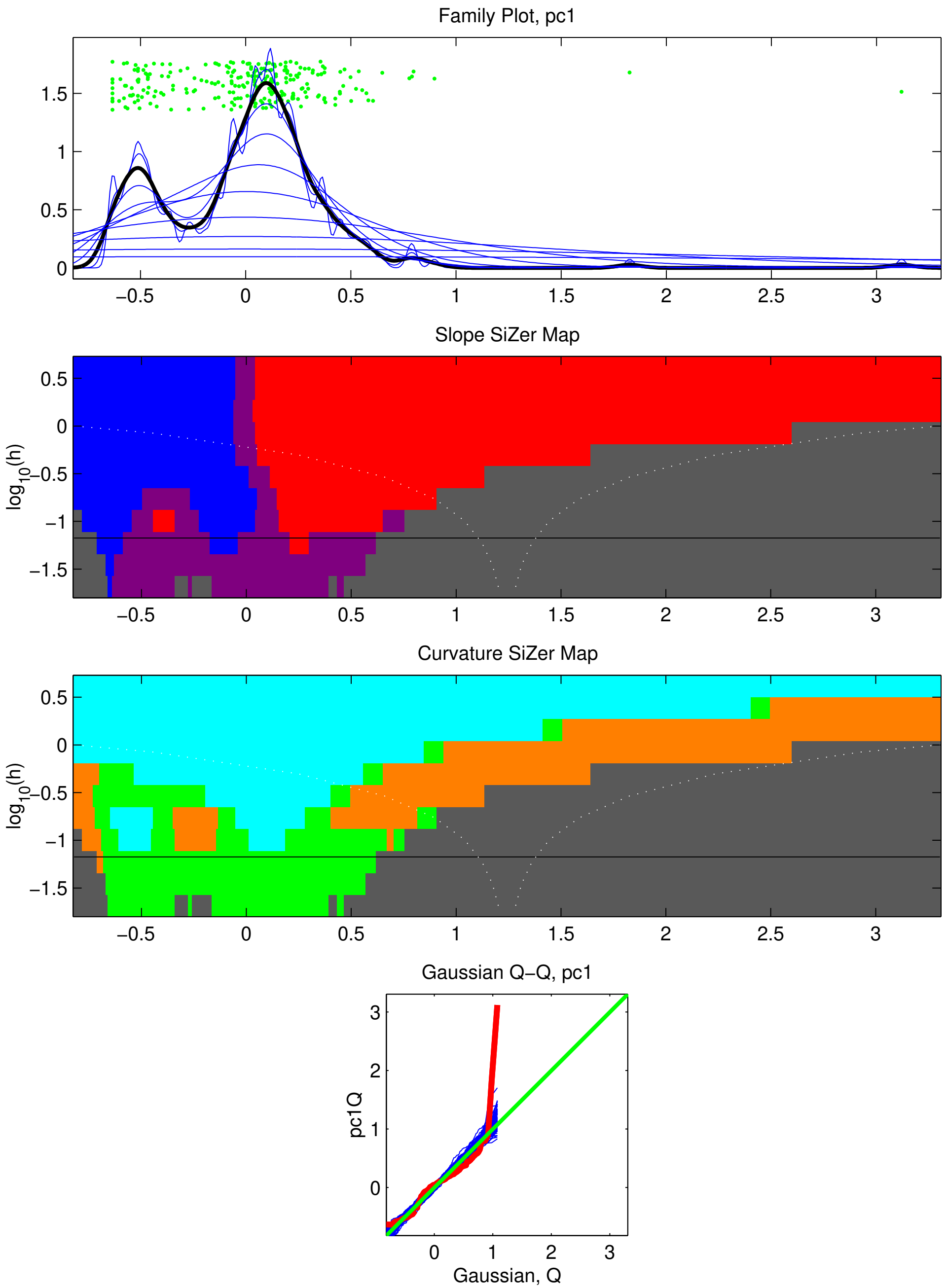
II. Principal Component Analysis (Continued)

- Significance in curvature map but not in slope SiZer map.

Reference:
http://www.unc.edu/depts/statistics/
postscript/papers/marron/SiCur/SiCon1.pdf

- Most difference observed in 6 categories.


Raw Data
MEAN
Center: L1 M−est.
61.29% of Tot
Center Resid.
38.71% of Tot
PC R−squareds
Spherical PCA
R^2
PC #
PC1 Proj.
Center +− PC1
49.67% of CR
Resid, PC1
50.33% of above
Scores kde, PC1



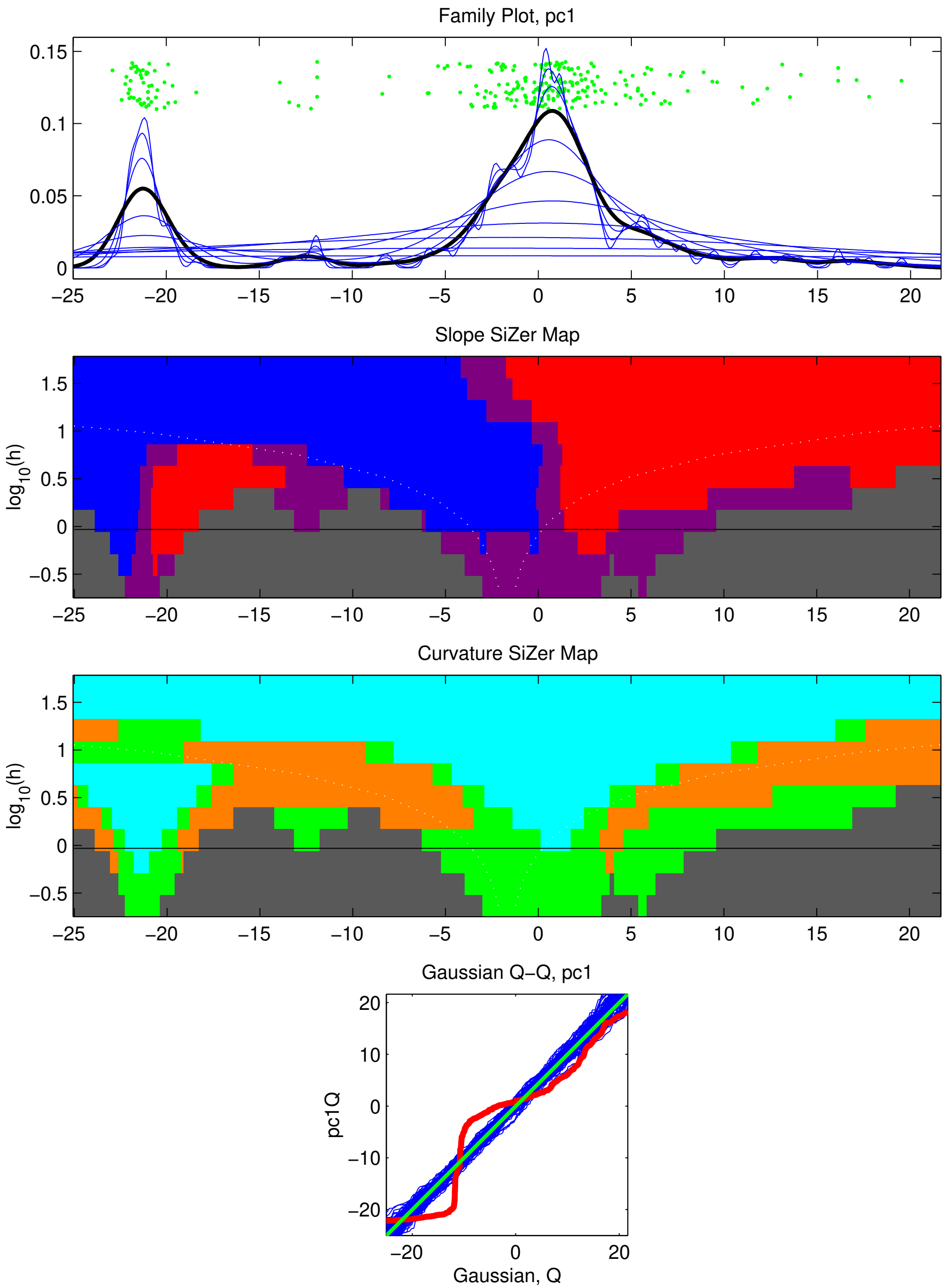
Family Plot, pc1
Slope SiZer Map
log10(h)
Curvature SiZer Map
log10(h)
Gaussian Q−Q, pc1
pc1Q
Gaussian, Q

II. Principal Component Analysis (Continued)

- The analysis is sharpened using 6 categories.

- Why there are two subpopulations?

(1) Empirically checking suggests that one cluster is of pediatricians and another is of doctors with adult patients.
(2) Further confirmed by SiZer analysis on MEAN(age)data.


Family Plot, pc1
Slope SiZer Map
$\log_{10}(h)$
Curvature SiZer Map
$\log_{10}(h)$
Gaussian Q–Q, pc1
pc1Q
Gaussian, Q

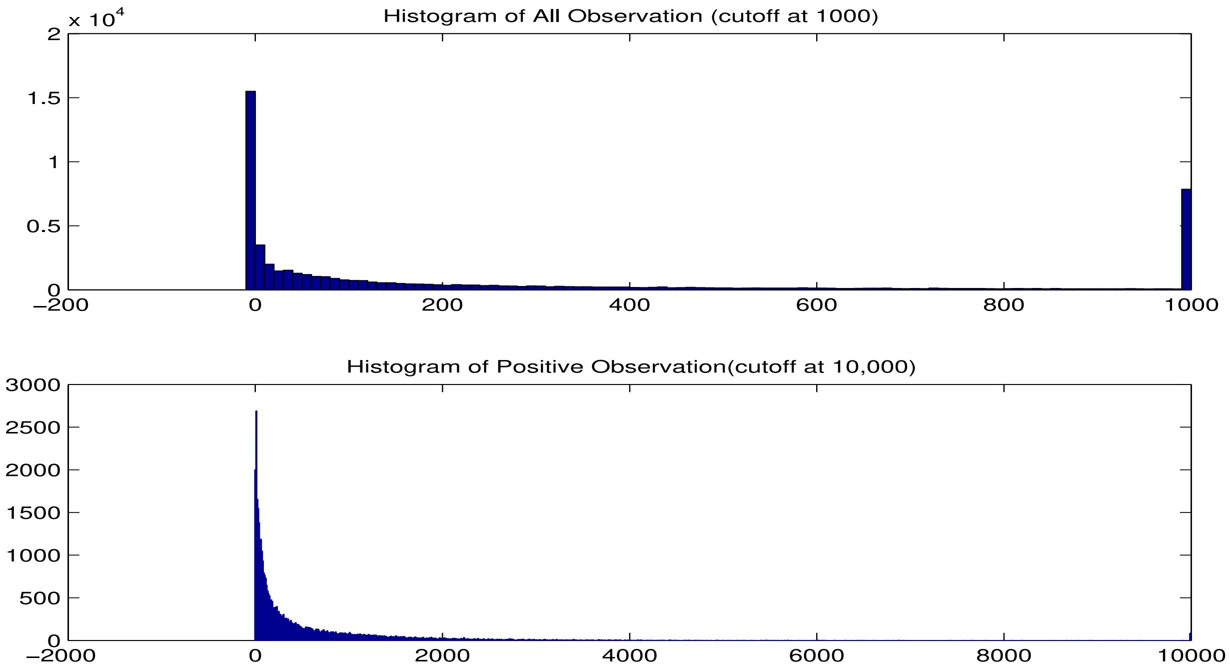
II. Principal Component Analysis (Continued)

- Summary

(1) Both MEAN and SD data show there are two significantly different groups in population.
(2) The two groups differs most in the 6 categories: analgesics/anti-inflammatory; cardiovascular; H2/PPI/motility; anti-histamines; anti-depressants; estrogens and progestins.

III. Parametric Model

Cost data in medical care system with following properties:
(a) A large portion of observations are zero (inflated zeros);
(b) The positive observations are highly skewed.



III.1 Two-Part Model

Stage 1. (Patient-Level/Within-Doctor Model)

$$Z_{ij} = I[Y_{ij}^{(1)} > 0] \times \exp(Y_{ij}^{(2)})$$

$$\begin{cases} Y_{ij}^{(1)} = b_i^{(1)} + \alpha_0 + \alpha^T X_{1ij} + \epsilon_{1ij} \\ Y_{ij}^{(2)} = b_i^{(2)} + \beta_0 + \beta^T X_{2ij} + \epsilon_{2ij} \end{cases}$$

where X_{1ij}, X_{2ij} are observed characteristics of patient j within doctor i; Z_{ij} is the observed cost of patient j within doctor i. In particular,

$$\begin{pmatrix} \epsilon_{1ij} \\ \epsilon_{2ij} \end{pmatrix} \overset{iid}{\sim} N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_e^2 \end{pmatrix})$$

III.1 Two-Part Model (Continued)

Stage 2. (Between-Doctor Model)

$$\begin{pmatrix} b_i^{(1)} \\ b_i^{(2)} \end{pmatrix} \overset{iid}{\sim} N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega)$$

where

- $b_i^{(1)}$ – Doctor i's tendency for his/her patients' to take medication;

- $b_i^{(2)}$ – Doctor i's tendency for his/her patients' to take more (expensive) medication.

III.2 Bayesian Analysis

- Non-Informative Priors

$$
\begin{aligned}
&\alpha_k \sim N(0, 10^6), k = 0, 1, \ldots \\
&\beta_k \sim N(0, 10^6), k = 0, 1, \ldots \\
&\sigma_e^{-2} \sim \text{Gamma}(10^{-3}, 10^{-3}) \\
&\Omega^{-1} \sim \text{Wishart}(\begin{pmatrix} 10^{-3} & 0 \\ 0 & 10^{-3} \end{pmatrix})
\end{aligned}
$$

III.2 Bayesian Analysis (Continued)

• Posterior Distributions: Basically, the posterior distributions are in the form of

$$[\theta|\text{data} = Z] = f(\theta|Z) \propto f(Z|\theta)\pi(\theta)$$

and the following conditional densities are available

$$\begin{aligned} & [\theta_i|Z;\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_p] = f_i(\theta_i|Z,\theta_{-i}) \\ \propto\ & f(\theta|Z) \end{aligned}$$

However, there are no closed-form posterior distributions. We have to consider the Gibbs Sampler.

III.2 Bayesian Analysis (Continued)

- Gibbs Sampler

Denote $\theta_{-i}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_p^{(t-1)})$.
Given $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_p^{(t)})$ and Z, generate

1. $\theta_1^{(t+1)} \sim f_1(\theta_1 | Z, \theta_{-1}^{(t)})$;

2. $\theta_2^{(t+1)} \sim f_2(\theta_2 | Z, \theta_{-2}^{(t)})$;

$\vdots$

p. $\theta_p^{(t+1)} \sim f_p(\theta_p | Z, \theta_{-p}^{(t)})$.

III.3 Performance Index

- Let θ include α_0, α, β_0, β, σ_e^2, Ω.

<u>**First Performance Index**</u>

- **Adjusted cost** of a patient of i-th doctor (adjusted for patient mix)

$$\mu_i^A = \frac{1}{n_i} \sum_{j=1}^{n_i} E[Z_{ij} | X_{1ij}, X_{2ij}, b_i, \theta]$$

- **Standardized cost** of a patient of i-th doctor when doctor i's patients are treated at a reference doctor

$$\mu_i^S = \frac{1}{n_i} \sum_{j=1}^{n_i} E[Z_{ij} | X_{1ij}, X_{2ij}, \theta]$$

III.3 Performance Index (Continued)

- Let $\mu^{A-S} = \{\mu_i^A - \mu_i^S; i = 1, 2, \ldots, M\}$, and $H(\cdot)$ denote a "benchmark" function. Then the posterior probability

$$P_i^{A-S} = P(\mu_i^A - \mu_i^S > H(\mu^{A-S})|Z)$$

evaluates the performance of doctor i relative to doctors of similar patients.

- Disadvantage: it is meaningful only when similar doctors are pooled in the data, i.e., the covariates X_{1ij}, X_{2ij} should not make much difference across doctors.

III.3 Performance Index (Continued)

Second Performance Index

- Consider the expected cost of a reference patient with characteristic X_1^* and X_2^*

$$Z_i^* = E[Z|X_1^*, X_2^*, b_i, \theta]$$

- The posterior probability

$$P_i^* = P(Z_i^* > H(Z*)|Z)$$

evaluates the performance of doctor i relative to all doctors in the sample.

III.4 Preliminary Results

- Convergence Diagnostics:

(1) Stationary Test: 1 out of 207 parameters failed.
(2) Halfwidth Test: 2 out of 207 parameters failed.

- Estimation Results:

- Performance Index:

Discussion

- Comparison of modeling with whole data vs. stratified data.

- SiZer is an efficient way to decide "which features are really there".

- Combining PCA and Parametric Models to define new performance index?