Understanding and Visualization of Feature Spaces for Biometrics Applications*

David W. Scott, Dept. Statistics, Rice Univ., Houston

CSTB Biometrics Workshop
March 15, 2005 The National Academies

*Research supported in part by NSF grant DMS 02-04723 (non-parametric methodology) and NSF contract EIA-9983459 (digital government).

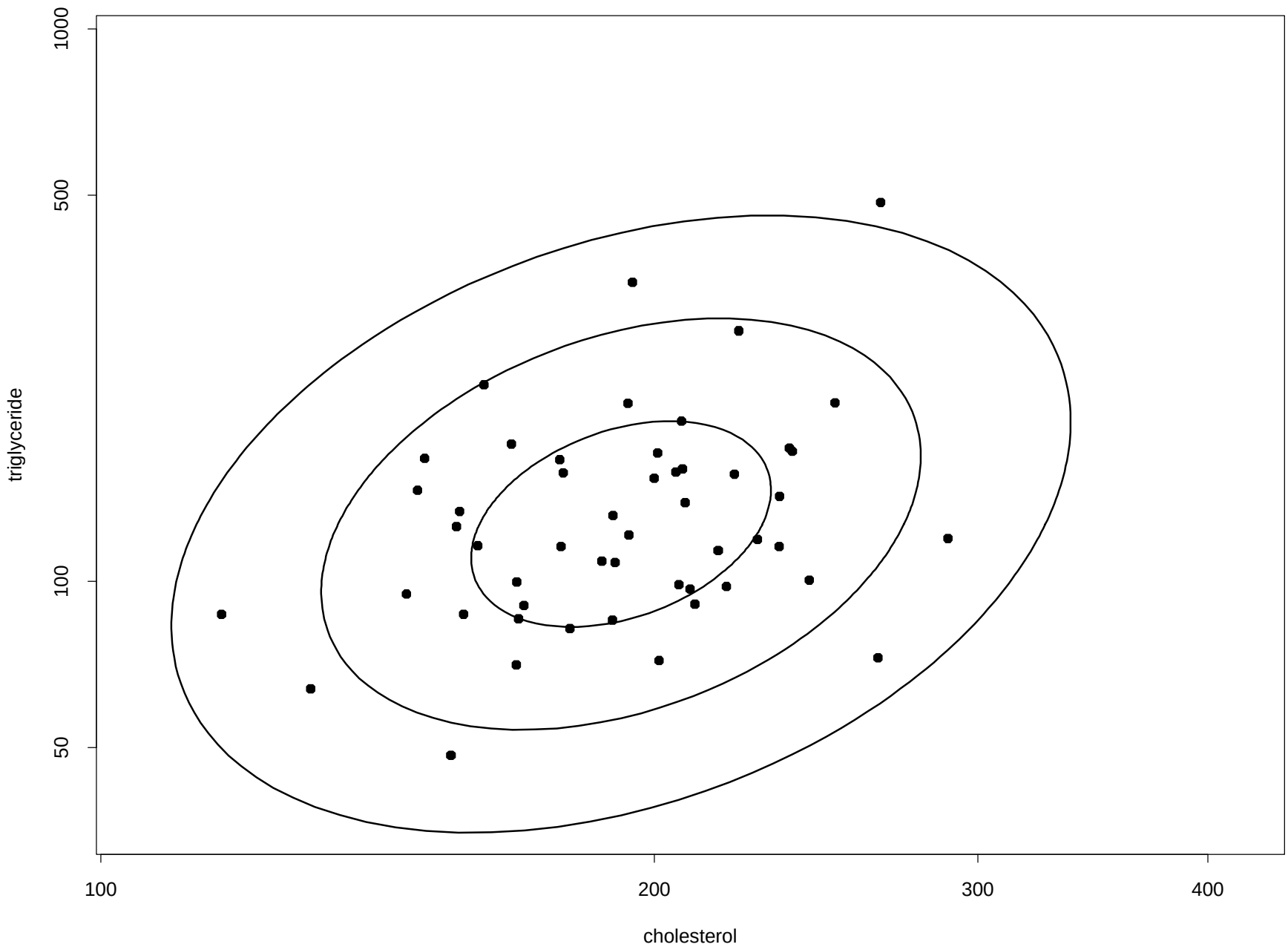
1 Introduction

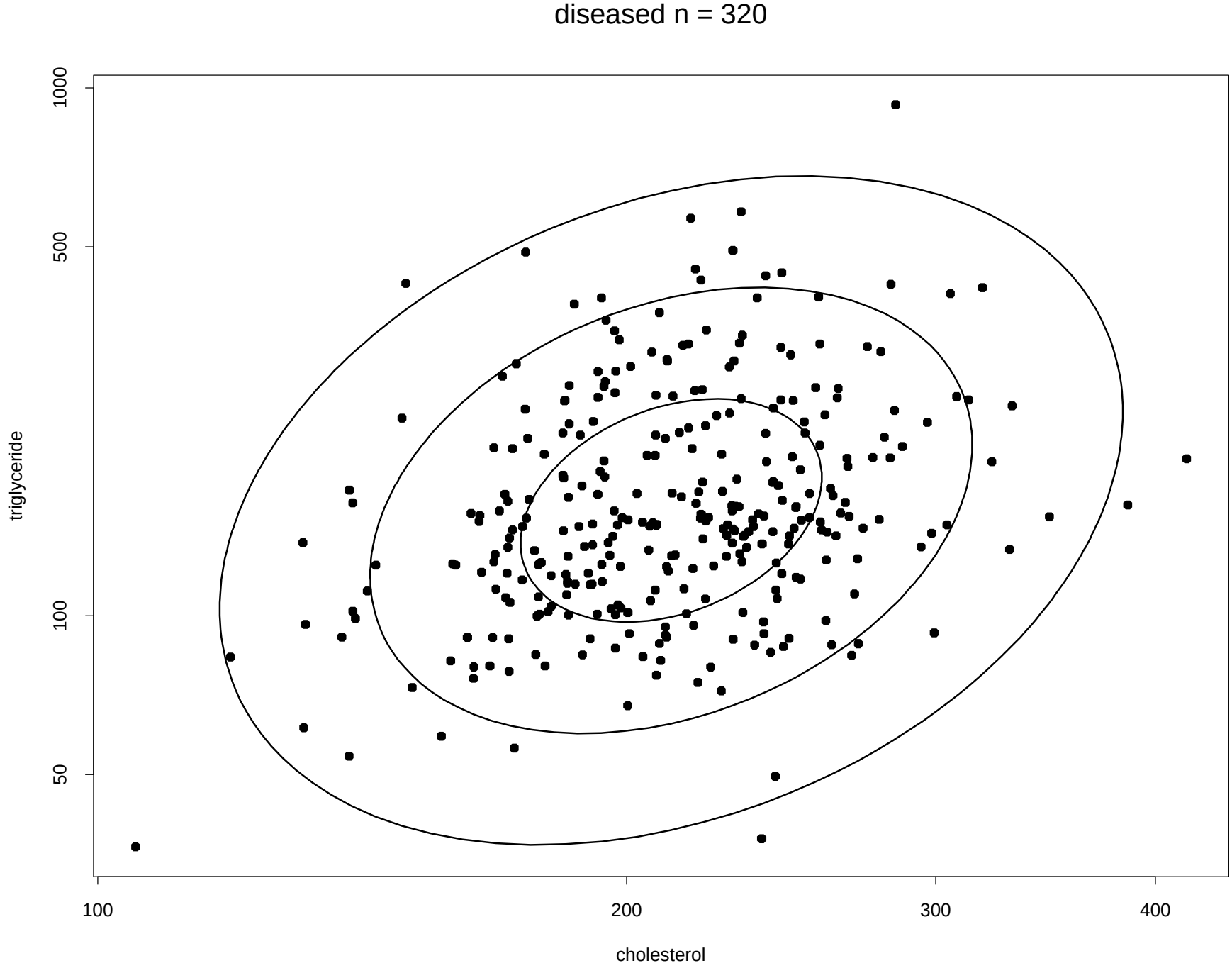
- engineering the right feature space
- visualization and understanding of the feature space
- much to be enthusiastic about
- how good is good enough?
- statistical polling used to be highly reliable
  - TV glamorized pollsters and encouraged participation
  - ignorance of the technology helped
  - nonstationary technology (cell phones)
  - public understanding of the impact of polls on election outcomes has led to decreased participation (learning/feedback)
  - similar to (but not exactly like) colored balls in urn

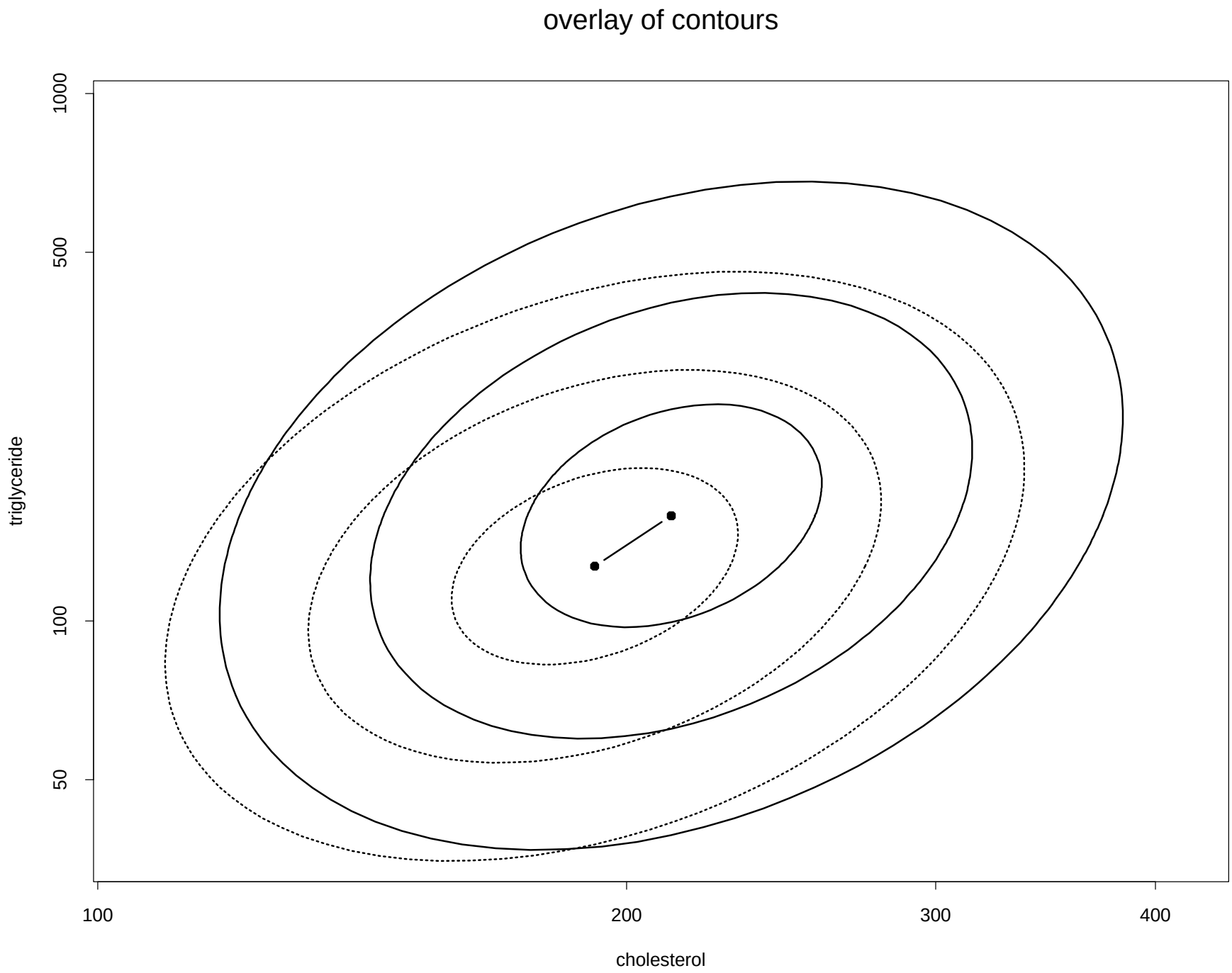
- consider automatic license plate scanning at border
  - nice big block alphabet (but not uniform)
  - what if mud splattered plate?
  - different than an altered plate?
  - (should an iris scan work on a black eye?)
- human cooperation assumed ("look straight into the camera")
- key for my Dad's new 1965 Oldsmobile Delta 88 opened another at the airport (I was sure each key was unique!)
- it is hard enough to get equipment to work; should we worry about countermeasures?
- biostatisticians have relevant experience with human trials:
  - multicenter studies & equipment variation
  - human evaluation limitations
  - analysis tools

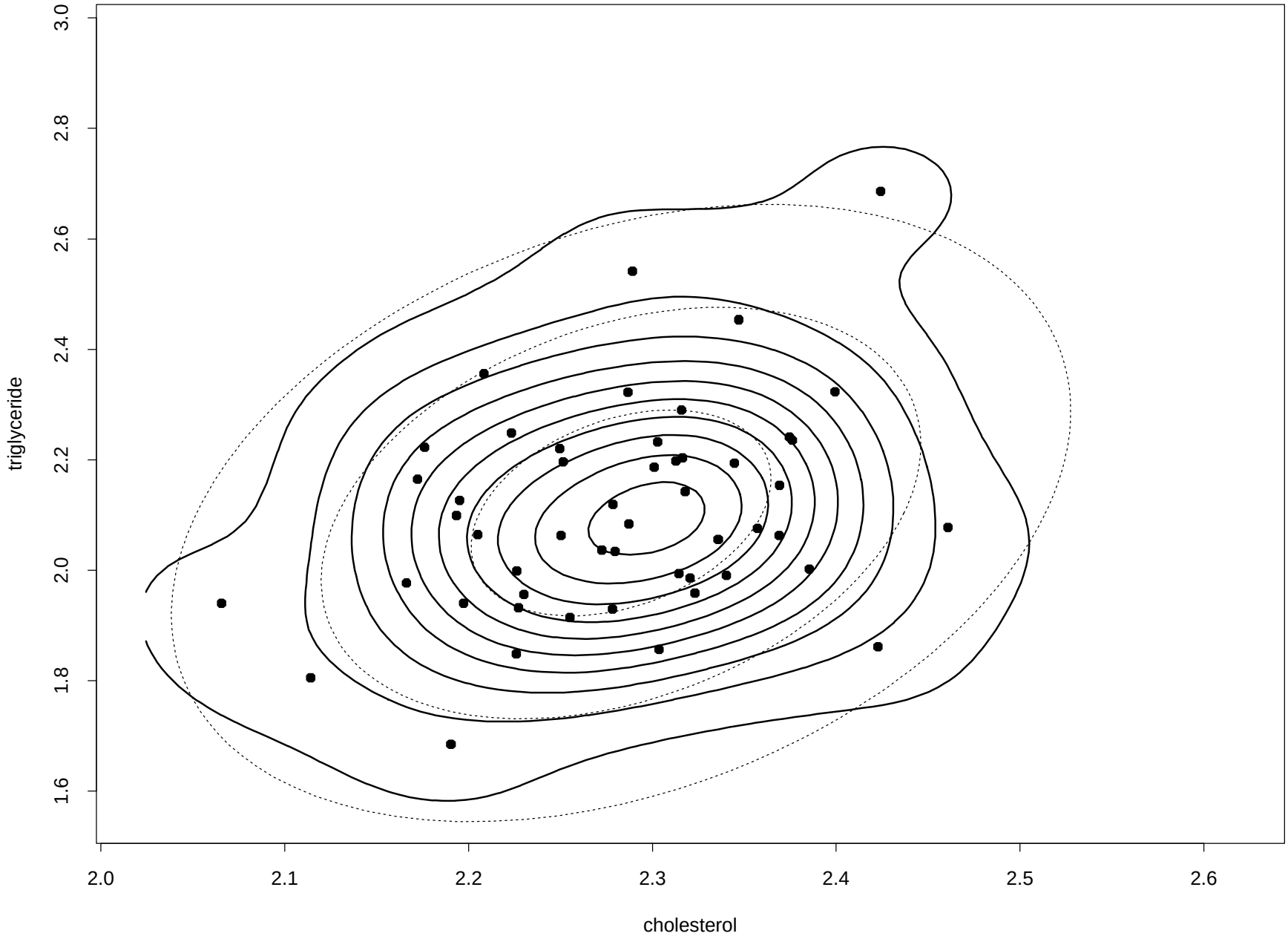
2 Statistical Discrimination and Feature Spaces

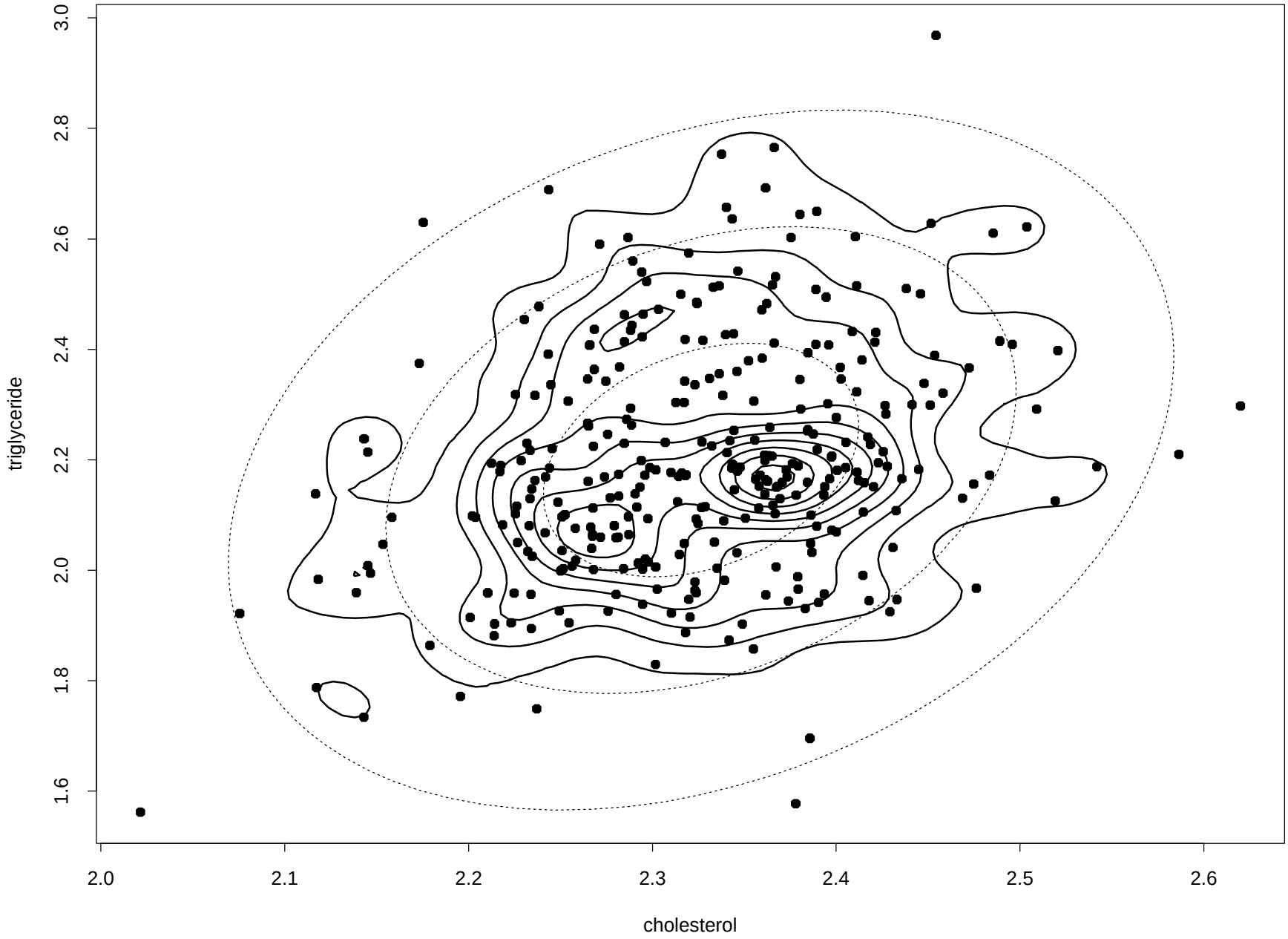
- discriminating diseases is like discriminating individuals
- in common statistical application, looking for structure (grouping) in feature space
- do blood lipids/fats discriminate heart patients?
- Reference: Scott, Gotto, Cole, and Gorry (1978). "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease: A Study of 371 Males with Chest Pain." *Journal of Chronic Diseases*, 31:337-345.

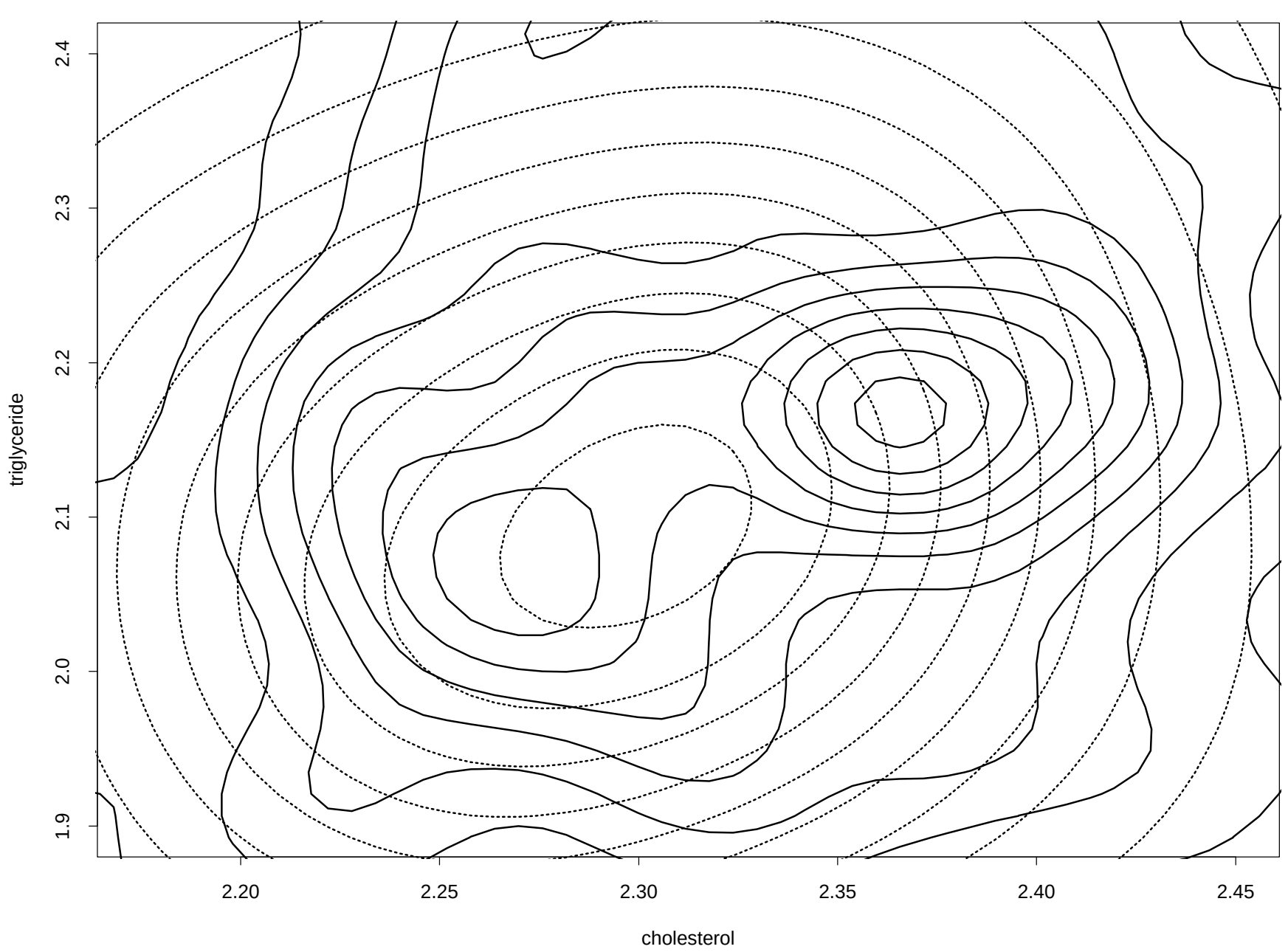

normal n = 51
triglyceride
cholesterol
1000
500
100
50
100
200
300
400


diseased n = 320
triglyceride
1000
500
100
50
100
200
300
400
cholesterol


overlay of contours
triglyceride
1000
500
100
50
100
200
300
400
cholesterol


normal n = 51
triglyceride
cholesterol
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
2.0
2.1
2.2
2.3
2.4
2.5
2.6


normal n = 320
triglyceride
cholesterol
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
2.0
2.1
2.2
2.3
2.4
2.5
2.6


overlay of contours (modes at 186 and 233)
triglyceride
2.4
2.3
2.2
2.1
2.0
1.9
2.20
2.25
2.30
2.35
2.40
2.45
cholesterol

- lipids do not even discriminate heart disease, much less individuals
- lipids change over time
  - lipids increase naturally about 5-10% each decade
  - lipid drugs can lower almost 50%
- biometrics feature spaces will drift (nonstationary)

- need richer feature spaces
- LANDSAT IV LACIE (agricultural crop inventory system)
- raw data: $\mathbf{x} \in \Re^4$ spectral bands
- principal components: $\Re^4 \rightarrow \Re^2$ sufficient
- 6 overpasses during growing season ($6 \times 4 = 24$ measurements per pixel)
- agronomist proposed nonlinear transformation: $\Re^{24} \rightarrow \Re^3$
- $K = 3$ crops: sunflower, spring wheat, spring barley

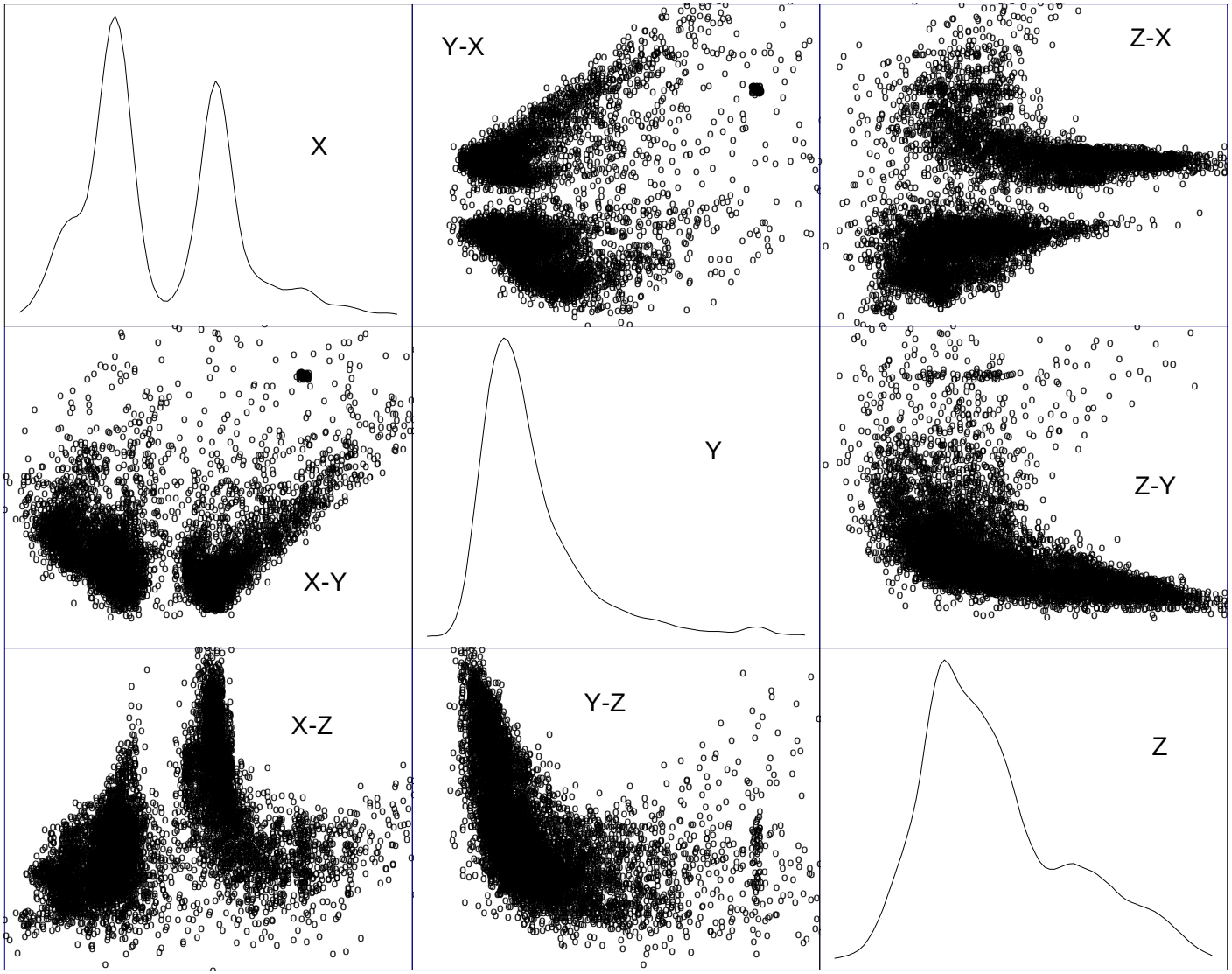


Figure 1: Landsat IV: 3 features of 3 crops (sunflower, spring wheat, spring barley)

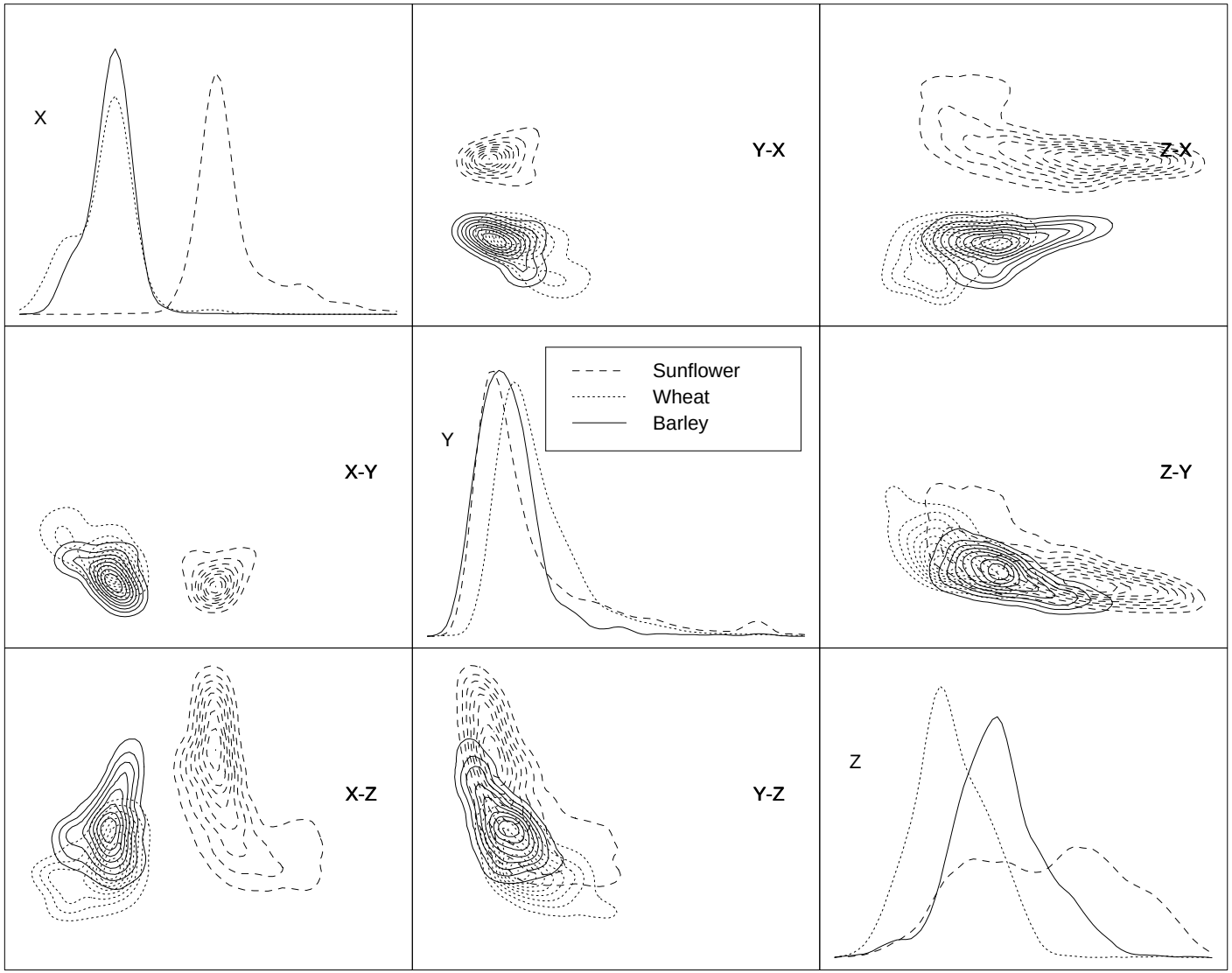


Figure 2: Landsat: 3 crops and 3 features

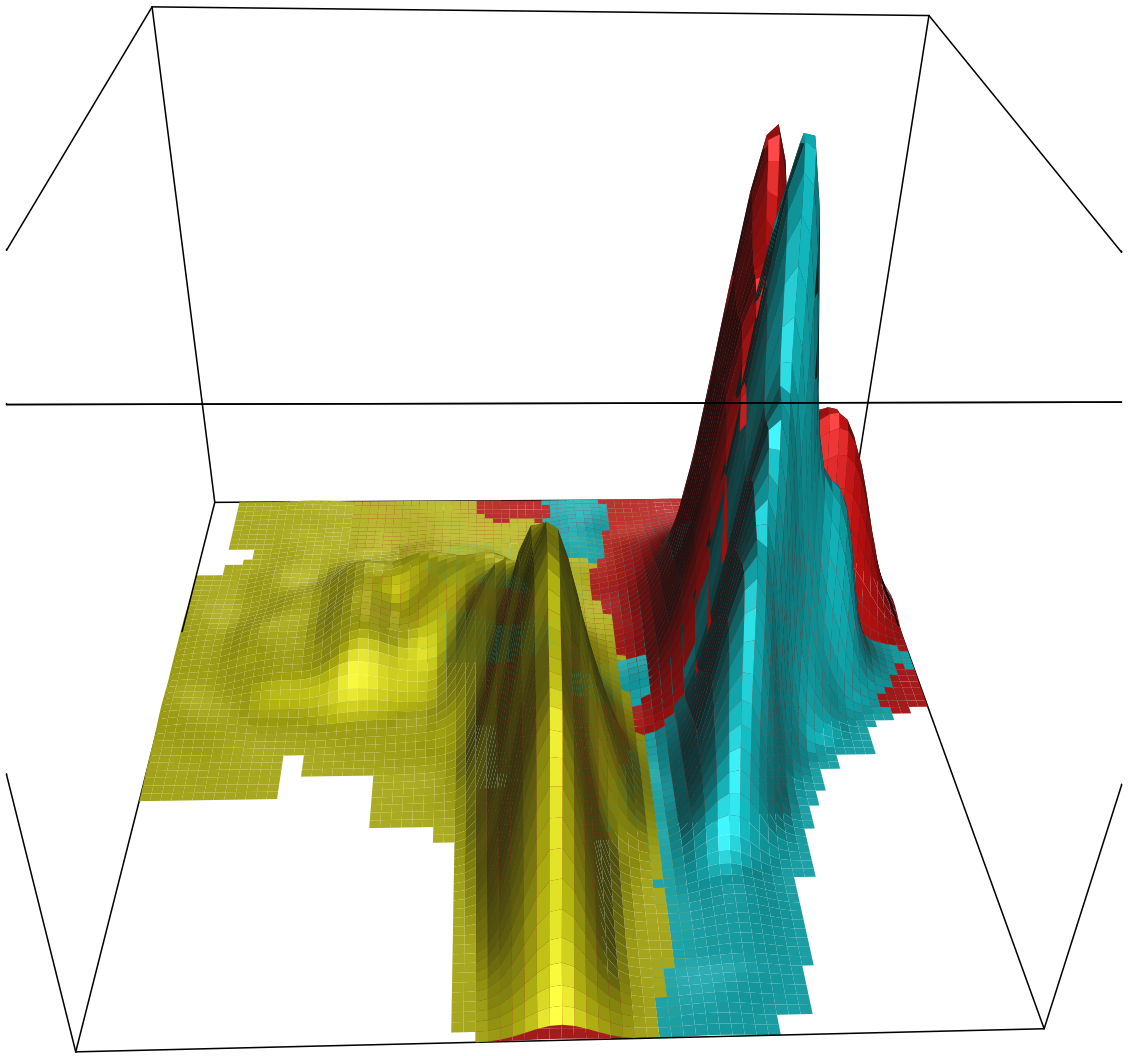


Figure 3: Landsat: 3 crops and pairwise features

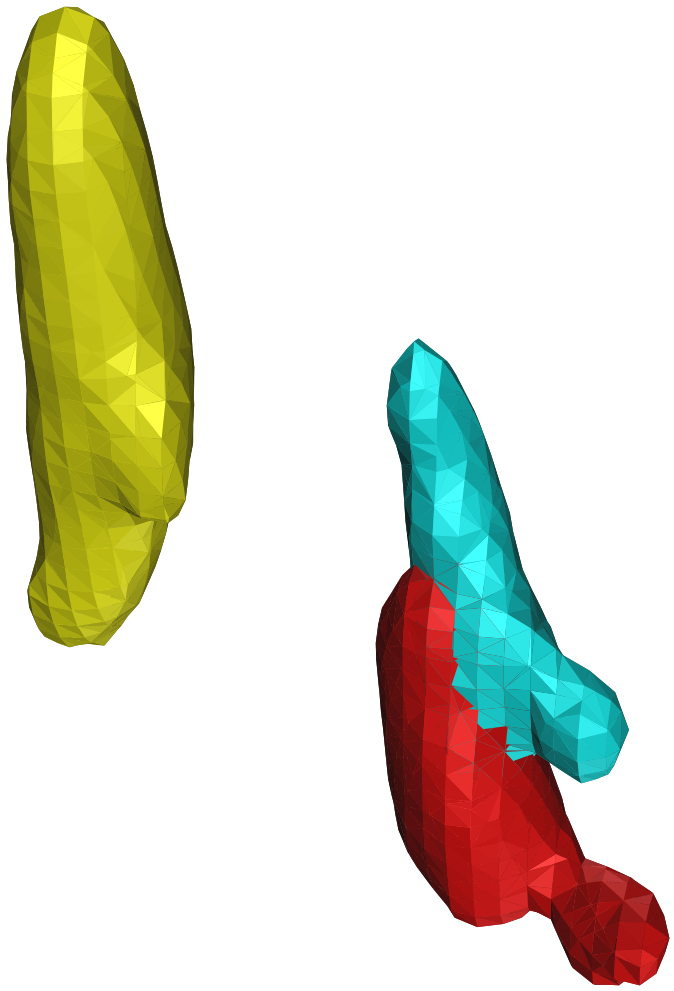


Figure 4: Landsat: 3 crops and trivariate features

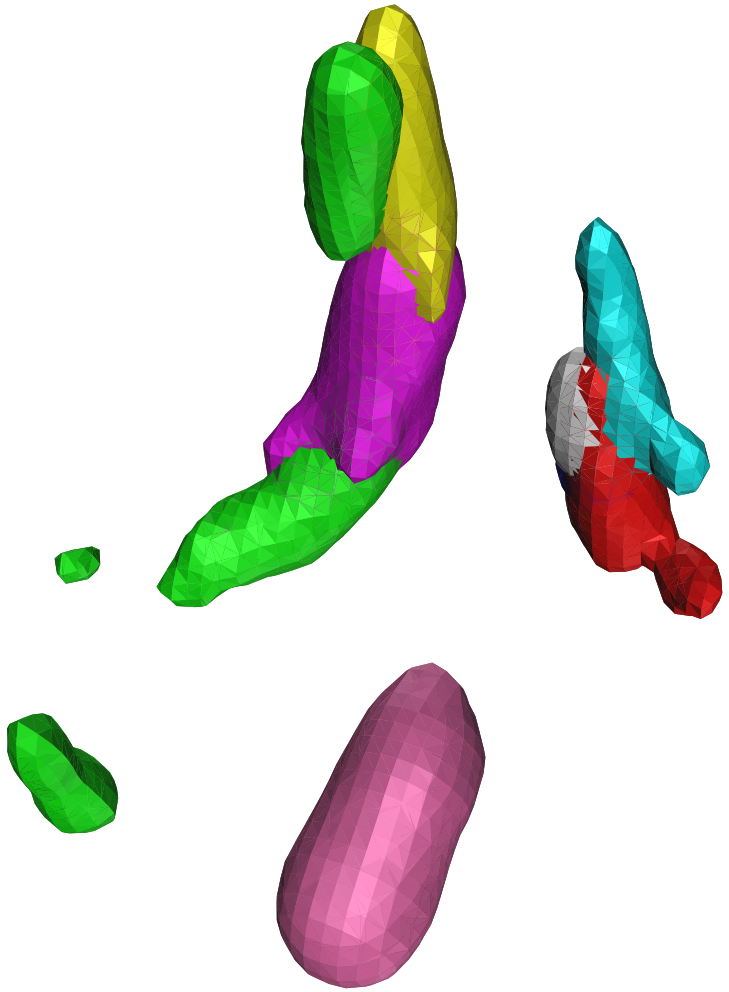


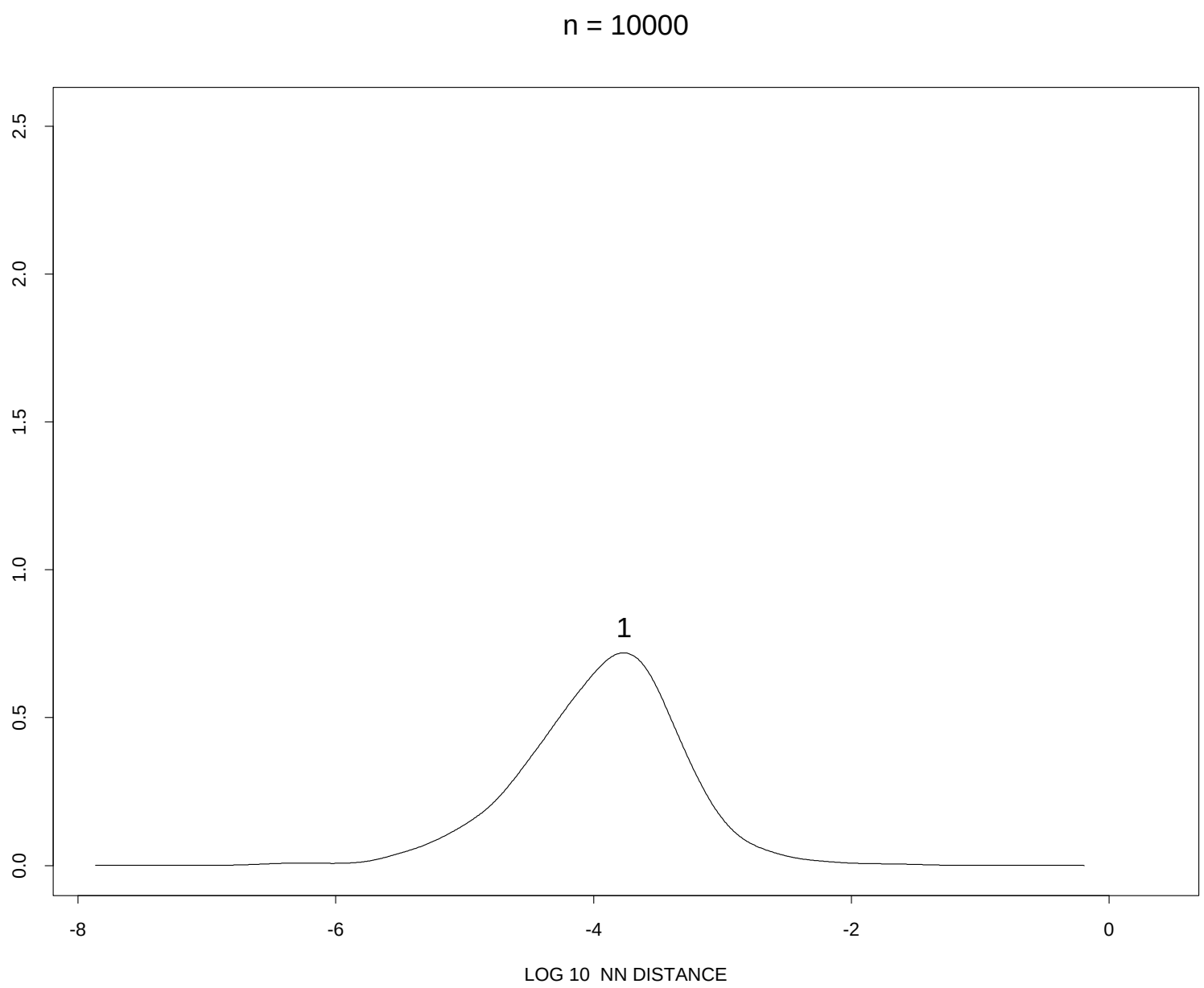
Figure 5: Landsat: all crops and trivariate features

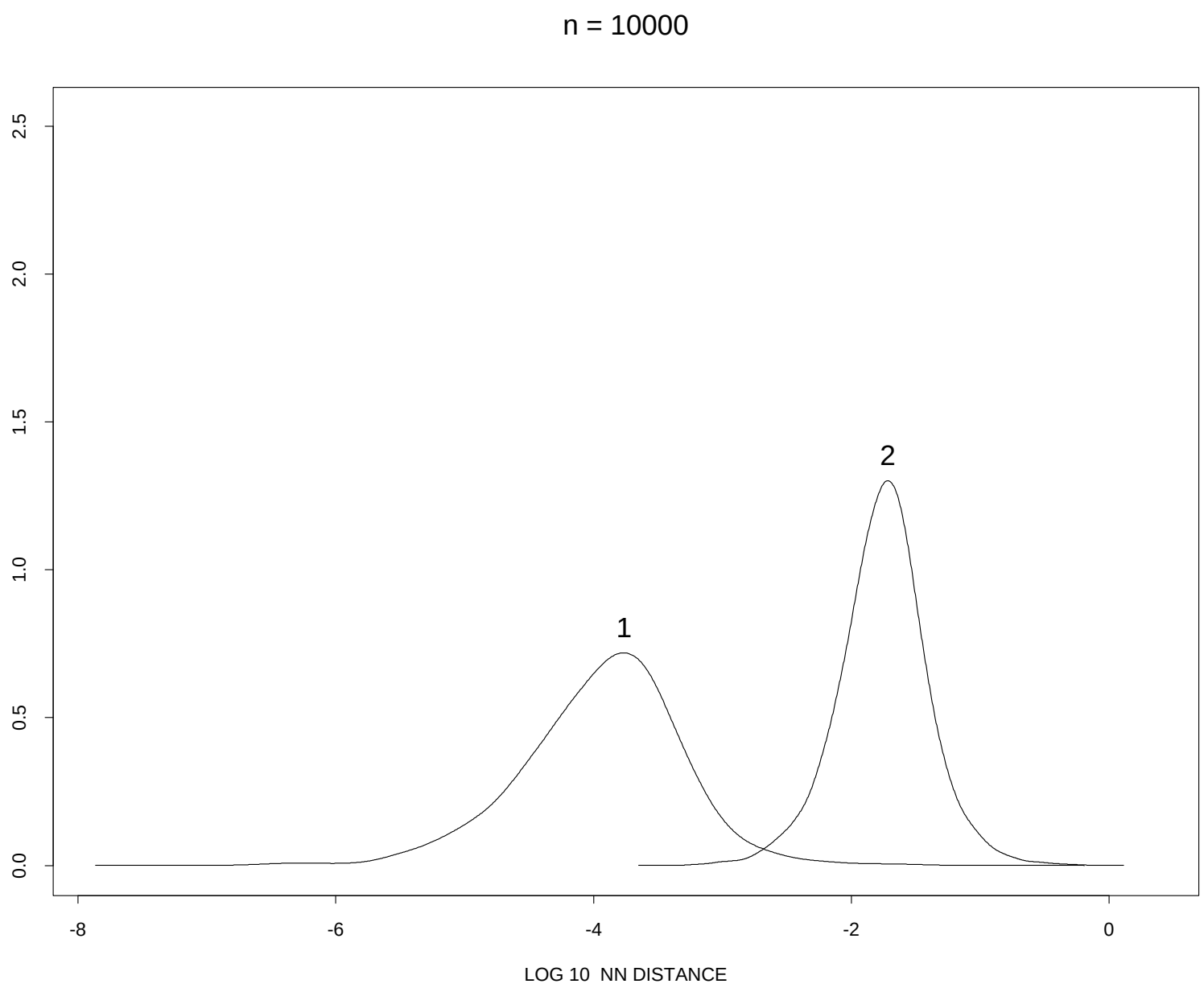
- not enough dimensions (yet) — hyperspectral satellites today
- visualization of $p > 3$ dimensions obviously hard
- how do we understand models in $\Re^{40}$?
- e.g. NIST-1996 and ICASSP-99
- one of Joe Campbell & Doug Reynolds' speech models
- EM fitting of millions of training points with $K = 2048$ normal mixtures
- found very strong structure
  - determinants of Σ_k varied by factor of over 10^{15}
  - many components not necessary (engineering solution: it works)
  - IPRA algorithm found $K = 1024$ and 512 models as good
- useful feature space need not be optimal

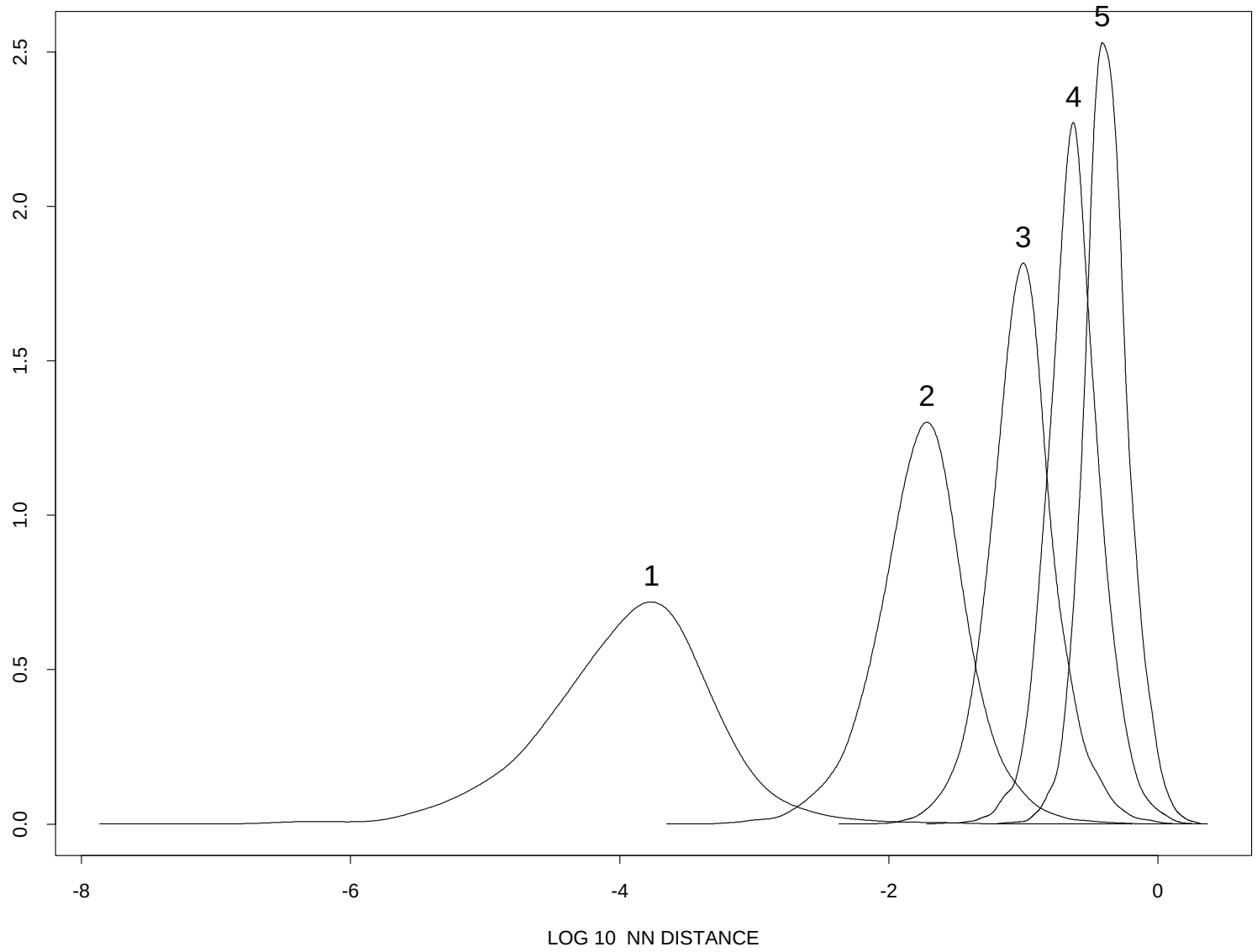
3 How Large Should A Feature Space Be?

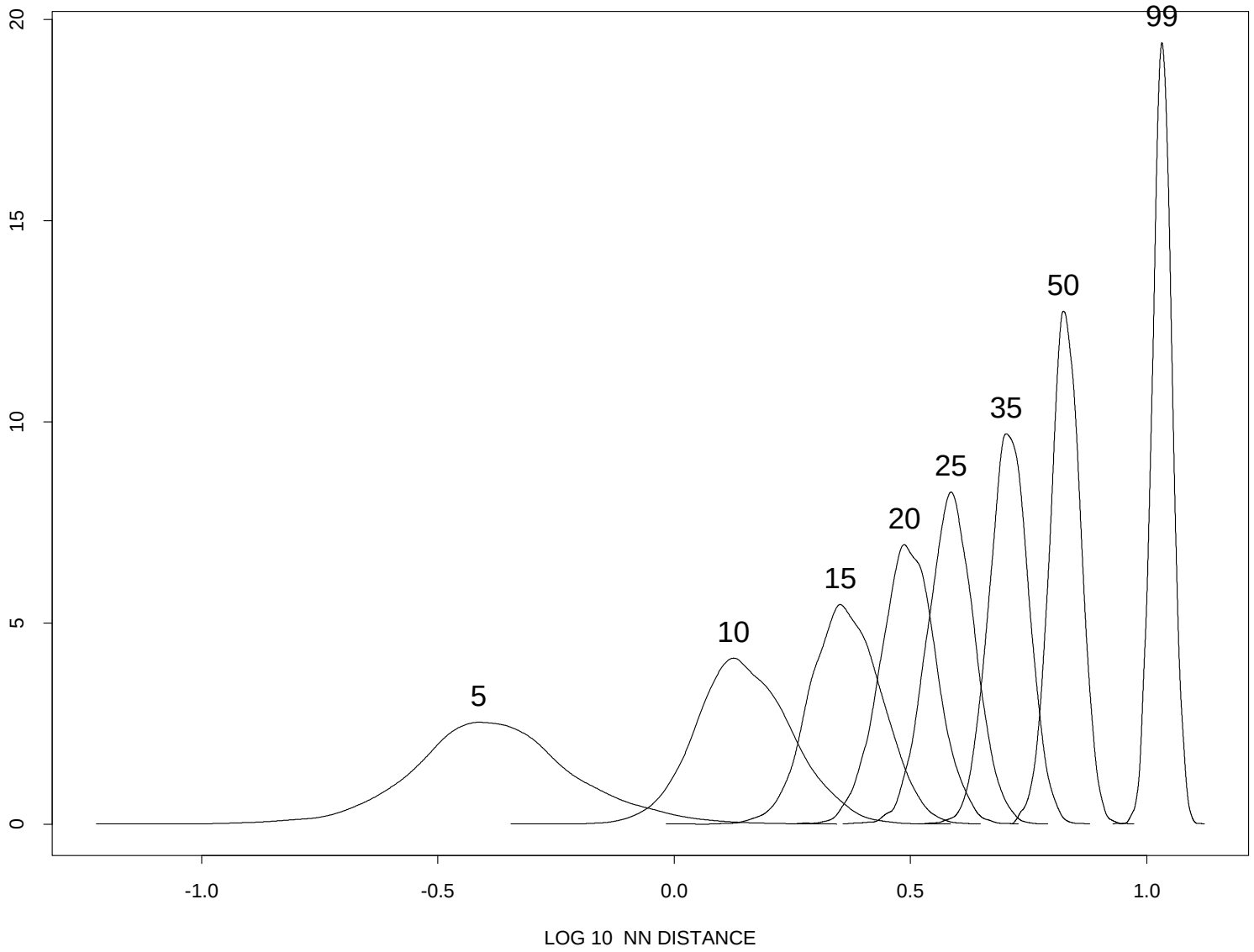
- For LANDSAT, $n = 10^7$ and $3 < p < 20$ with $K = p$
- For Biometrics, $K = n!!$ Very different than usual problem.
- Questions:
  - *Can multiple samples from one individual be reliably distinguished from similar samples on all other* $n - 1$ *individuals?*
  - As n increases?
  - As p increases?
  - How much does multimodal biometrics help?

- What happens in high-dimensional feature spaces?
- $\mathbf{x}_k \sim N(\mu_k, \Sigma_k)$ — model for individual's measurements
- $\mu_k \sim N(\mu, \Sigma)$ — model for individuals' means
- More generally, $\mu_k \sim \sum_{k=1}^{K} w_k \, N(\gamma_k, \Gamma)$
  - for example, differences by sex, etc.
  - $K > 1$ helps in overall performance
  - still need to understand performance in a single pure group
- easiest case: $\mu \sim N(0, I_p)$ with $K = 1$
- (aren't all biometric measurements normal? Galton, Pearson)
- generate n biometrics measurements:
  - how close is the nearest alternative?
  - what is the distribution of such distances over all n individuals?

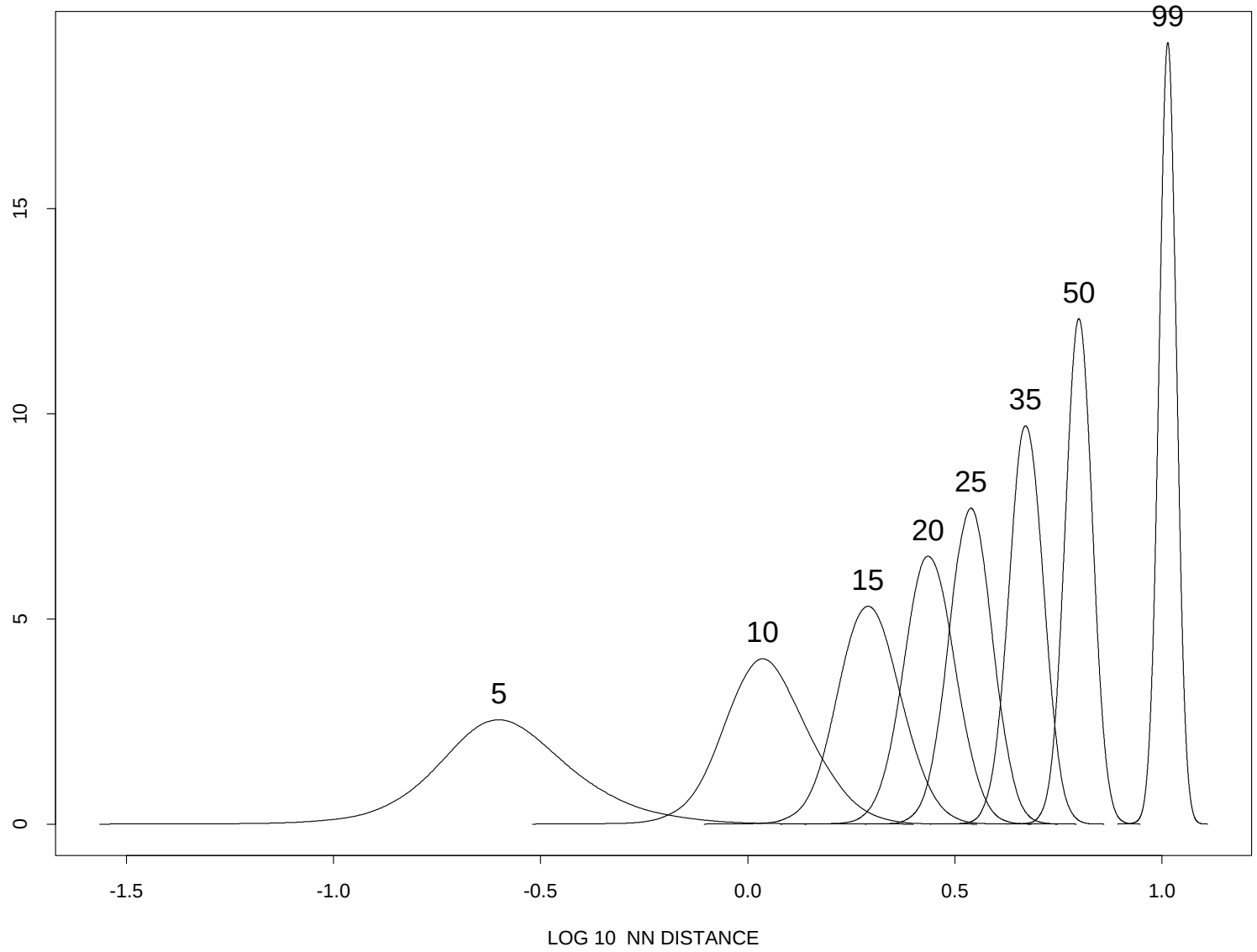

n = 10000
2.5
2.0
1.5
1.0
0.5
0.0
1
-8
-6
-4
-2
0
LOG 10 NN DISTANCE

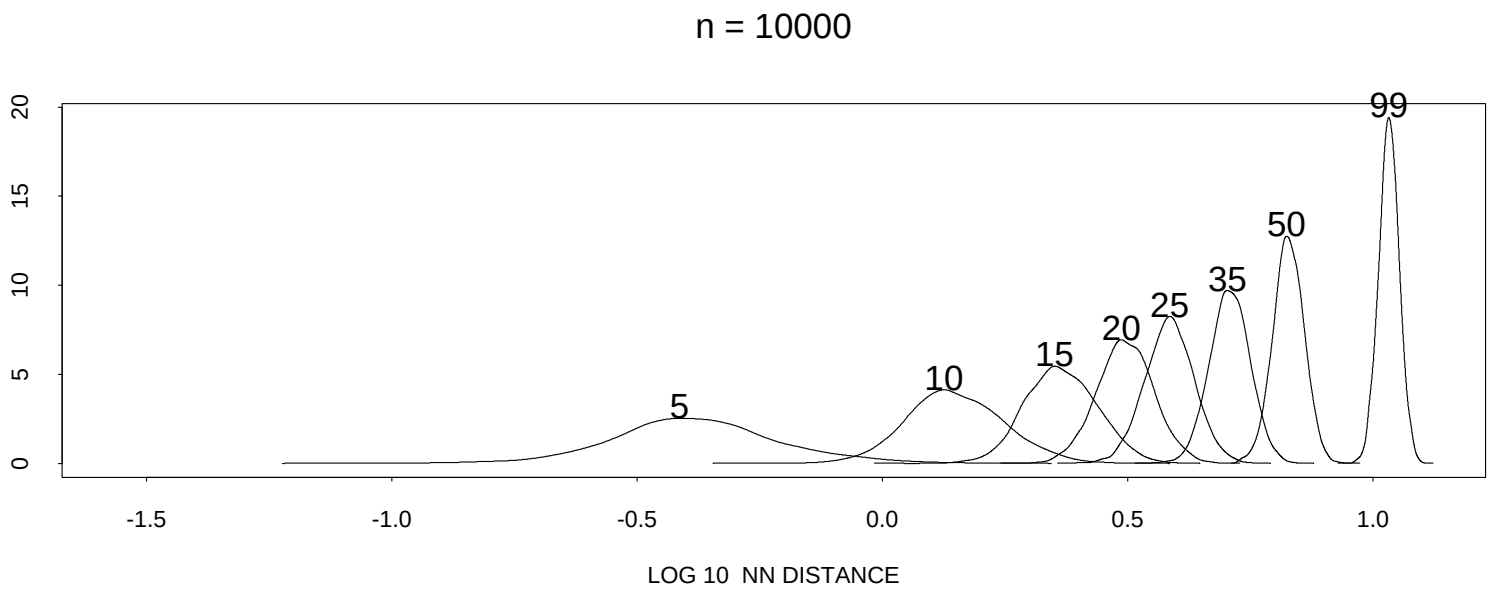

n = 10000
2.5
2.0
1.5
1.0
0.5
0.0
1
2
-8
-6
-4
-2
0
LOG 10 NN DISTANCE

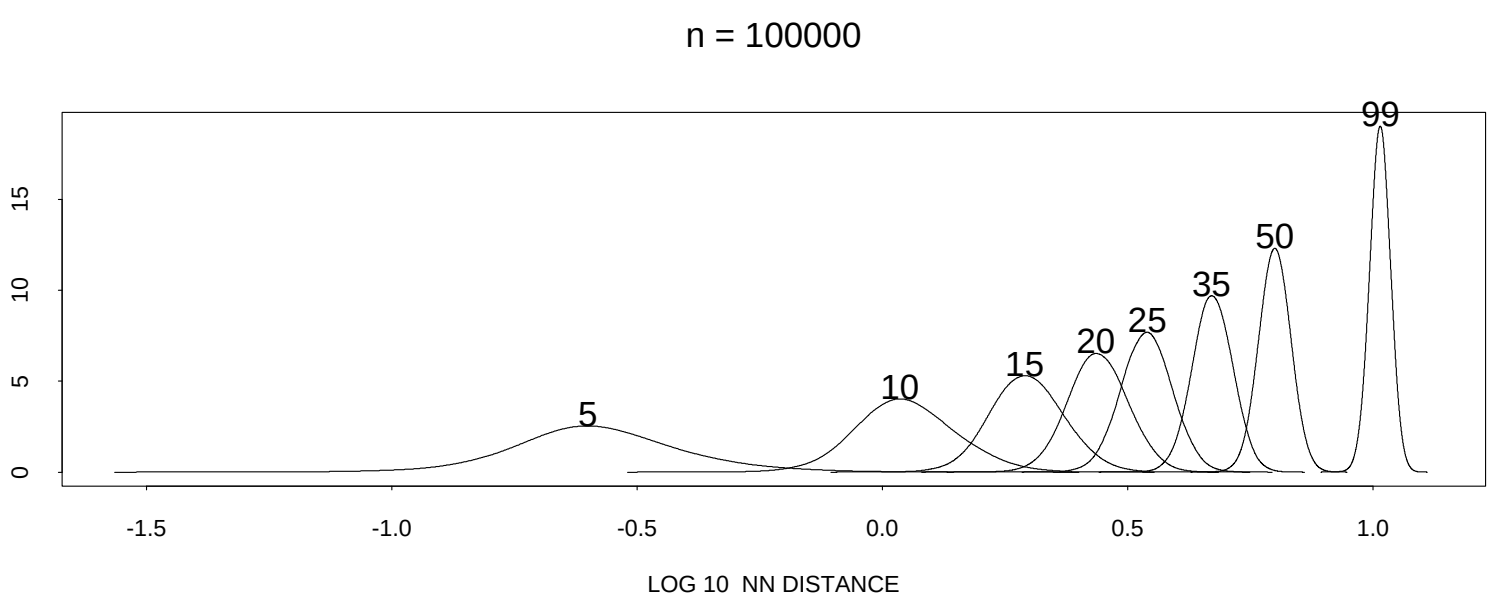

n = 10000
0.0
0.5
1.0
1.5
2.0
2.5
-8
-6
-4
-2
0
LOG 10 NN DISTANCE
1
2
3
4
5

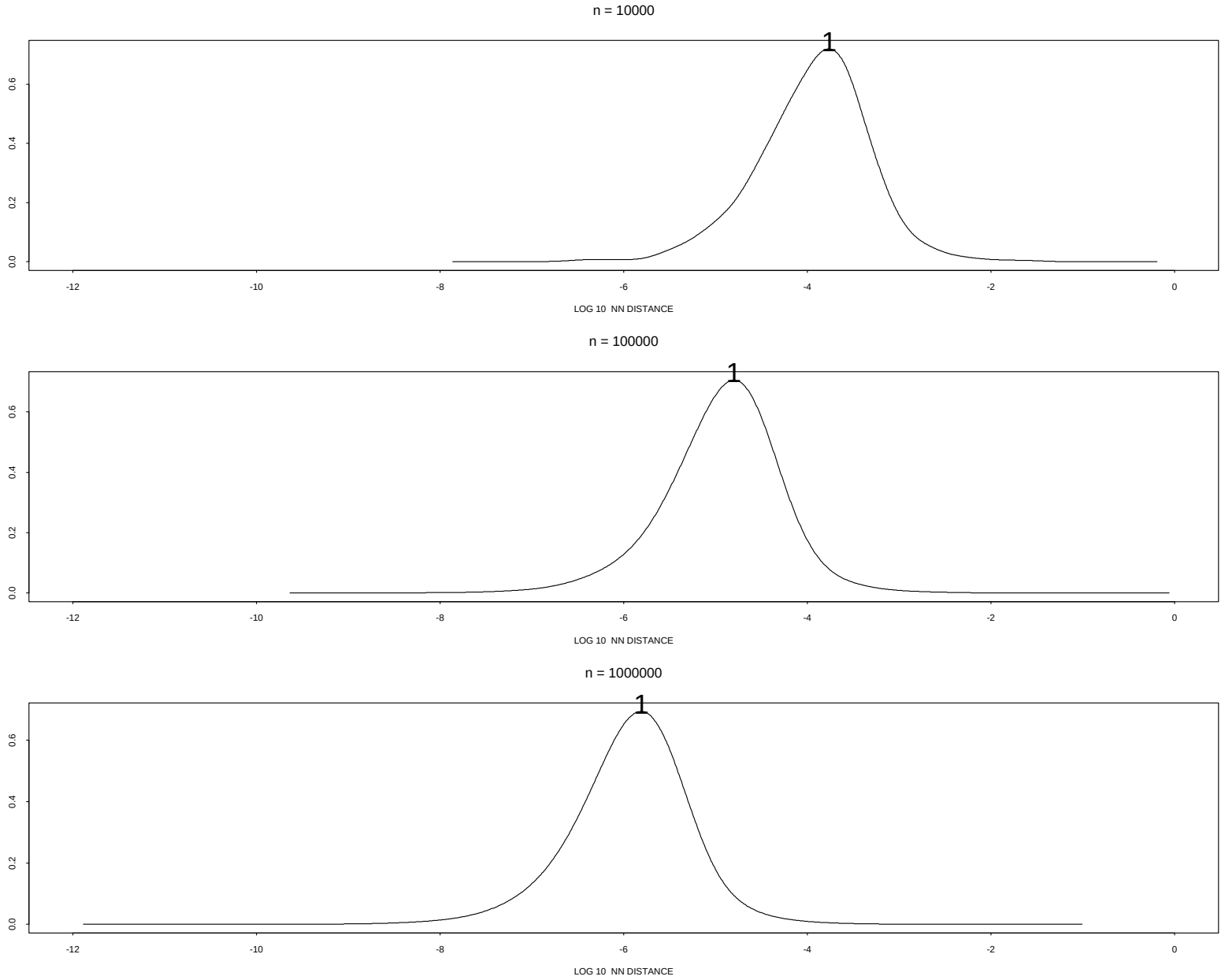

n = 10000
99
50
35
25
20
15
10
5
0
5
10
15
20
-1.0
-0.5
0.0
0.5
1.0
LOG 10 NN DISTANCE

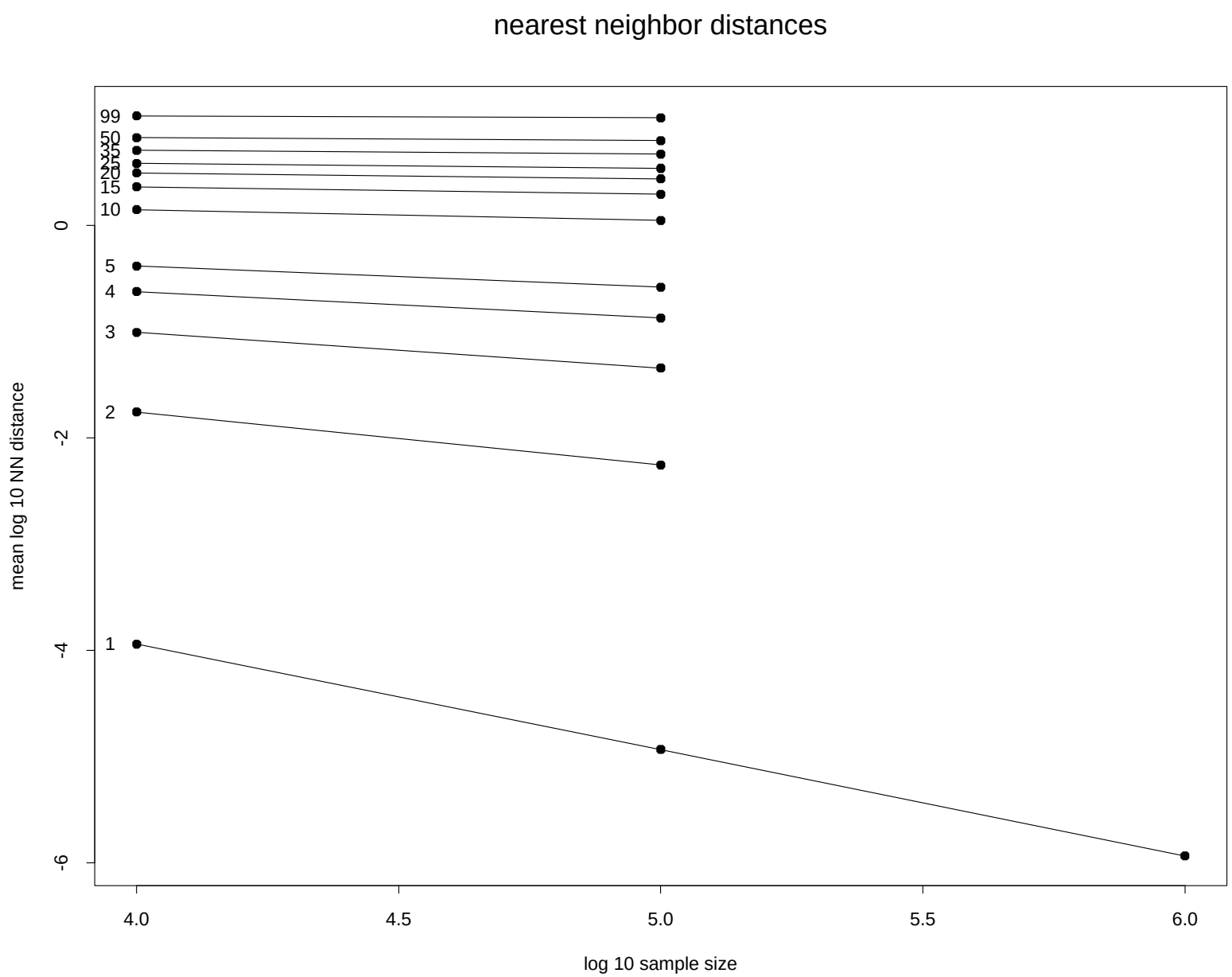
- with 99 independent features, the most unfavorable individual is still a distance of 10 units away from her nearest "match"
- with only 15 independent features, that distance is 1 unit
- *if* can measure $\mathbf{x}_k$ reliably with smaller σ, all set
- the picture is much the same if the population is larger


n = 100000
99
50
35
25
20
15
10
5
0
5
10
15
-1.5
-1.0
-0.5
0.0
0.5
1.0
LOG 10 NN DISTANCE


n = 10000
0
5
10
15
20
99
50
35
25
20
15
10
5
-1.5
-1.0
-0.5
0.0
0.5
1.0
LOG 10 NN DISTANCE


n = 100000
0
5
10
15
99
50
35
25
20
15
10
5
-1.5
-1.0
-0.5
0.0
0.5
1.0
LOG 10 NN DISTANCE


n = 10000
1
0.0
0.2
0.4
0.6
-12
-10
-8
-6
-4
-2
0
LOG 10 NN DISTANCE
n = 100000
1
0.0
0.2
0.4
0.6
-12
-10
-8
-6
-4
-2
0
LOG 10 NN DISTANCE
n = 1000000
1
0.0
0.2
0.4
0.6
-12
-10
-8
-6
-4
-2
0
LOG 10 NN DISTANCE

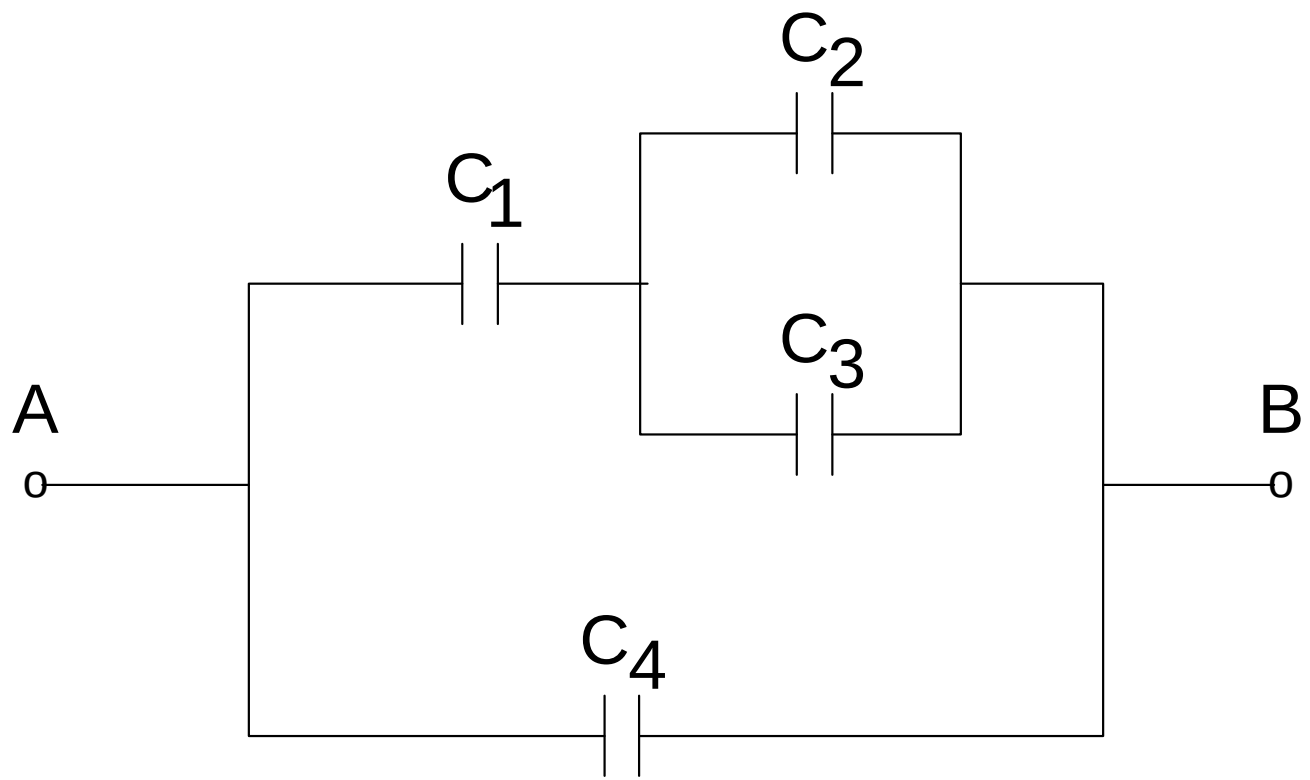

nearest neighbor distances
99
50
35
25
20
15
10
5
4
3
2
1
0
-2
-4
-6
mean log 10 NN distance
4.0
4.5
5.0
5.5
6.0
log 10 sample size

- here, each increase in the population by a factor of 10, reduces the mean of the log10 distance by a factor of about 10 (not so for larger p)
- context helps (a more limited population)
- however, the "curse of dimensionality" helps for Biometrics
- points become sparse in high dimensions, even with "normal" data
- if drift and nonlinearities and biases are less than a certain threshold, then reliable biometrics is feasible
- well-designed studies (see biostatistics) can answer this question

- high-dimensional feature spaces are
  - *necessary*, so very small chance of confusion
  - *not sufficient* — if highly correlated features, then the effective dimension is much reduced
  - if drift away from flat feature space, then extrapolation can destroy performance
  - big question: will statistical modelling outperform non-model-based algorithms in this situation? better extrapolation?

4 Multimodal Biometrics

- increase both sensitivity and specificity
- use in series or in parallel?
- pass any single Biometrics test (parallel)
- pass all Biometrics tests (series)
- too stringent or relaxed
- in a large-scale operation, one test likely to mess up on different individuals
- to increase both sensitivity and specificity, use multimodal Biometrics in a combined decision form, such as


C2
C1
C3
A
B
C4

5 Summary

- finding a useful feature space may require work and understanding
- visualization techniques and modeling may aid here
- does the normal distribution capture human variation?
- it is easy to compute sample histograms of nearest-neighbor distances and try to match to these simulated densities in order to find the "effective dimension" (account for feature correlation)
- application in large-scale settings should assume that one or more tests will not work for an individual for various legitimate reasons; a combination of series and parallel application will still allow high performance
- the tools of biostatistics should have application here: analysis, design, randomization, and evaluation
- www.stat.rice.edu/∼scottdw