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?

normal n = 51
diseased n = 320

cholesterol

triglyceride

100 200 300 400

50 100 500 1000
normal n = 320
overlay of contours (modes at 186 and 233)
• 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: \( \mathbf{x} \in \mathbb{R}^4 \) spectral bands

• principal components: \( \mathbb{R}^4 \rightarrow \mathbb{R}^2 \) sufficient

• 6 overpasses during growing season (\( 6 \times 4 = 24 \) measurements per pixel)

• agronomist proposed nonlinear transformation: \( \mathbb{R}^{24} \rightarrow \mathbb{R}^3 \)

• \( K = 3 \) 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 \( \mathbb{R}^{40} \)?
• 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 \( \Sigma_k \) 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?
$n = 10000$

Graph showing the distribution of log10 NN distances.
$n = 10000$
n = 10000

LOG 10 NN DISTANCE

-8 -6 -4 -2 0
0.0 0.5 1.0 1.5 2.0 2.5
1 2 3 4 5

LOG 10 NN DISTANCE
n = 10000
• 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 $x_k$ reliably with smaller $\sigma$, all set

• the picture is much the same if the population is larger
nearest neighbor distances

log 10 sample size

mean log 10 NN distance

4.0 4.5 5.0 5.5 6.0

-6 -4 -2 0 2 3 4 5 10 15 20 25 35 50 99

log 10 sample size

mean log 10 NN distance
• 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
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