



---

# The Cotton Ball Problem

Dr. Jim Wayman

With thanks to: David Herrington, Jim Maar, Joe McCloskey, Kang James, Barry James, Art Owen, Hani Dass, David Donoho

Typo corrected Oct. 20, 2004



# Technical Support Working Group



San José State  
UNIVERSITY

- 1986 V.P. Bush task force on CbT
- Finding: counter-terrorism response hurt by lack of R&D infra-structure.
- President Reagan created Interdepartmental Group on Terrorism (IG/T) with DoS as lead.
- TSWG created as sub-group of IG/T for technology development under DoD administration



# References



- H. Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946
- K. Fukunaga, Introduction to Statistical Pattern Recognition, 2<sup>nd</sup> ed. Academic Press, 1990
- R.Duda, P. Hart, D. Stork, Pattern Classification, 2<sup>nd</sup> ed., John Wiley, 2001
- A. Jain, R. Duin, and J.C. Mao, “Statistical Pattern Recognition: A Review”, PAMI, 32(1), Jan. 2000
- S. Raudys and D. Young, “Results in statistical discriminant analysis: A review of the former Soviet Union literature”, Journal of Multivariate Analysis, Vol. 89, pp. 1–35, 2004



# Mathematical Statistics in Biometrics



- Biometrics is a behavioral, not physical, science
- Mathematical statistics can be usefully applied to behavioral sciences PROVIDED THAT we:
  - Carefully note limitations of our simplifying assumptions
  - Respect unknown and known, but uncontrollable, factors impacting human behaviors
  - Acknowledge numerical uncertainty caused by above.



# The Open Universe Problem



San José State  
UNIVERSITY

- In application, biometrics is a one-class problem.
- Priors are unknown.
- Open universe identification is  $N$  sequential one-class problems.
- Consequently, methods from  $N$ -class analysis may not apply.



# Cotton Balls



- Imagine throwing identical round cotton balls into a box.
- What is the probability that the next ball hits one already in the box?
- Can we write probability of collision as function of dimension of space, within- and between-class distributions, size of cotton ball, number of balls?
- Can we ultimately write FMR as function of FNMR, dimensionality, and fundamental distributional parameters for the one-class problem?
- Current funding from TSWG and SAG programs



# Assumptions



- Euclidean space
- Gaussian within-class distributions
  - **Initially**, i.i.d.
  - Exactly known mean vector and covariance matrix
    - This implies infinite training data
- Class homogeneity
- Known between-class distributions
  - Gaussian or uniform over finite domain
  - No impact of detection/partitioning algorithms
- Continuous, time-invariant variables
  - No quantization effects
  - No “template aging”



# Why Continue Under These Assumptions?



San José State  
UNIVERSITY

- “The purpose of computation is insight, not numbers” – Richard Hamming
- Explore fundamental relationships between dimensionality, distributions and error rates.
- Understand trivial problems before taking on the harder, but more realistic ones.

Theorem (Onoyama, Sibuya and Tanaka (1983)) Let  $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$  be i.i.d.  $\mathbb{R}^d$ -valued random vectors with pdf  $f(\vec{x}) \in L^2(\mathbb{R}^d)$ , and let

$$Y_n = \text{minimum distance between points} = \min_{1 \leq i < j \leq n} \|\vec{x}_i - \vec{x}_j\|.$$

Then for any  $t > 0$ ,

$$\lim_{n \rightarrow \infty} P(n^2 Y_n^d > t) = e^{-ct} \quad (\text{i.e., } n^2 Y_n^d \rightarrow \text{EXP}(c)),$$

where

$$c = \frac{c_d \|f\|^2}{2} = \frac{c_d}{2} \int_{\mathbb{R}^d} f^2(\vec{x}) d\vec{x}$$

and  $c_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})}$  = volume of the d-dimensional unit ball.



# James and James (2002)



San José State  
UNIVERSITY

- For multivariate Gaussian data, the expected number of collisions is approximately

$$\mu = \frac{1}{2} \frac{n^2 r^d}{\Gamma(1 + \frac{d}{2}) \sqrt{\det \Sigma_{between}}}$$

- Where  $n$  is number of balls,  $r$  is radius,  $d$  is dimensionality,  $\Sigma$  is covariance matrix



# “Intertemplate” Result



- The expected number of pairs for which the centroid separation is less than the radius,  $r$ , is

,

$$\mu = \frac{n^2 r^d}{2^{d+1} \Gamma(1 + \frac{d}{2}) \sqrt{\det \Sigma_{between}}}$$



# Extension to Non-i.i.d. Case



San José State  
UNIVERSITY

- Rotate and rescale by inverse of within-class covariance matrix
- Probability,  $\beta$ , of observation being within  $r$  of class centroid is distributed as  $X^2$  with  $\text{dof} = d$ . So  $\text{FNMR} = 1 - \beta$
- There are  $n^2$  pairings of balls. If number of collisions is small, the probability of any pairing will be in collision is now approx

$$FMR_{\text{INTERTEMPLATE}} = P(i, j \text{ within } r) = \frac{(\chi^2_{1-\text{FNMR}, d})^d}{2^{d+1} \Gamma(1 + \frac{d}{2}) \sqrt{\frac{\prod^d \lambda_{\Sigma \text{between}}}{\prod^d \lambda_{\Sigma \text{within}}}}}$$



# Separability Criteria



San José State  
UNIVERSITY

$$\frac{\prod^d \lambda_{\Sigma_{between}}}{\prod^d \lambda_{\Sigma_{within}}}$$

- One of four separability measures given by Fukunaga (1990), pgs. 446-447
- Strong implications for data fusion problem



# Appealing Result, but.....



- Can eigenvalues be well estimated from limited data?
- Singularity of  $\Sigma_{\text{between}}$  does not mean zero separability
- Result may be heavily dependent upon Gaussian assumption



# Future Work



- Limited training data
- Between-class distribution as mixture of Gaussians
- Within-class distributions as homogeneous in shape, but non-homogeneous in variance
- Ultimate recourse to heuristic methods??