Tutorial on Forensic Voice Comparison and Forensic Acoustics
2nd Pan-American/Iberian Meeting on Acoustics

Geoffrey Stewart Morrison
Daniel Ramos

FORENSIC VOICE COMPARISON LABORATORY
SCHOOL OF ELECTRICAL ENGINEERING & TELECOMMUNICATIONS
UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

ATVS
UAM
UNIVERSIDAD AUTONOMA
DE MADRID

Announcements

Websites

- Tutorial and Special Session on Forensic Voice Comparison and Forensic Acoustics @ 2nd Pan-American/Iberian Meeting on Acoustics

  http://cancun2010.forensic-voice-comparison.net/

- Morrison, G.S. (2010). Forensic voice comparison. In I. Freckelton, & H. Selby (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters.

  http://expert-evidence.forensic-voice-comparison.net/

Half-price pdf between 8 November and 3 December 2010.
When ordering use promotion code: EEF2010

Events

- Tutorial

  Monday 15 November 2010, 7:00 – 9:00 pm

  Coral Kingdom 2/3

- Forensic Acoustics presentations in Signal Processing Session

  Tuesday 16 November 2010, 8:05 – 8:45 am

  Coral Garden 1

- Meeting to organize an ASA Forensic Acoustics Group

  Tuesday 16 November 2010 at 7:30 pm

  Coral Kingdom 1

Petition

Events

- Special Session

  Wednesday 17 November 2010

- Lecture presentations: 8:00 – 11:40 am

  Grand Coral 1A

- Poster presentations: 1:00 – 3:00 pm

  Grand Coral 3

Events

- Special Session Cancellations/Changes

- 3aSC1 Andrzej Drygajlo *Value and interpretation of biometric evidence in forensic automatic speaker recognition*

  will be presented by Michael Jessen

- 3aSC2 Didier Meuwly *Forensic speaker recognition: Comparison and validation of automatic systems over 3 generations*

  will be replaced by

  Geoffrey Stewart Morrison *Calibration and fusion*

- 3pSC9 Alejandro Wang *Forensic voice comparison based on nasal formants*

  is cancelled

Introduction

Geoffrey Stewart Morrison
Director
Forensic Voice Comparison Laboratory
School of Electrical Engineering & Telecommunications
University of New South Wales
Sydney, Australia

Invited Lectures
Judicial Phonetics Specialization
Master of Phonetics and Phonology Program
Consejo Superior de Investigaciones Científicas
/ Universidad Internacional Menéndez Pelayo
Madrid, Spain

Daniel Ramos
Assistant Professor
Biometric Recognition Group
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Spain

Requirements for Forensic-Comparison Science:

- Objective
- Replicable
- Demonstrated validity and reliability
- Logically correct

The Likelihood-Ratio Framework for the Evaluation of Evidence

Calculating a Forensic Likelihood Ratio

Testing the Validity of a Forensic-Comparison System

Basic Architecture of an Automatic Speaker Recognition System

Spectral Feature Extraction

Gaussian Mixture Model – Universal Background Model (GMM-UBM)

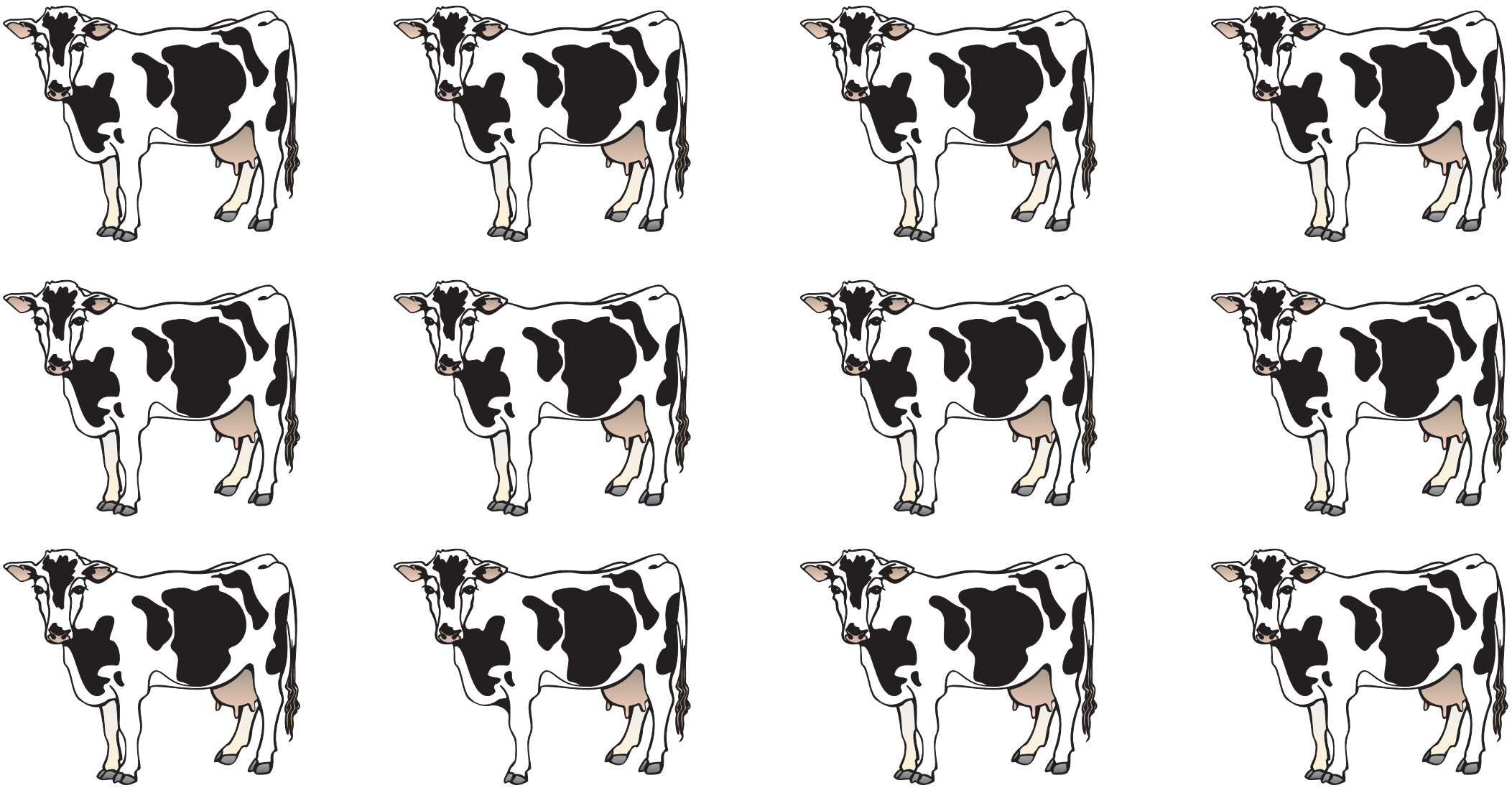
GMM Supervectors and SVM-SM Systems

Session- and Channel-Mismatch Compensation

The Likelihood-Ratio Framework for the Evaluation of Evidence

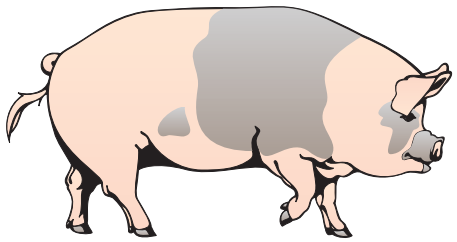
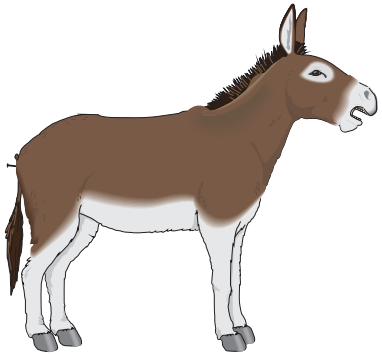
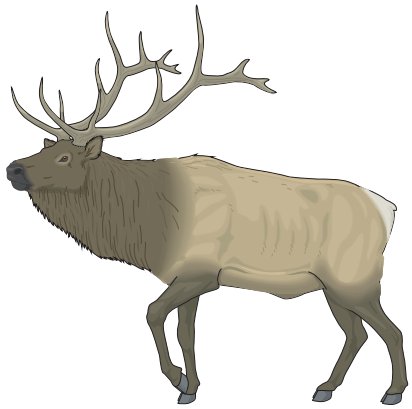
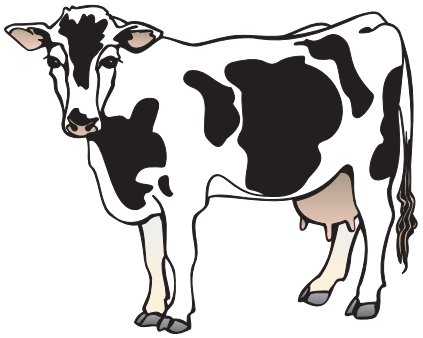
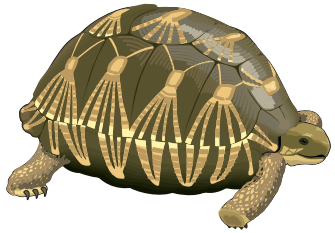
Given that it is a cow, what is the probability of it having four legs?

$$p(4 \text{ legs} \mid \text{cow}) = ?$$

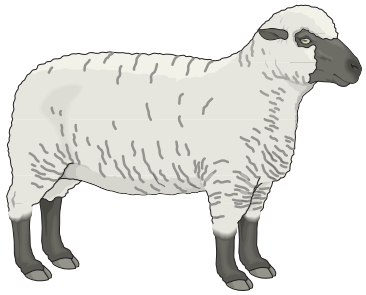
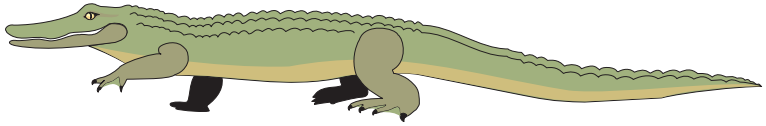
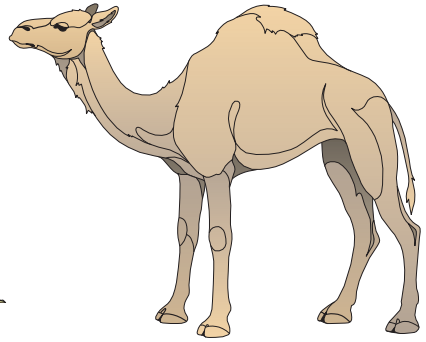


Given that it has four legs, what is the probability that it is a cow?

$$p(\text{ cow} \mid 4 \text{ legs }) = ?$$

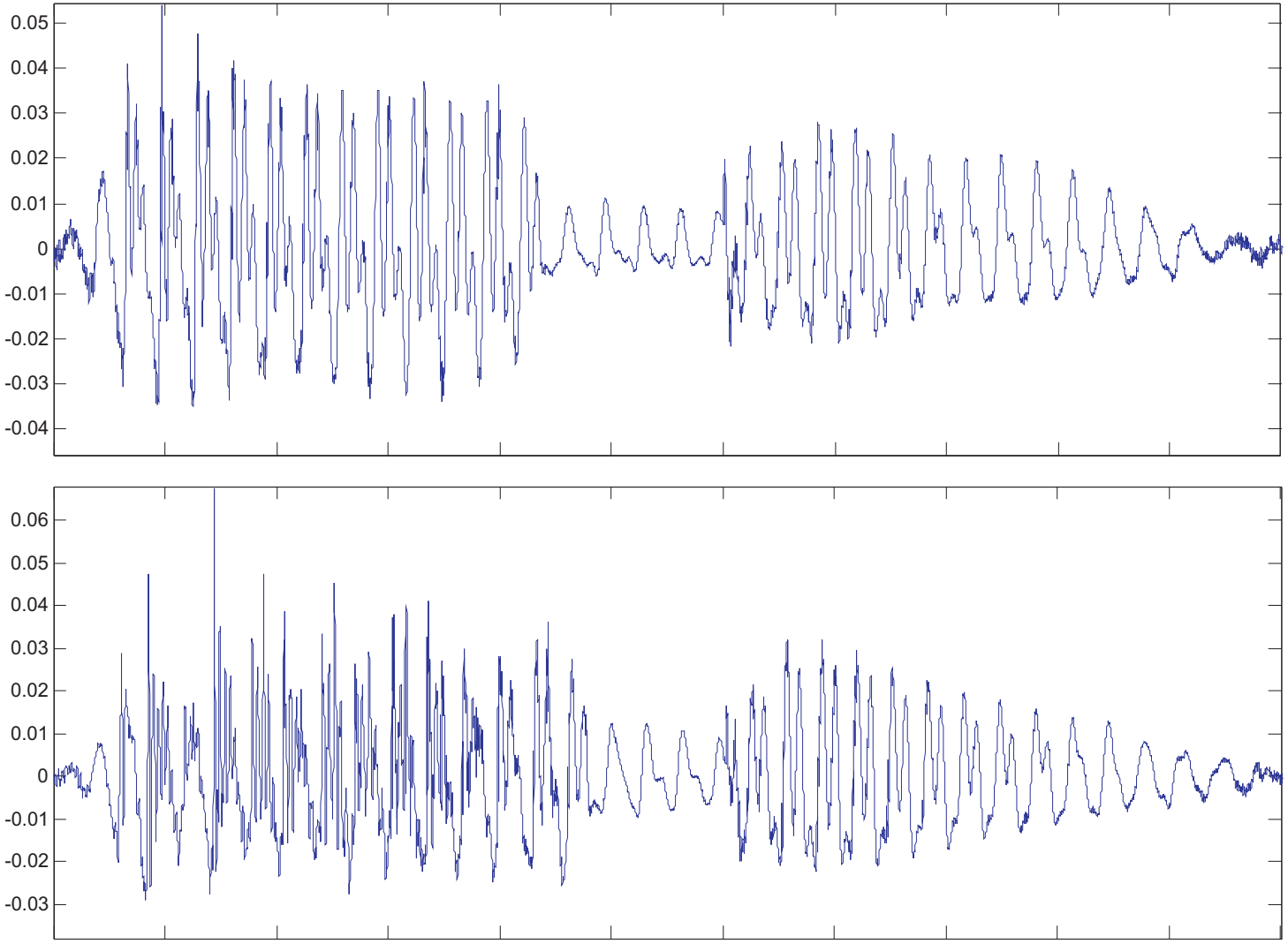








Given two voice samples with acoustic properties x_1 and x_2,
what is the probability that they were produced by the same speaker?

$$p(\text{ same speaker } | \text{ acoustic properties } x_1, x_2\ ) = ?$$



$$p(\text{ same speaker } | \text{ acoustic properties } x_1, x_2) = ?$$

$$p(\text{ cow } | \, x \text{ legs }) = ?$$

Missing information!

Bayes' Theorem:

posterior probability

$$p(\text{ same speaker } | \text{ acoustic properties } x_1, x_2)$$

$$=$$

$$\frac{p(\text{ acoustic difference } x \mid \text{ same speaker }) \times p(\text{ same speaker })}{\left(\begin{array}{c} p(\text{ acoustic properties } x_1, x_2 \mid \text{ different speaker }) \times p(\text{ different speaker }) \\ + p(\text{ acoustic properties } x_1, x_2 \mid \text{ same speaker }) \times p(\text{ same speaker }) \end{array}\right)}$$

conditional probability

prior probability

Bayes' Theorem:

posterior odds

$$\frac{p(\text{ same speaker } | \text{ acoustic properties } x_1, x_2)}{p(\text{ different speaker } | \text{ acoustic properties } x_1, x_2)}$$

$$=$$

$$\frac{p(\text{ acoustic properties } x_1, x_2 | \text{ same speaker })}{p(\text{ acoustic properties } x_1, x_2 | \text{ different speaker })} \times \frac{p(\text{ same speaker })}{p(\text{ different speaker })}$$

likelihood ratio

prior odds

¡¡¡ However !!!

The forensic scientist acting as an expert witness canNOT give the posterior probability. They canNOT give the probability that two speech samples were produced by the same speaker.

Why not?

- The forensic scientist does not know the priors.
- Determining the probability of guilt (same speaker) is the task of the judge or jury, not the forensic scientist.
- The task of the forensic scientist is to present the *strength of evidence* which can be extracted from the speech samples.

posterior odds

$$\frac{p(\text{ same speaker } | \text{ acoustic properties } x_1, x_2)}{p(\text{ different speaker } | \text{ acoustic properties } x_1, x_2)}$$

$$= \frac{p(\text{ acoustic properties } x_1, x_2 | \text{ same speaker })}{p(\text{ acoustic properties } x_1, x_2 | \text{ different speaker })} \times \frac{p(\text{ same speaker })}{p(\text{ different speaker })}$$

likelihood ratio

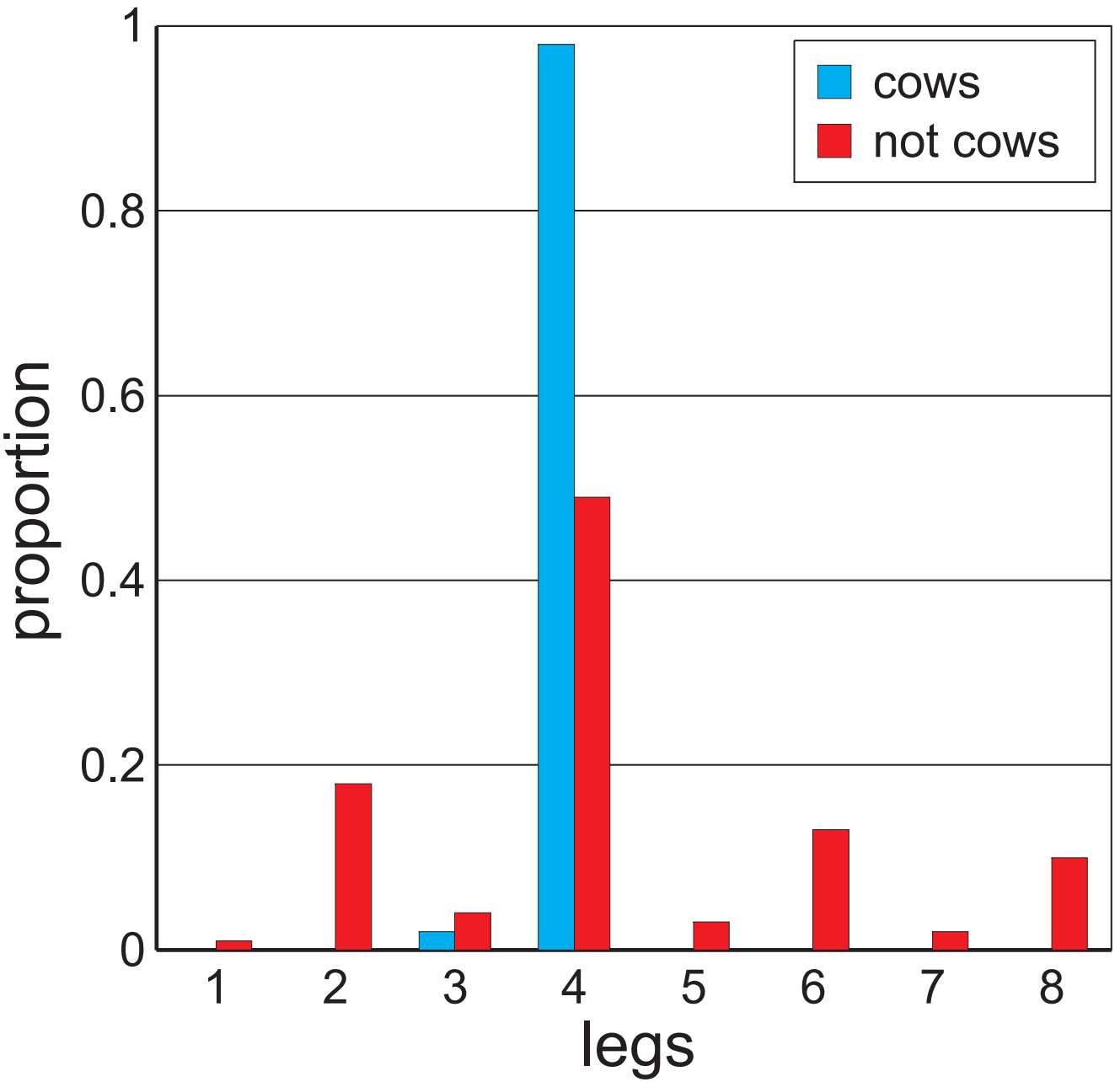
prior odds

Calculating a Forensic Likelihood Ratio

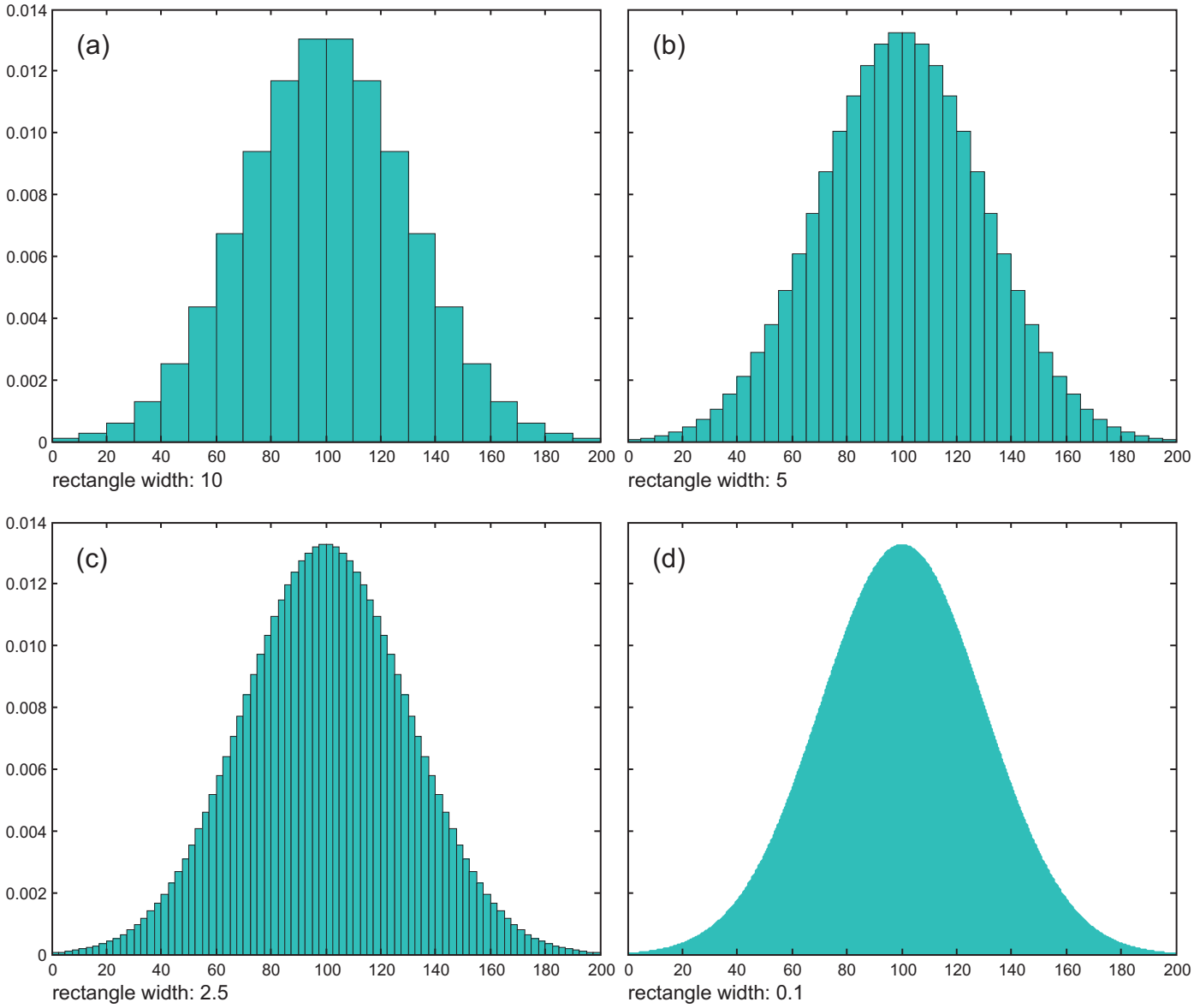
Likelihood Ratio:

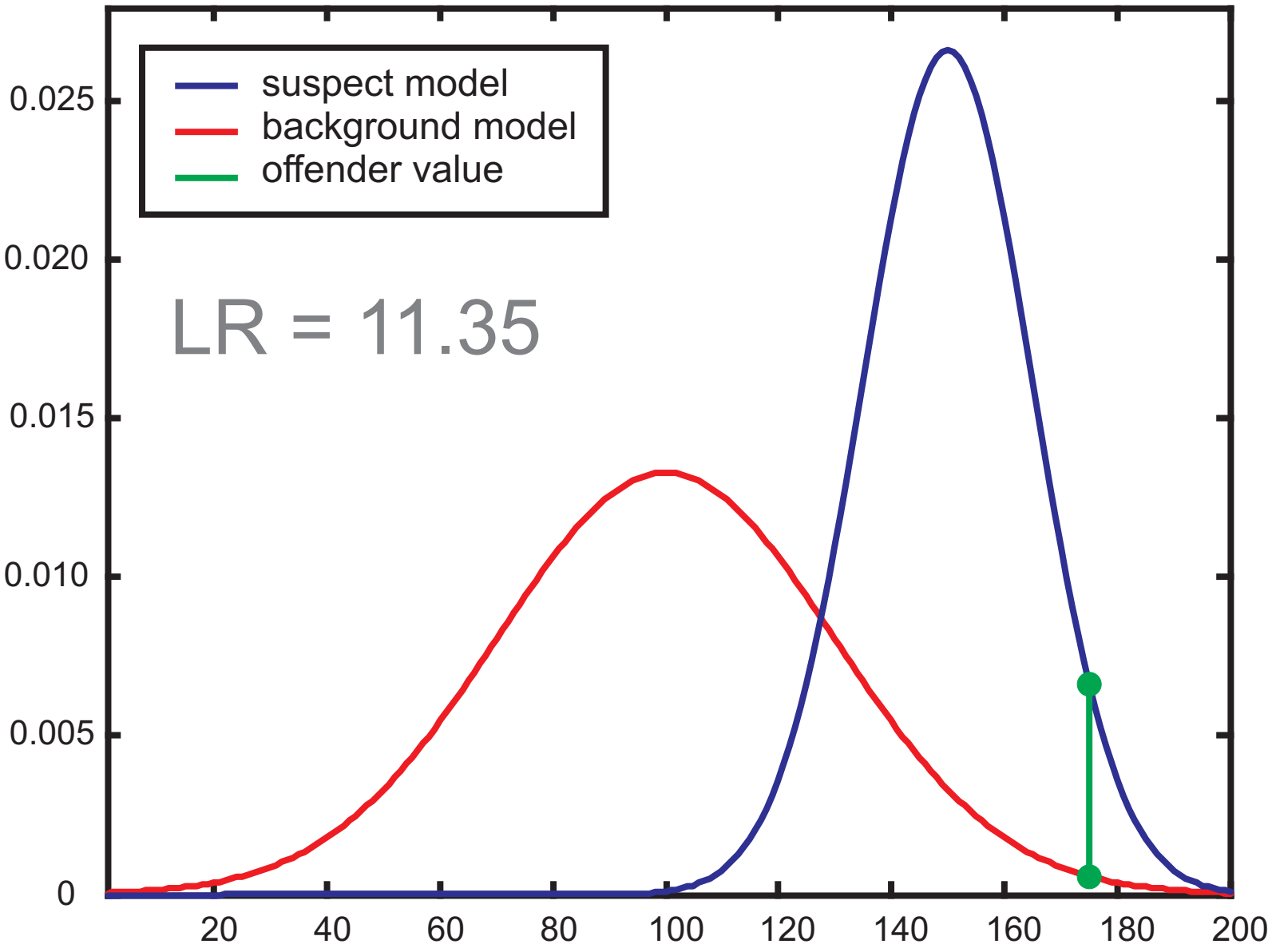
$$\frac{p(\text{ acoustic properties } x_1, x_2 \mid \text{same speaker })}{p(\text{ acoustic properties } x_1, x_2 \mid \text{different speaker })}$$

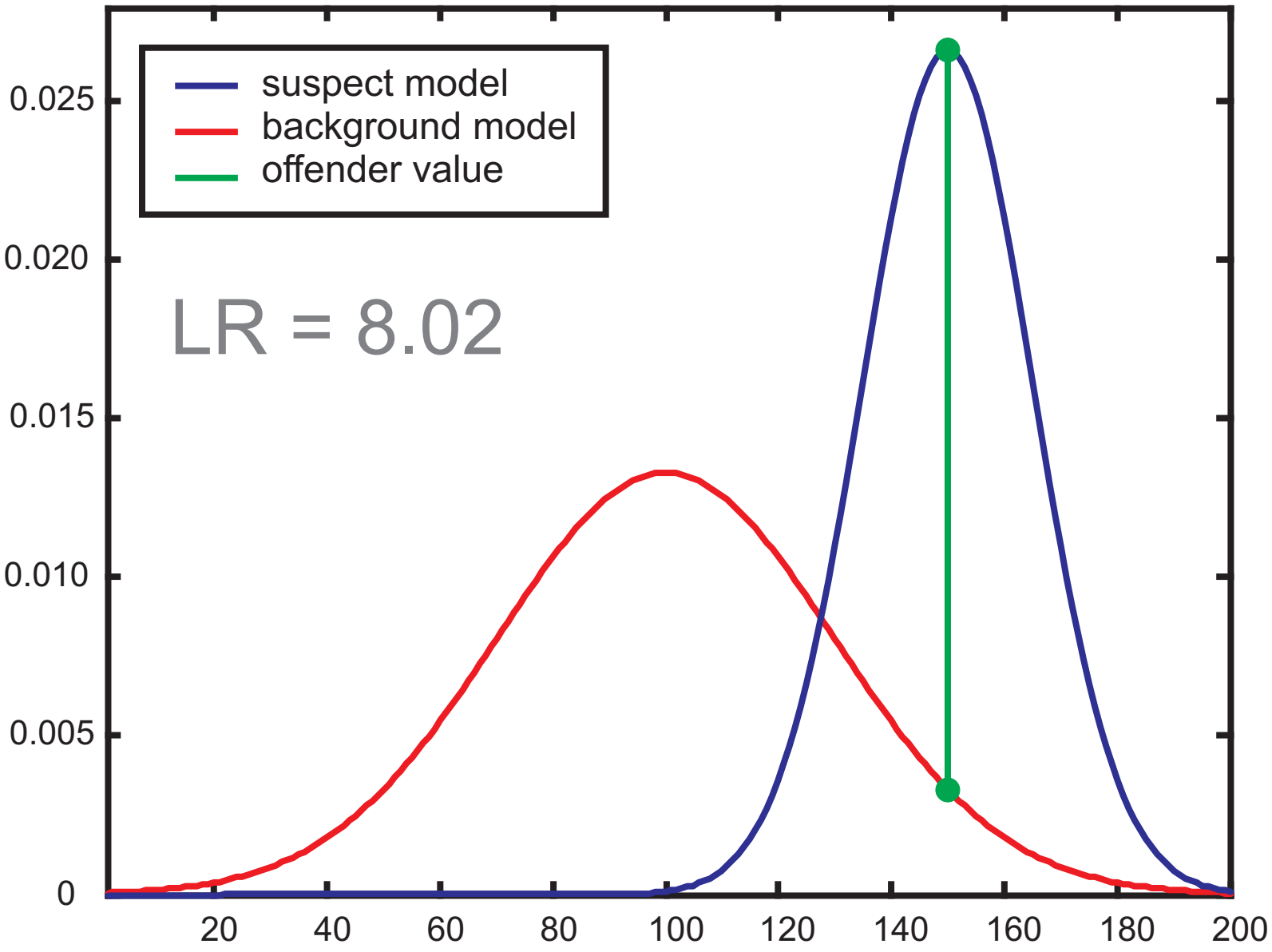
$$\frac{p(\text{ } x \text{ legs} \mid \text{cow })}{p(\text{ } x \text{ legs} \mid \text{not a cow })}$$

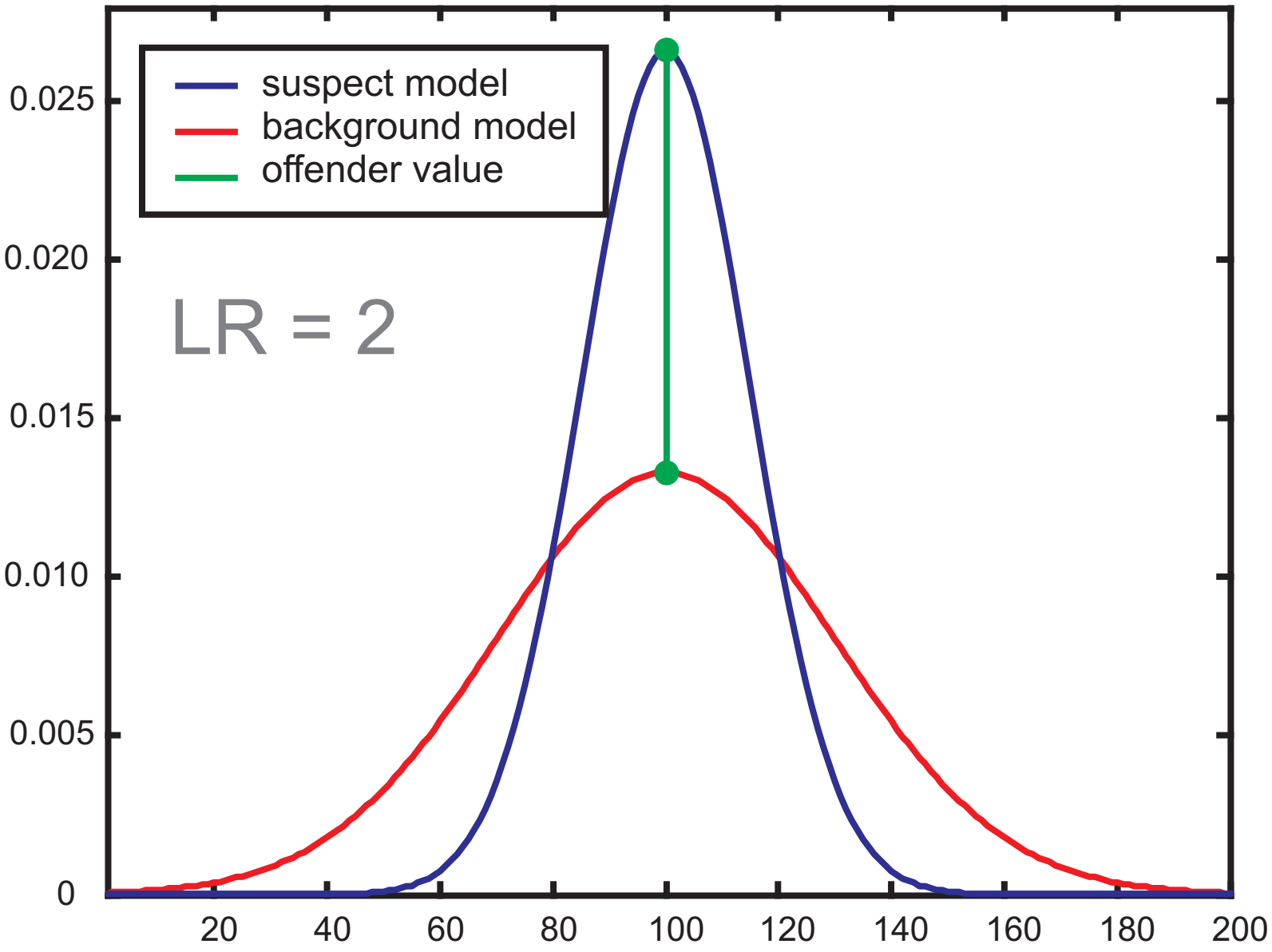

cows
not cows
proportion
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
8
legs

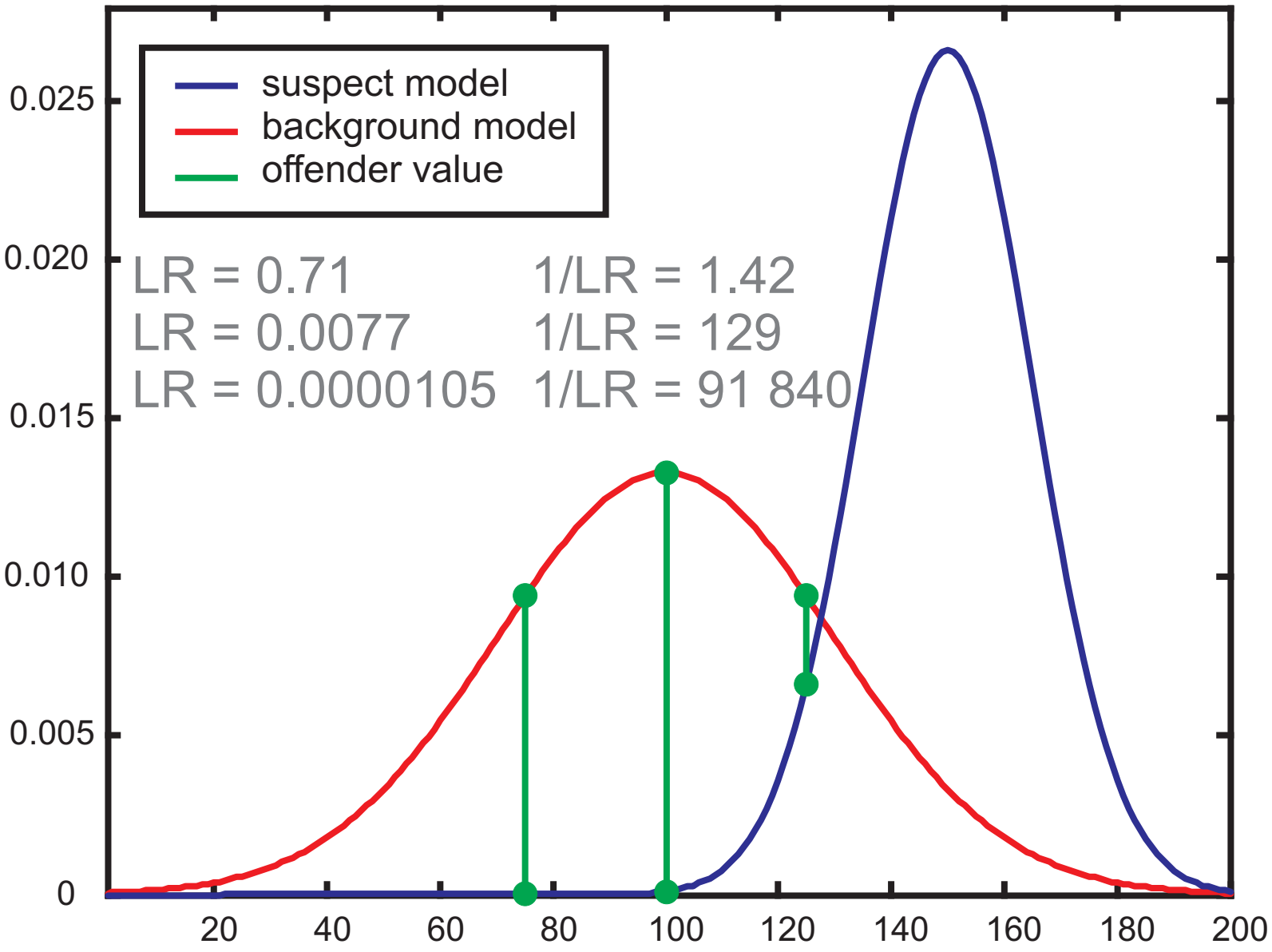
For continuous data rather than histograms, probability density functions (PDFs) must be used.

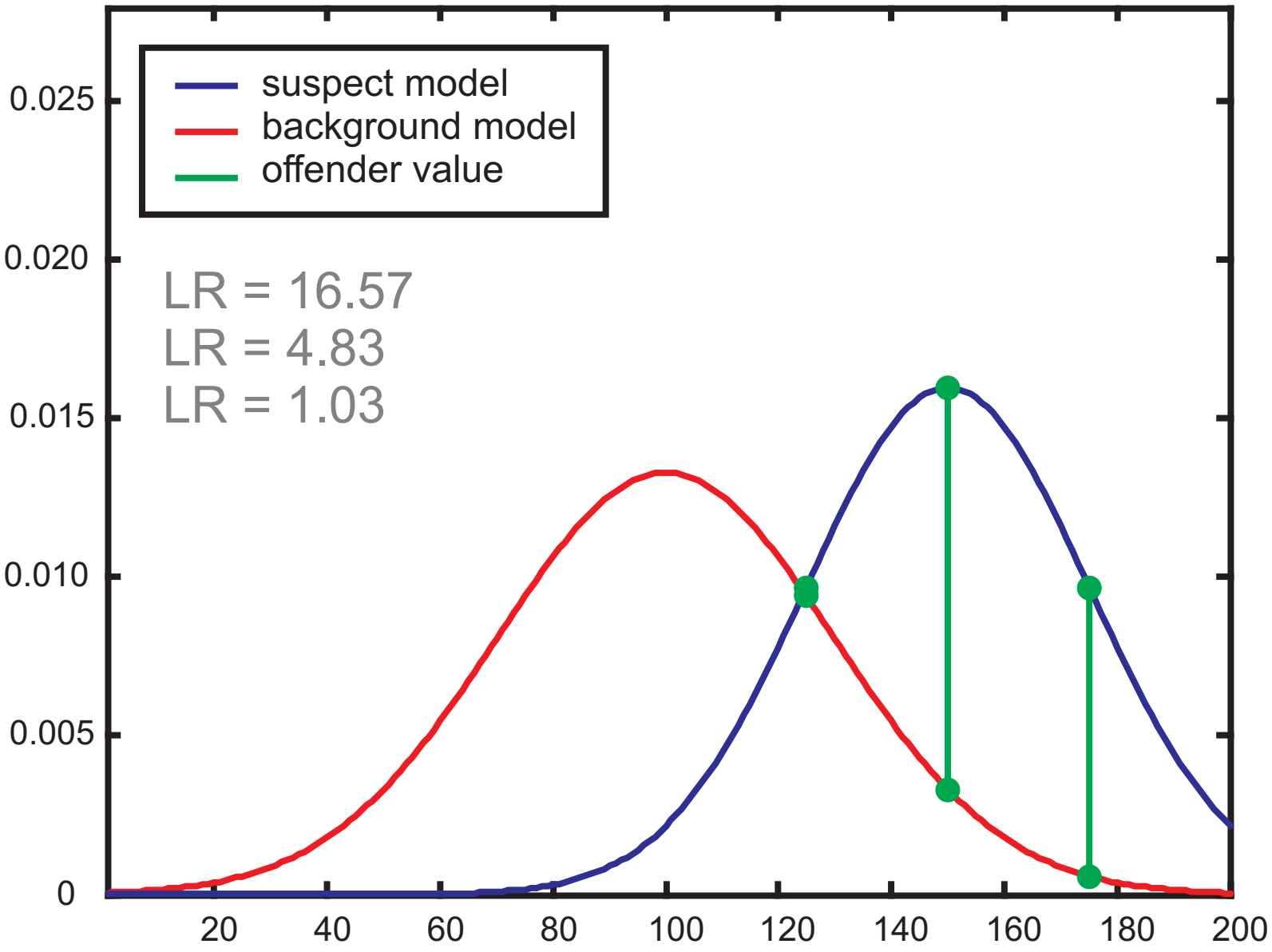



suspect model
background model
offender value
LR = 11.35
0.025
0.020
0.015
0.010
0.005
0
20
40
60
80
100
120
140
160
180
200

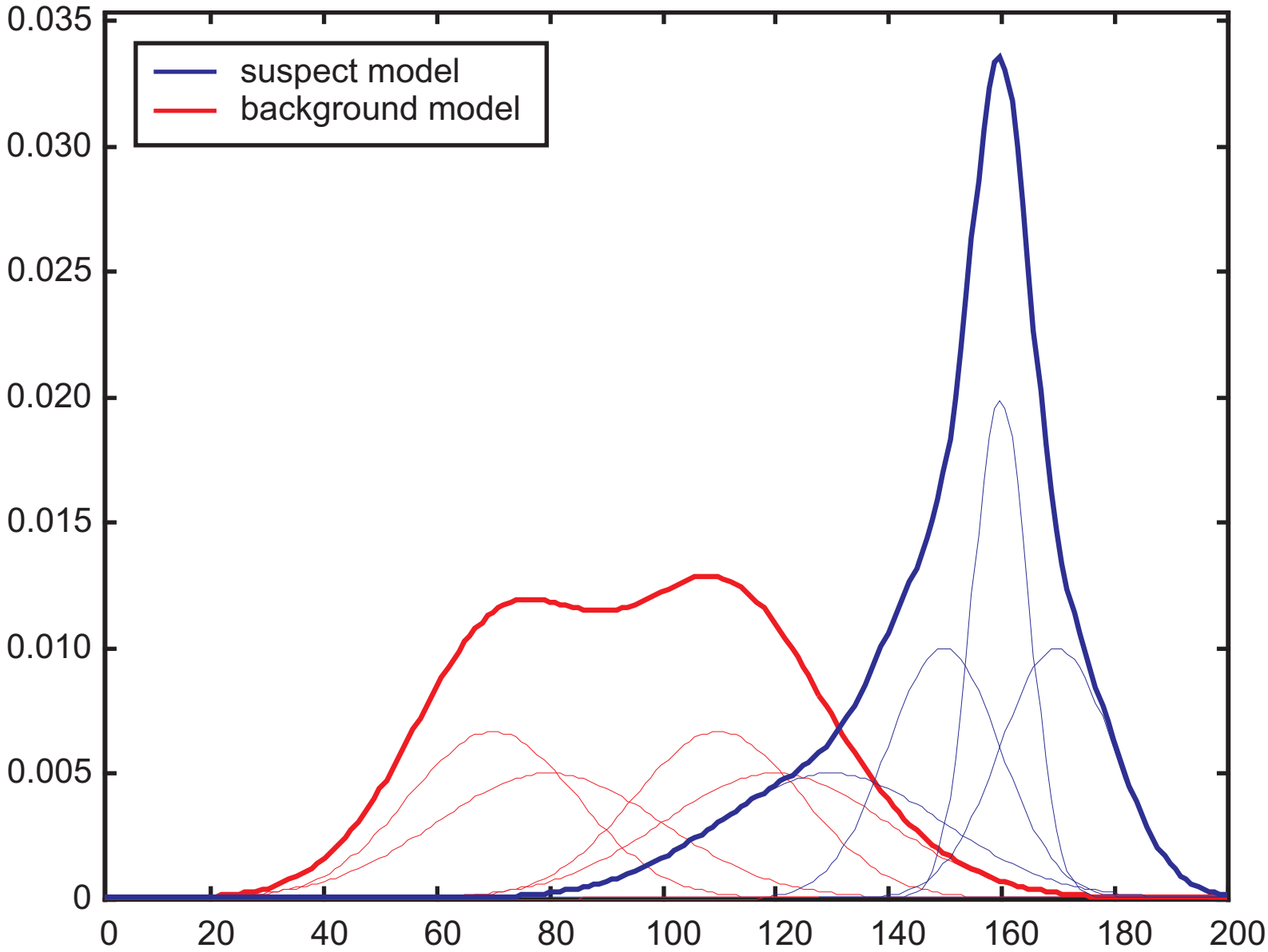

suspect model
background model
offender value
LR = 8.02
0.025
0.020
0.015
0.010
0.005
0
20
40
60
80
100
120
140
160
180
200


suspect model
background model
offender value
LR = 2
0.025
0.020
0.015
0.010
0.005
0
20
40
60
80
100
120
140
160
180
200

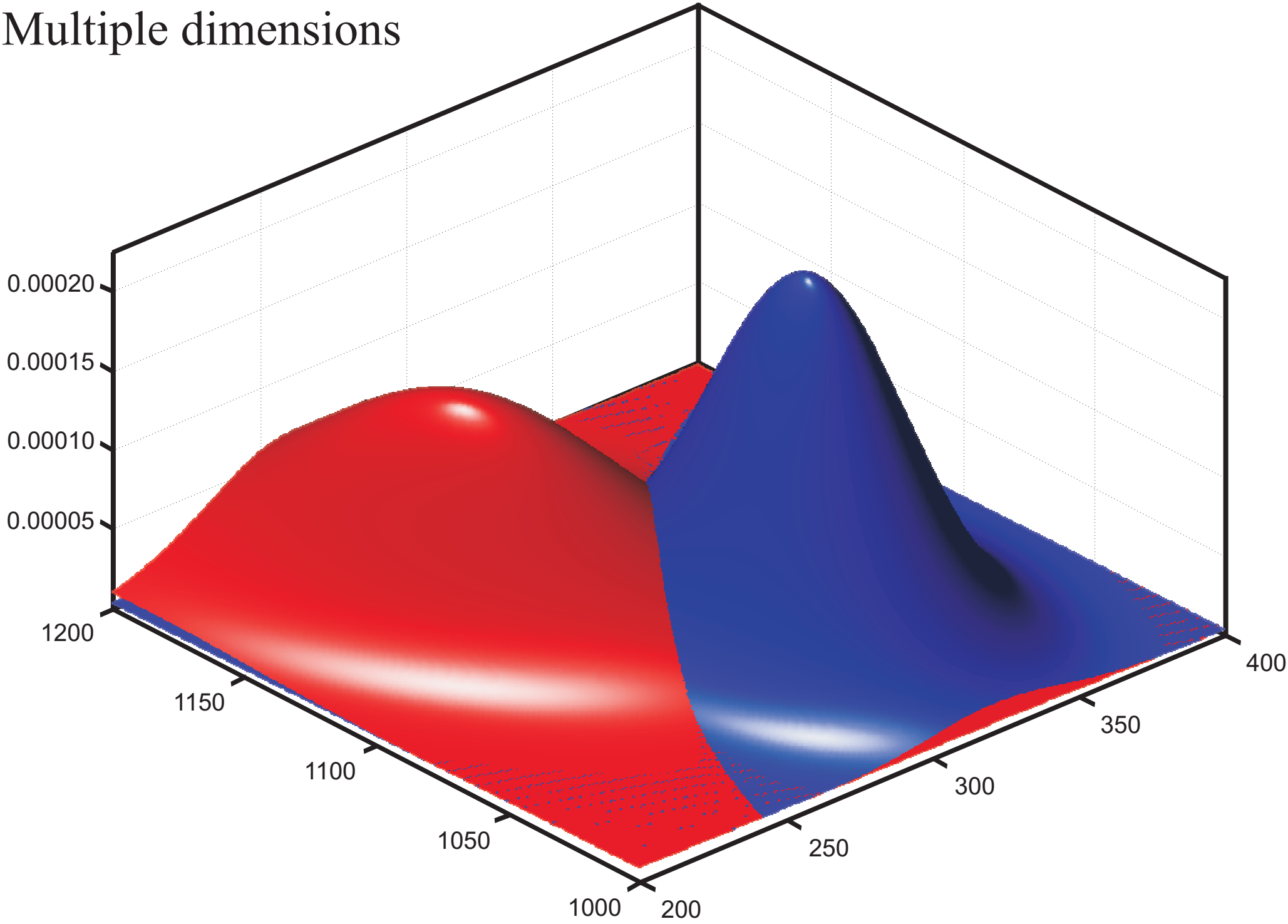

suspect model
background model
offender value
LR = 0.71
1/LR = 1.42
LR = 0.0077
1/LR = 129
LR = 0.0000105
1/LR = 91 840
0.025
0.020
0.015
0.010
0.005
0
20
40
60
80
100
120
140
160
180
200


suspect model
background model
offender value
LR = 16.57
LR = 4.83
LR = 1.03
0.025
0.020
0.015
0.010
0.005
0
20
40
60
80
100
120
140
160
180
200

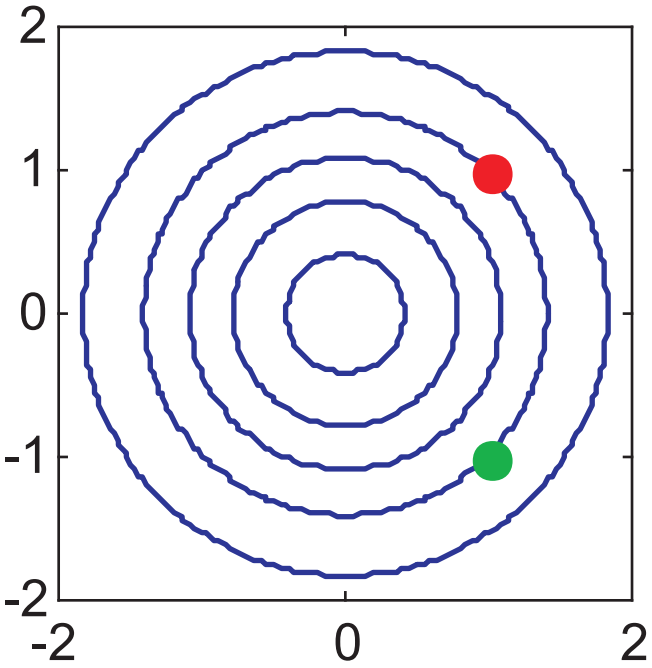
Gaussian Mixture Models (GMMs)

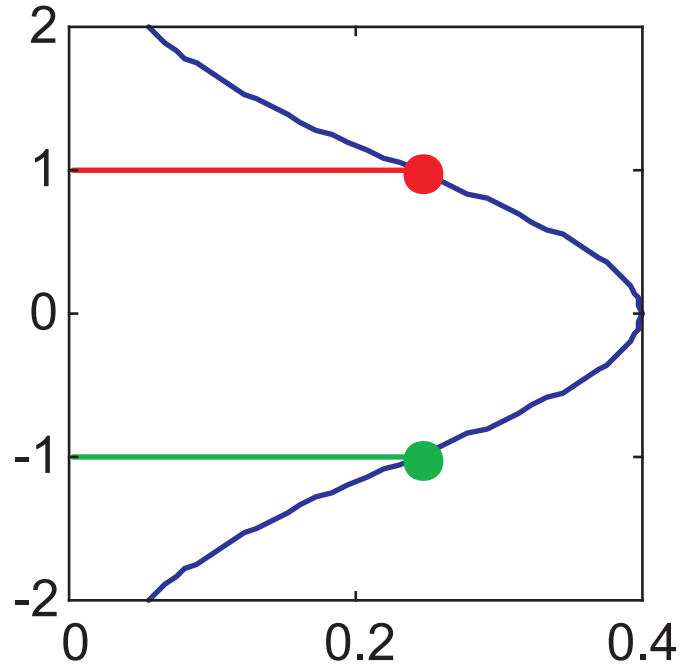


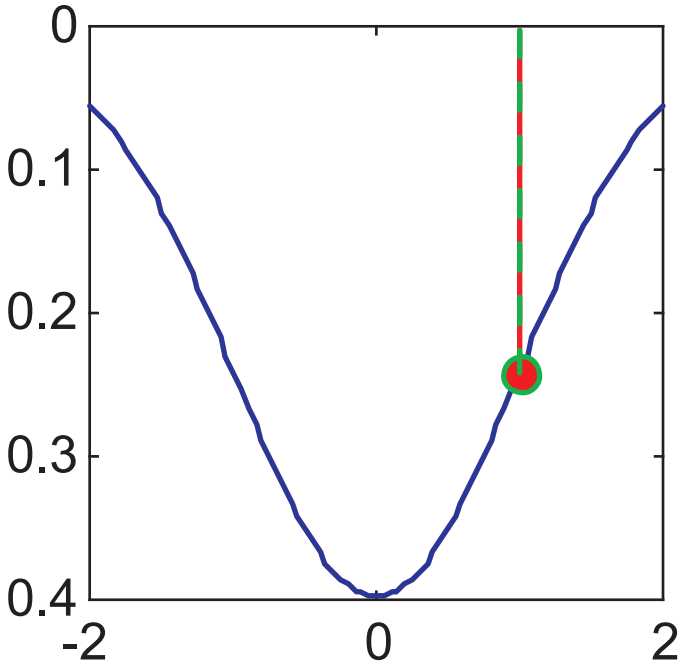
Multiple dimensions



no correlation



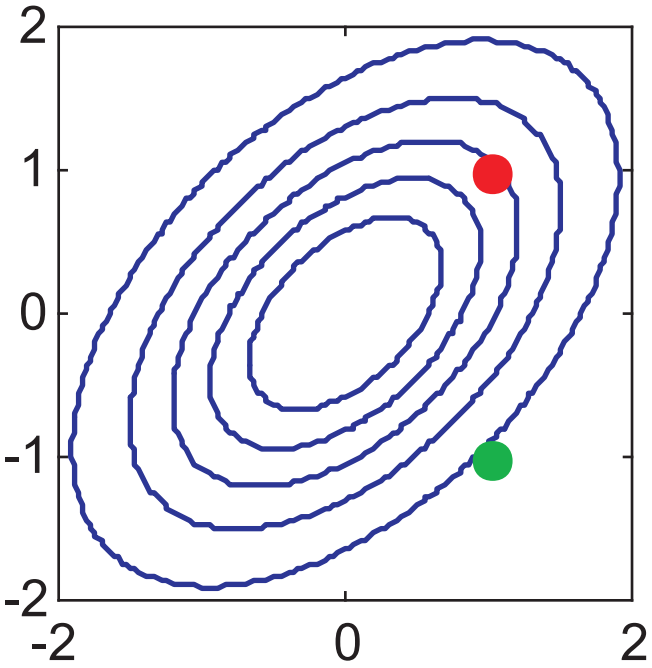


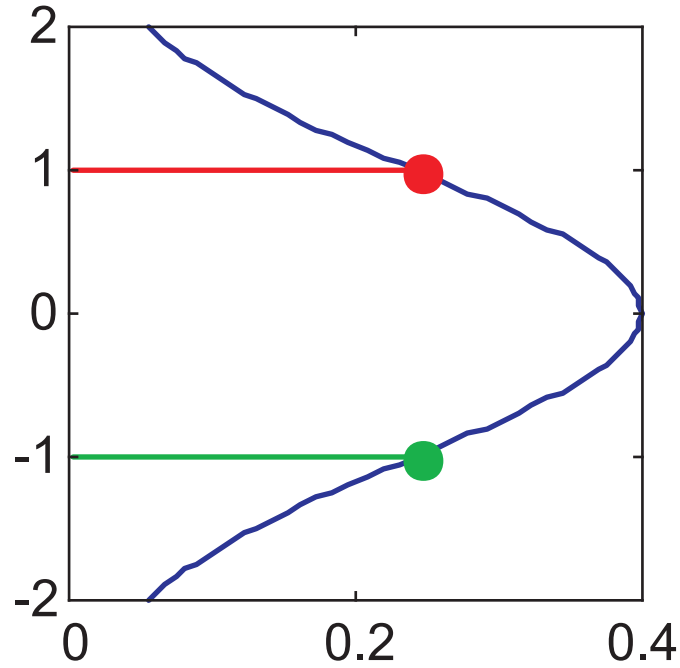


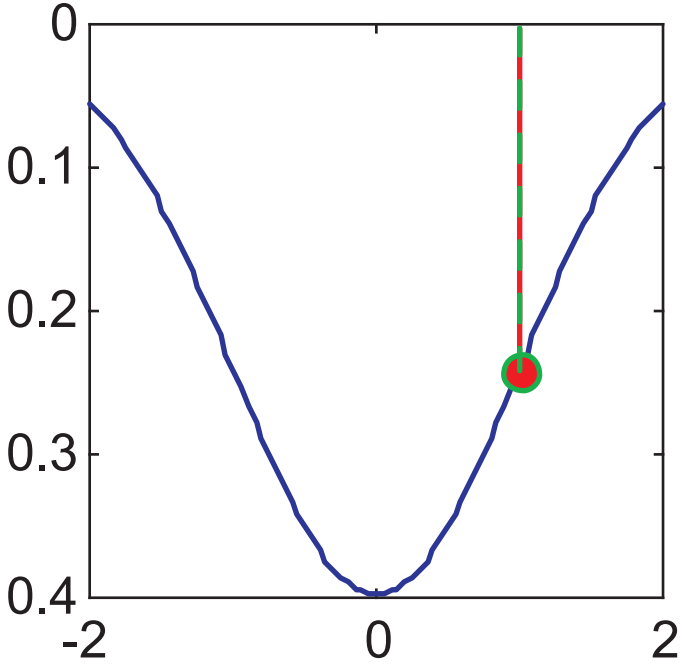
Probe A:
univariate x pdf = 0.242
univariate y pdf = 0.242
naïve Bayes pdf (x*y) = 0.059
bivariate pdf = 0.059

Probe B:
univariate x pdf = 0.242
univariate y pdf = 0.242
naïve Bayes pdf (x*y) = 0.059
bivariate pdf = 0.059

correlation







Probe A:
univariate x pdf = 0.242
univariate y pdf = 0.242
naïve Bayes pdf (x*y) = 0.059
bivariate pdf = 0.094

Probe B:
univariate x pdf = 0.242
univariate y pdf = 0.242
naïve Bayes pdf (x*y) = 0.059
bivariate pdf = 0.025

Testing the Validity of a Forensic-Comparison System

Measuring Validity

- Test set consisting of a large number of pairs known to be same origin and a large number of pairs known to be different origin
- Use forensic-comparison system to calculate LR for each pair
- Compare output with knowledge about input

Measuring Validity

- Correct-classification / classification-error rate is not appropriate
  - based on posterior probabilities
  - hard threshold rather than gradient

fact	decision: same	decision: different
same	correct acceptance	incorrect rejection
different	incorrect acceptance	correct rejection

Measuring Validity

- Goodness is **extent** to which LRs from same-origin pairs > 1, and LRs from different-origin pairs < 1

- A metric which captures the gradient goodness of a set of likelihood ratios derived from test data is the log-likelihood-ratio cost, C_{llr}

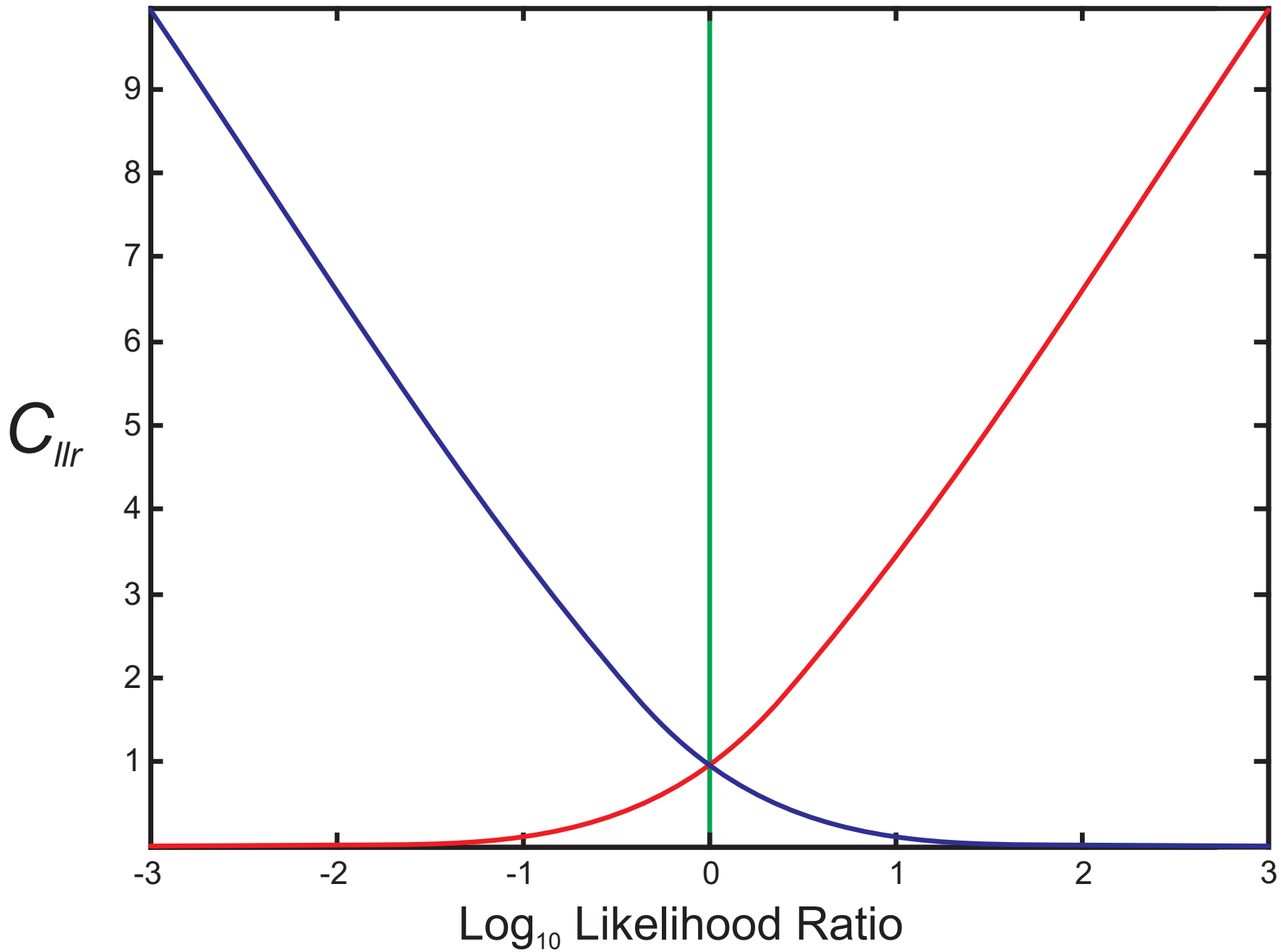
Measuring Validity

- Goodness is **extent** to which LRs from same-origin pairs > 1, and LRs from different-origin pairs < 1

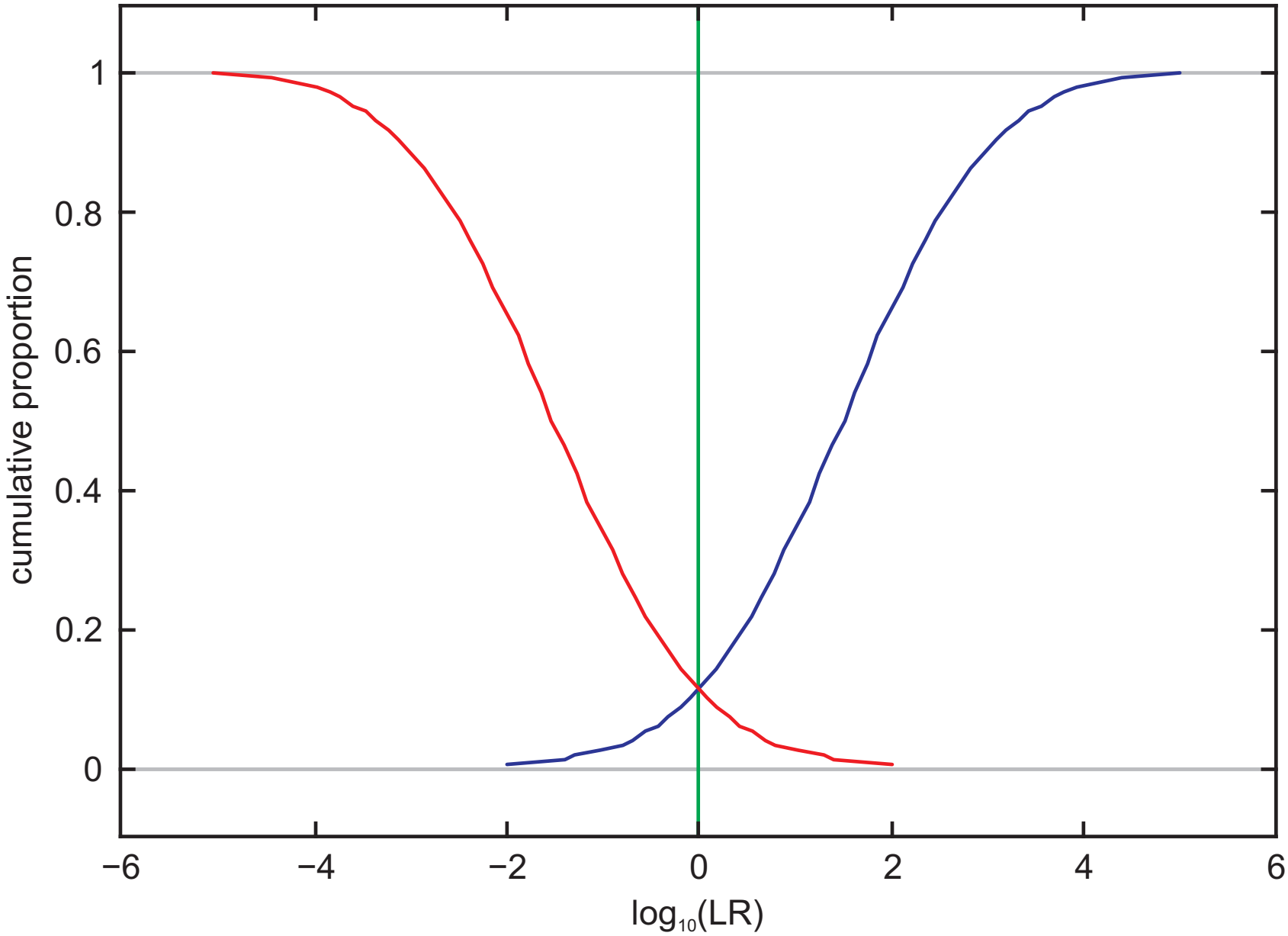
- Goodness is **extent** to which log(LR)s from same-origin pairs > 0, and log(LR)s from different-origin pairs < 0

LR						
1/1000	**1/100**	**1/10**	**1**	**10**	**100**	**1000**
-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
$\log_{10}$(LR)						

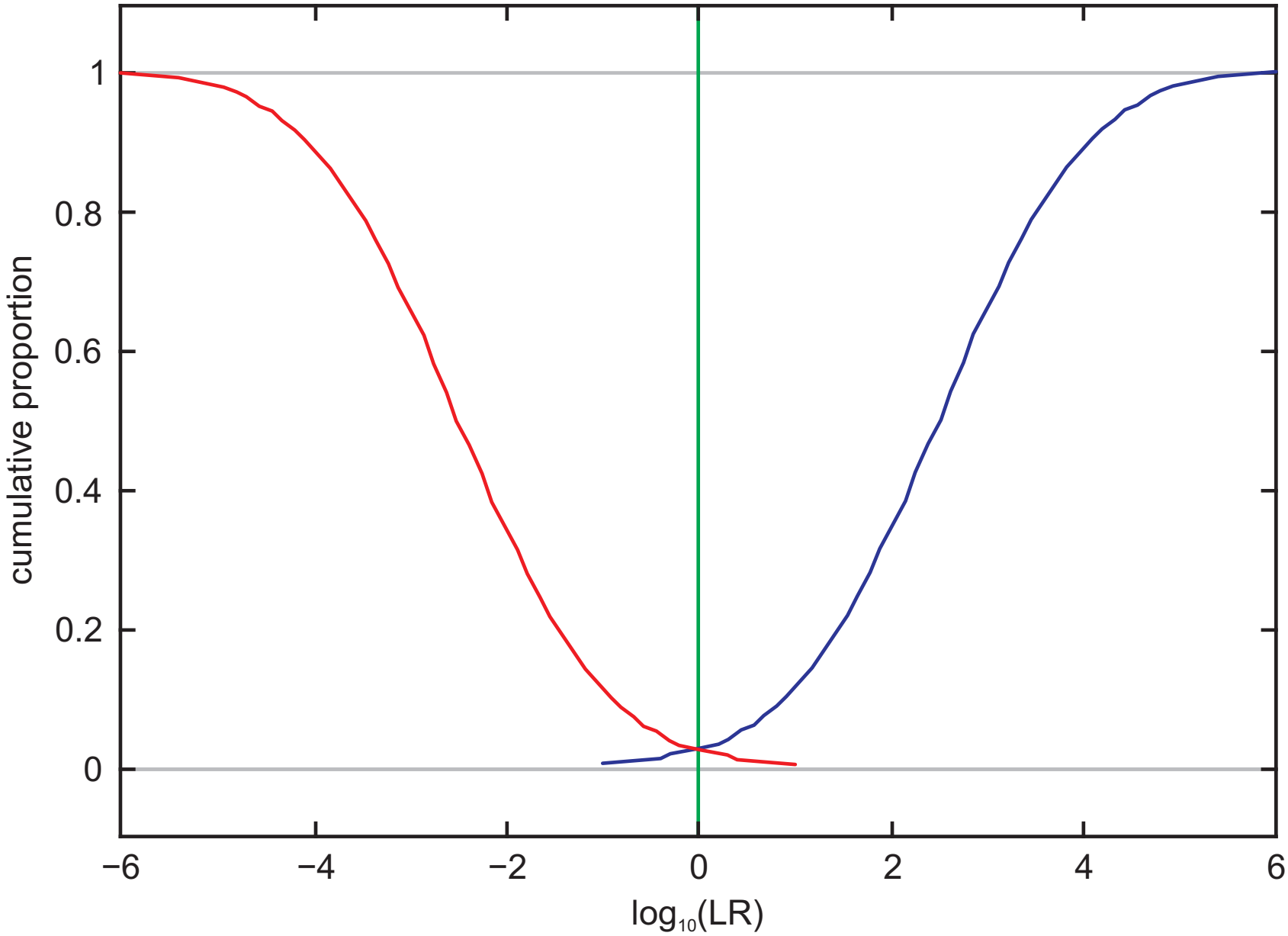
$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}} \sum_{i=1}^{N_{ss}} \log_2 \left(1 + \frac{1}{LR_{ss_i}} \right) + \frac{1}{N_{ds}} \sum_{j=1}^{N_{ds}} \log_2 \left(1 + LR_{ds_j} \right) \right)$$


C_{llr}
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log_{10} Likelihood Ratio

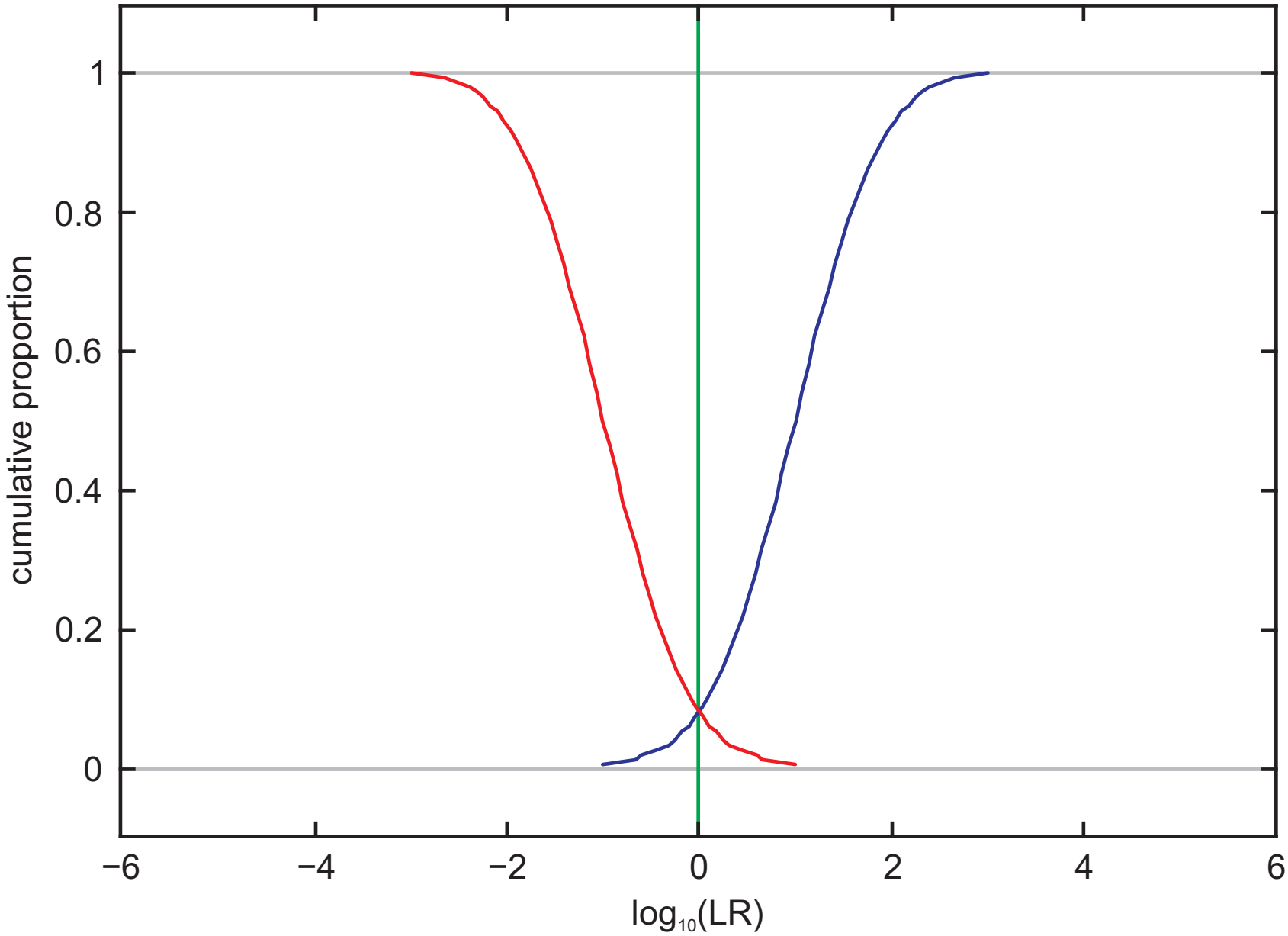
Tippett Plots



Tippett Plots



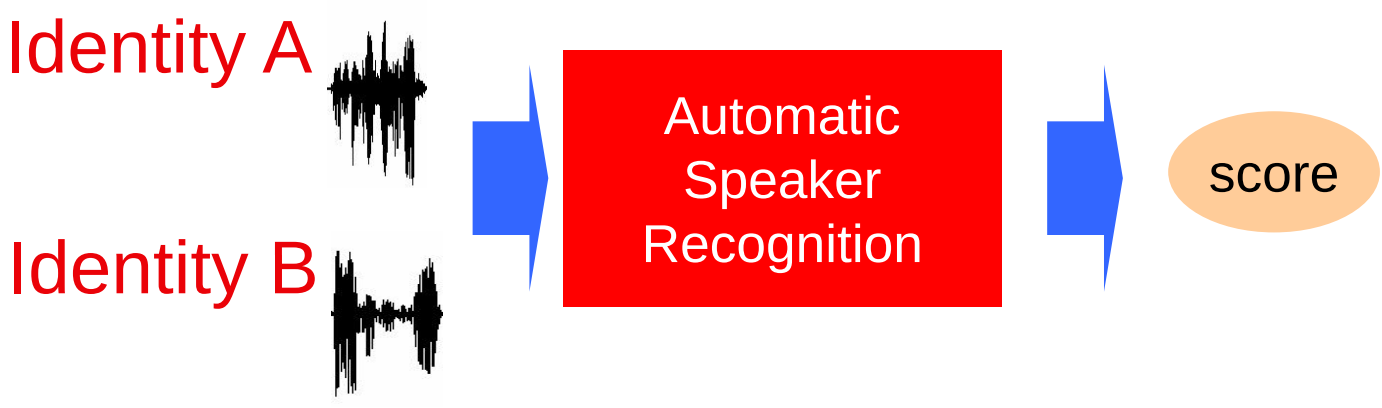
Tippett Plots



Basic Architecture of an Automatic Speaker Recognition System

Similarity: score

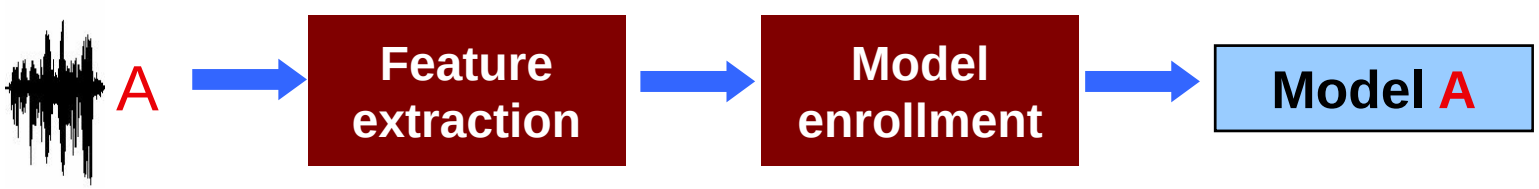
- Wide majority of systems generate scores
- Similarity of the identity in two speech utterances



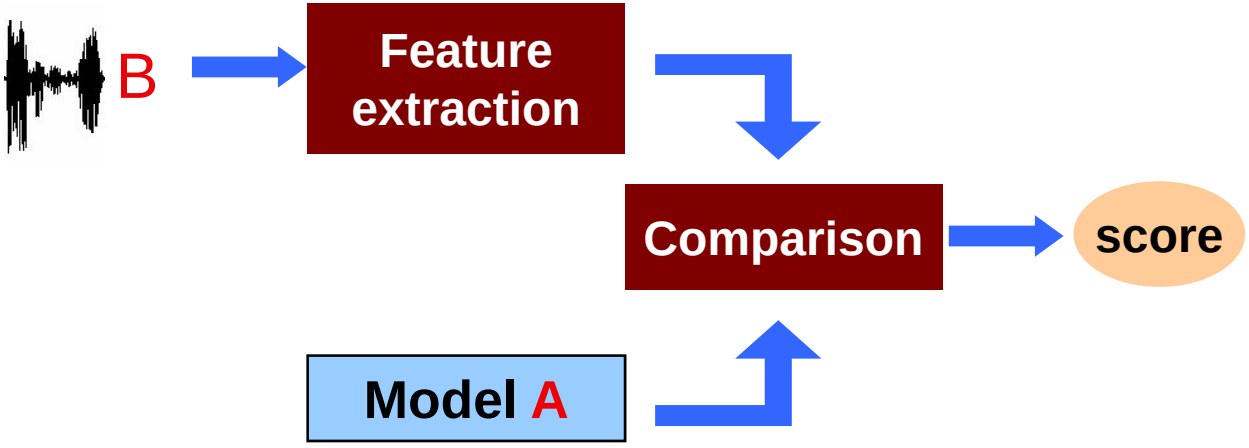
  - Ideally:
    - A and B are the same: higher score
    - A and B are different: lower score
  - It is possible to discriminate
    - Same-ID cases should yield higher scores than different-id cases

GMM-UBM: stages

- Enrollment



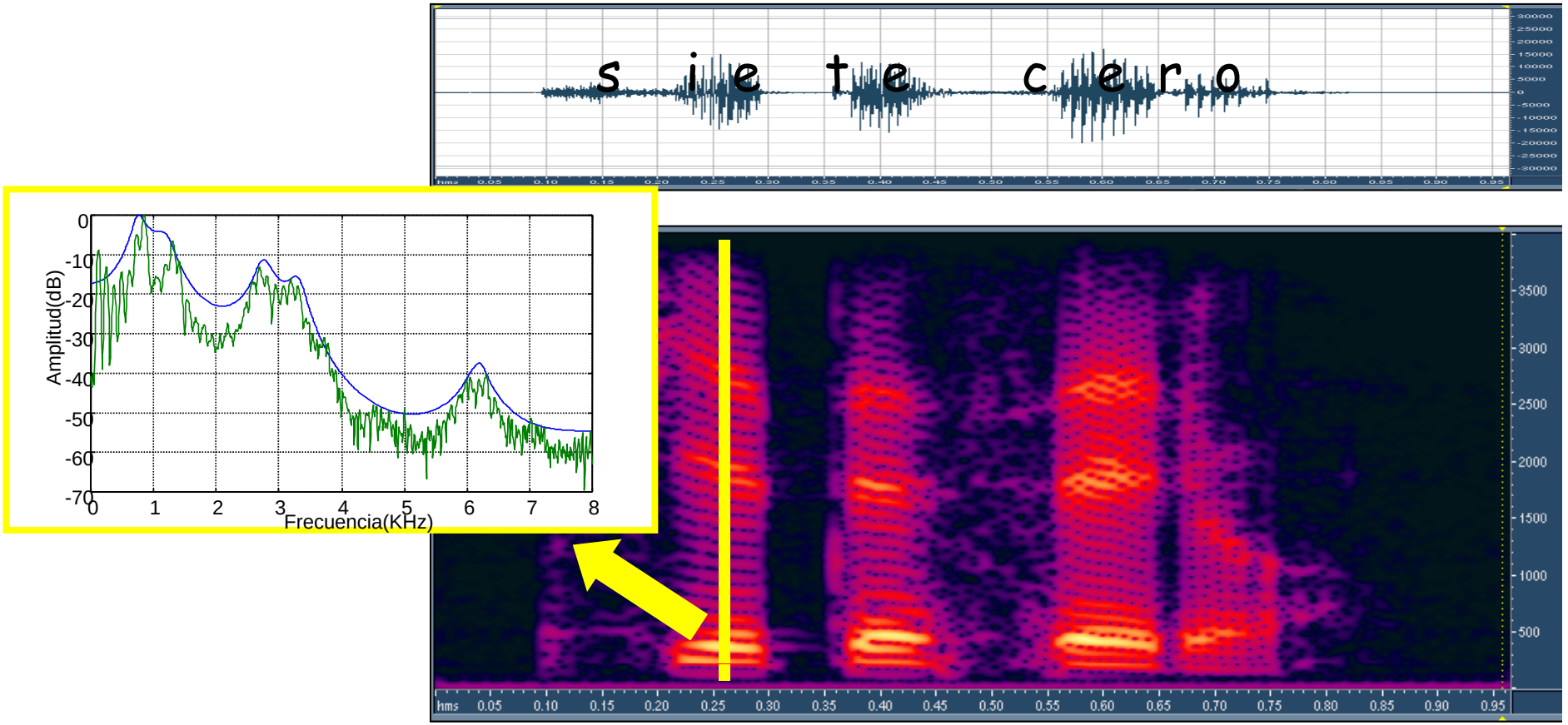
- Score computation



Spectral Feature Extraction

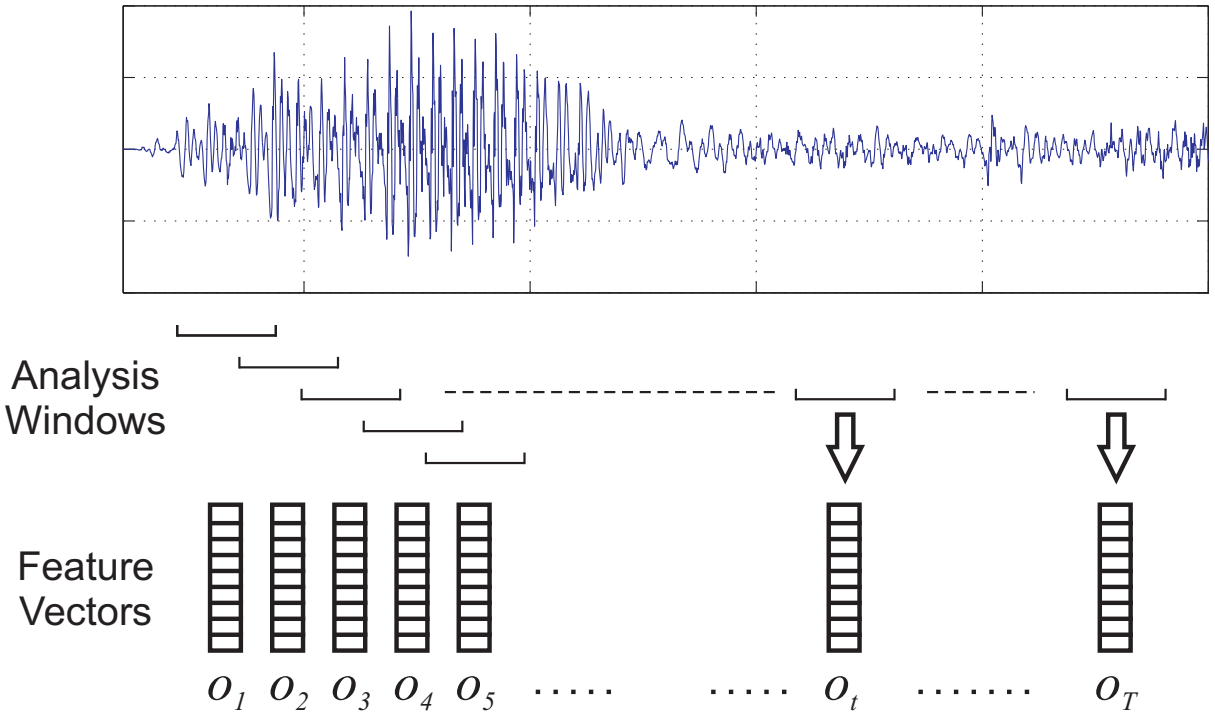
Speech spectrum

- Spectral systems extract information from the speech spectrum
  - And its variation across time



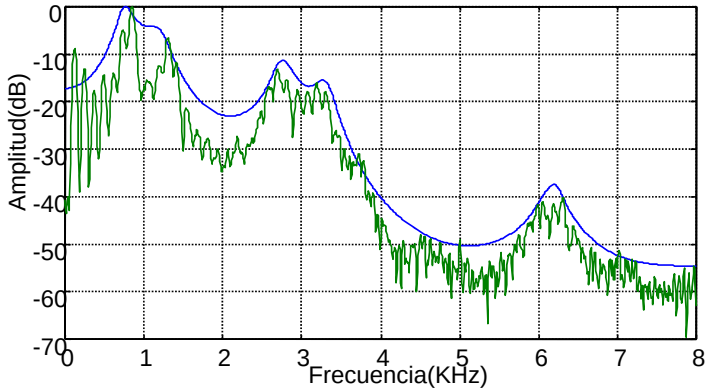
Spectral feature extraction

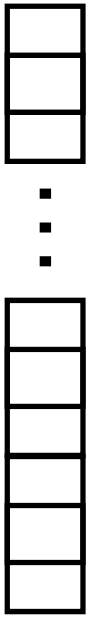
- First stage: windowing
  - T short-term speech frames
- Second stage: feature extraction
  - Each frame is represented by a vector of D numbers



Feature extraction

- Objective: obtaining discriminating information from a speech frame
  - The envelope is particular of the speaker

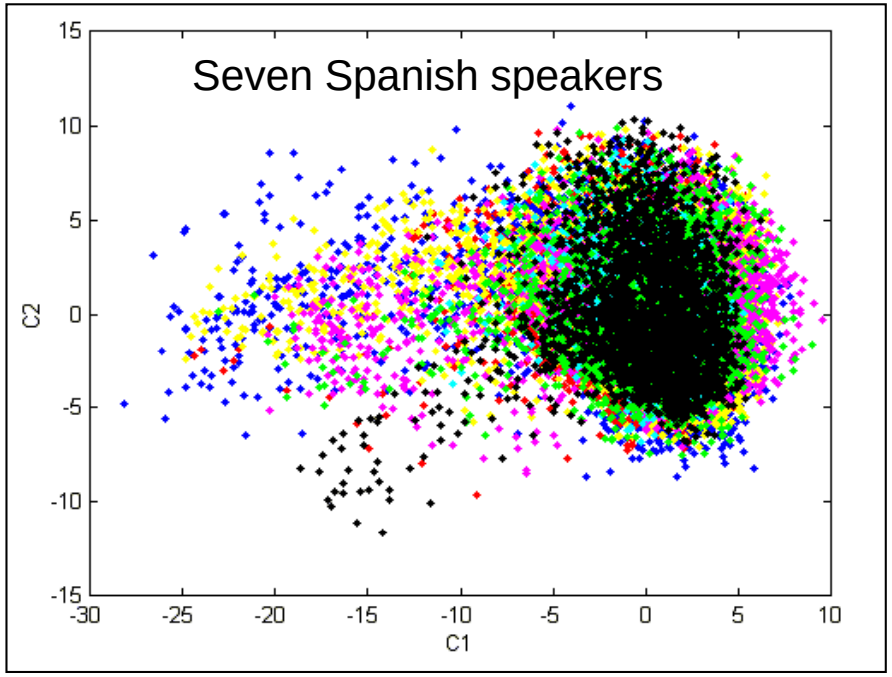


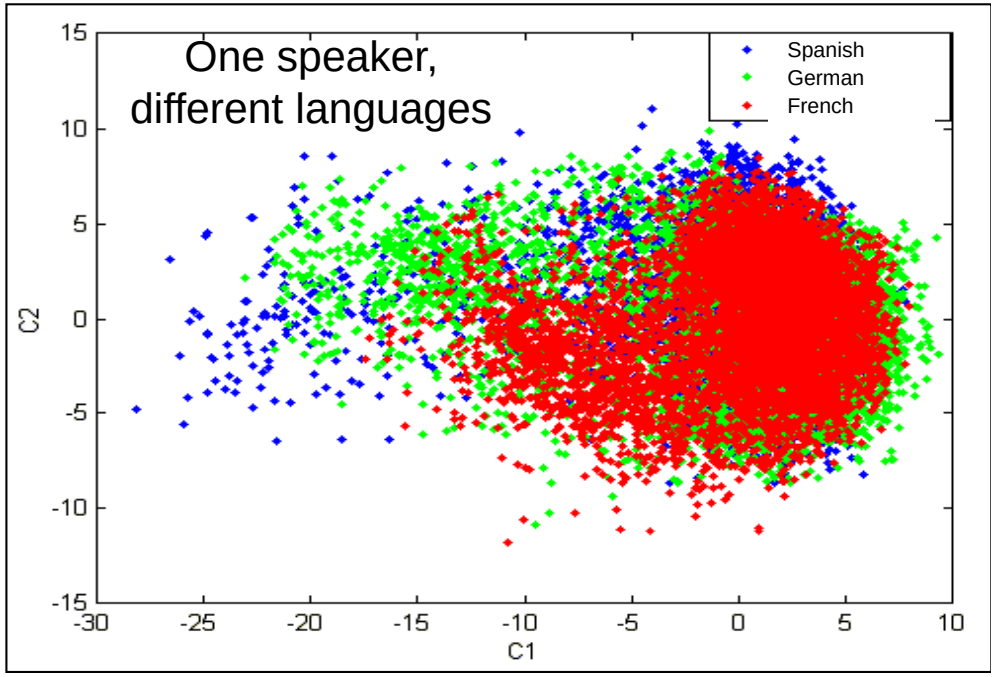


- Several strategies
  - Mel-Frequency Cepstral Coefficient (MFCC)
  - Linear Prediction Cepstral Coefficients (LPCC)
  - Perceptual Linear Prediction (PLP)

Spectral feature space

- Problems:
  - Overlapping between speakers
  - Variability due to channel, language, noise, etc.

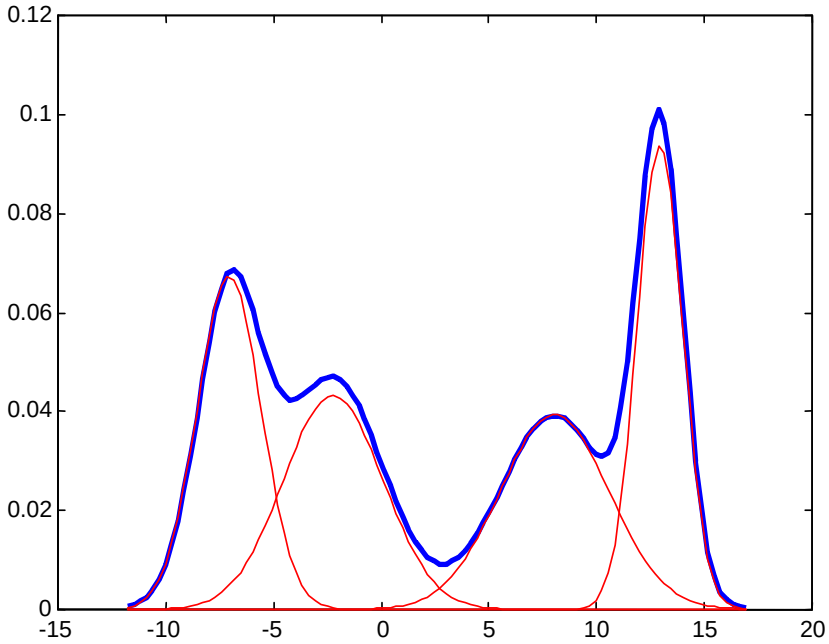




Gaussian Mixture Model Universial Background Model (GMM-UBM)

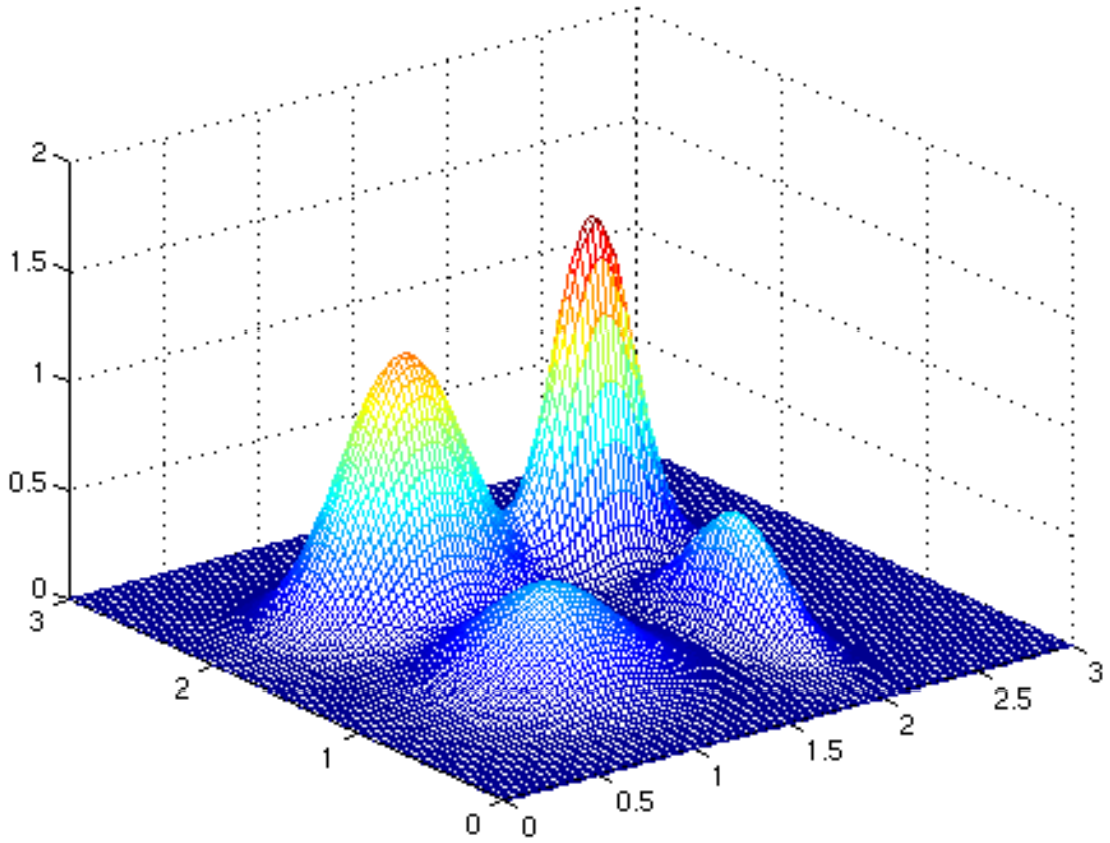
Gaussian Mixture Models (GMM)

- Multidimensional probability density function
- Models feature space using a mixture of Gaussians
- Generative model
  - [Reynolds00]



Gaussian Mixture Models (GMM)

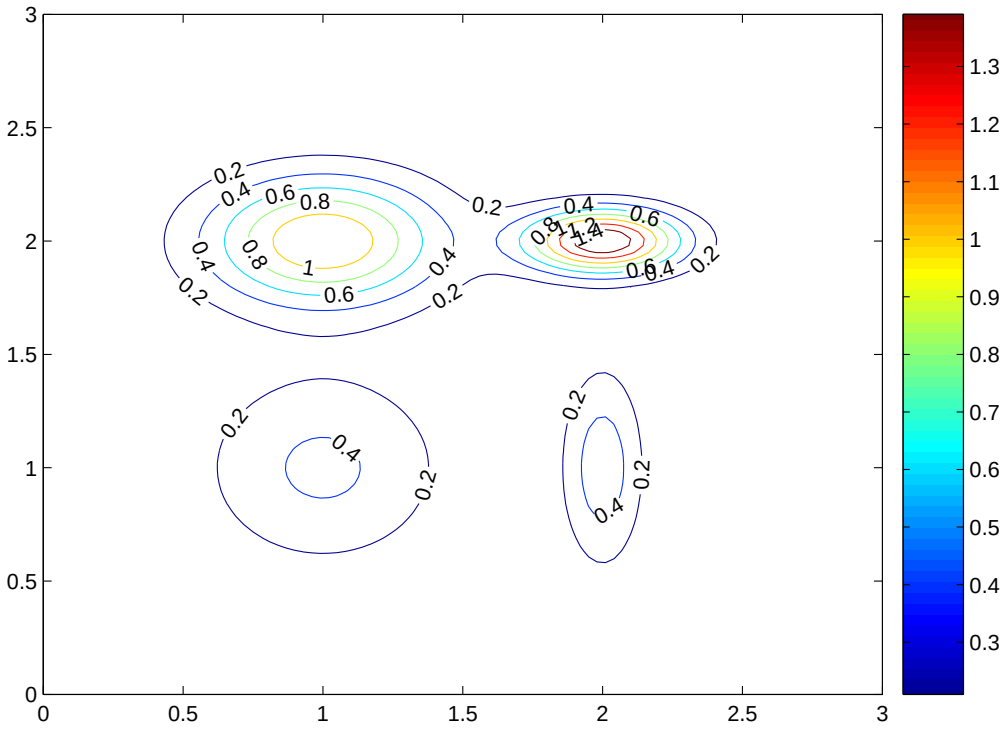
- 4-mixtures GMM in a 2-dimensional feature space



Gaussian Mixture Models (GMM)

- 4-mixtures GMM in a 2-dimensional feature space

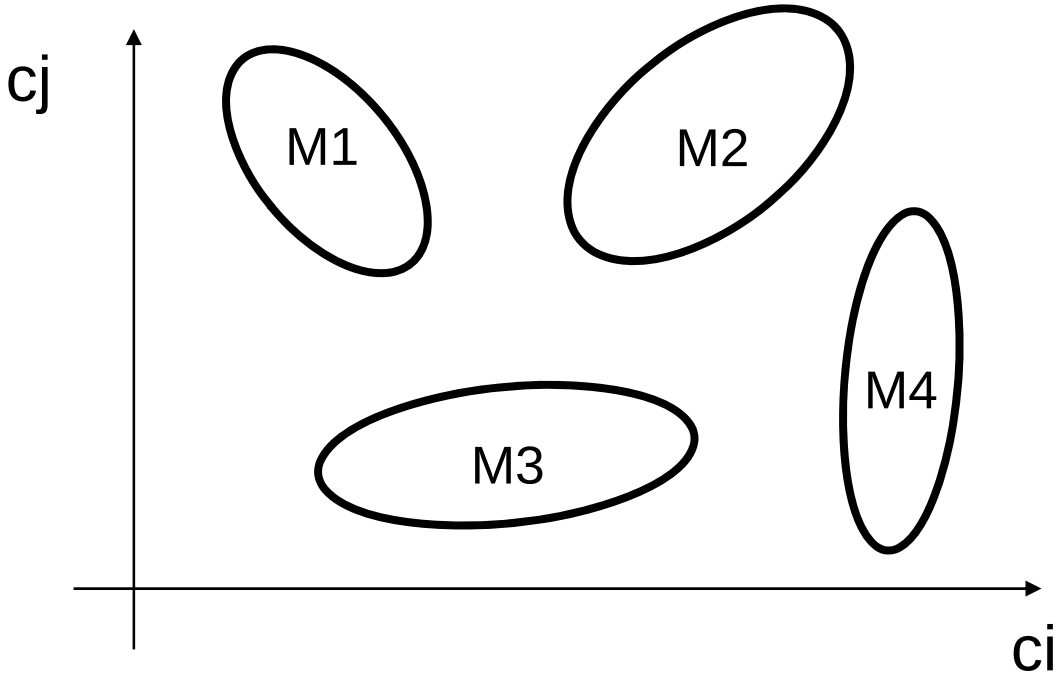
- Different regions of the space related to different vocal tract configurations
  - Therefore different values of the feature vectors



Gaussian Mixture Models (GMM)

- Formulation and elliptical representation

$$g(\mathbf{o}) = \frac{1}{(2\pi)^{D/2} \cdot |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{o}-\boldsymbol{\mu})\right] = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



$$p(\mathbf{o}|\lambda_p) = \sum_{i=1}^{M} \omega_{ip} g_{ip}(\mathbf{o})$$

GMM (parametric)

- GMM defined by a fixed number of parameters

Mean vector (mixture i): $\mu_p=\{\mu_{ip}\}$

Covariance matrix (mixture i): $\Sigma_p=\{\Sigma_{ip}\}$

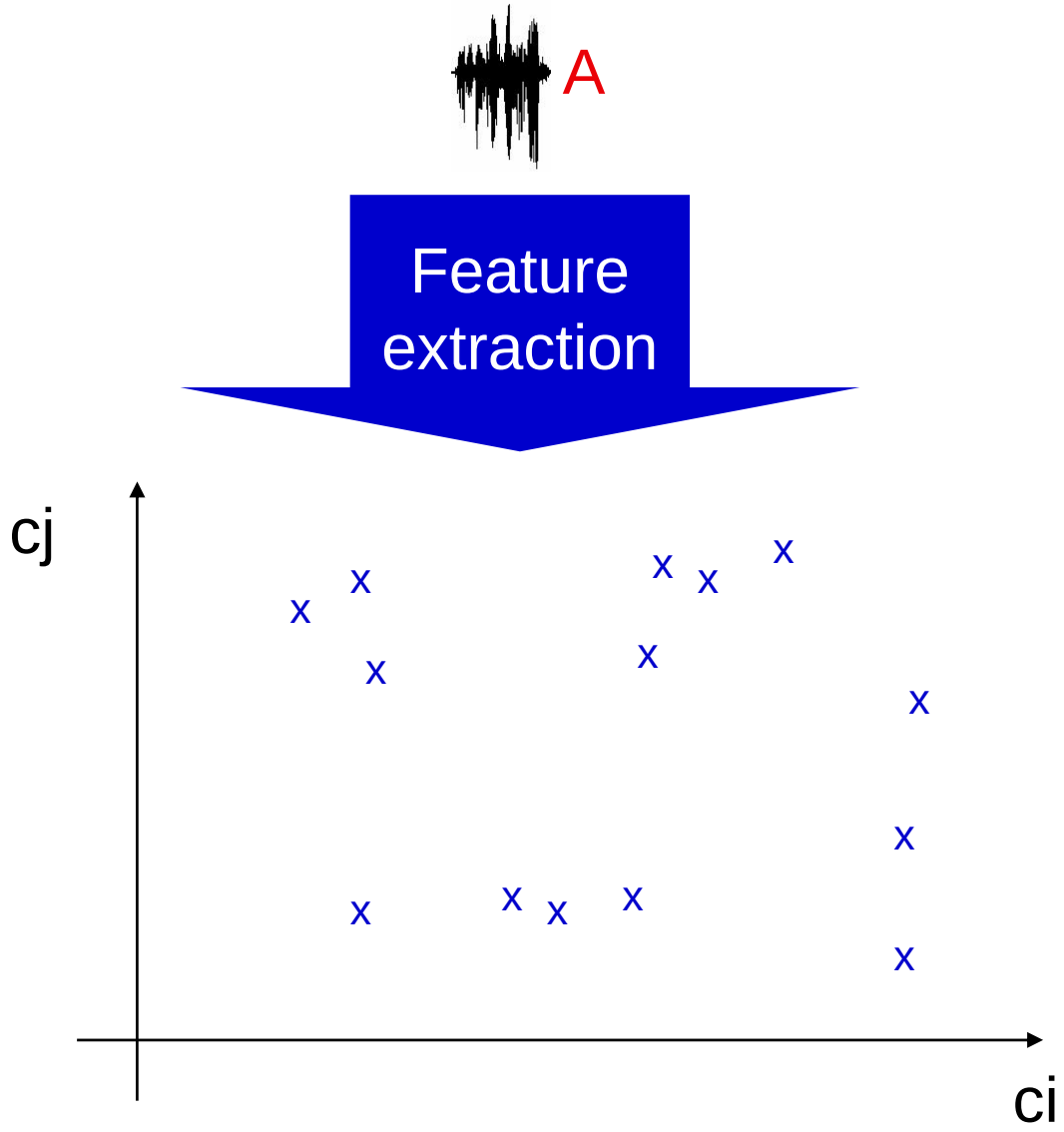
Weights vector (mixture i): $\omega_p=\{\omega_{ip}\}$, $\Sigma_i\ \omega_{ip}=1$

Model of speaker p: $\lambda_p=\{\mu_{ip},\Sigma_{ip},\omega_{ip}\}$

$$p\left(\mathbf{o}\middle|\lambda_p\right)=\sum_{i=1}^{M}\omega_{ip}g_{ip}\left(\mathbf{o}\right) \qquad g_{ip}\left(\mathbf{o}\right)=N\left(\boldsymbol{\mu}_{ip},\Sigma_{ip}\right)$$

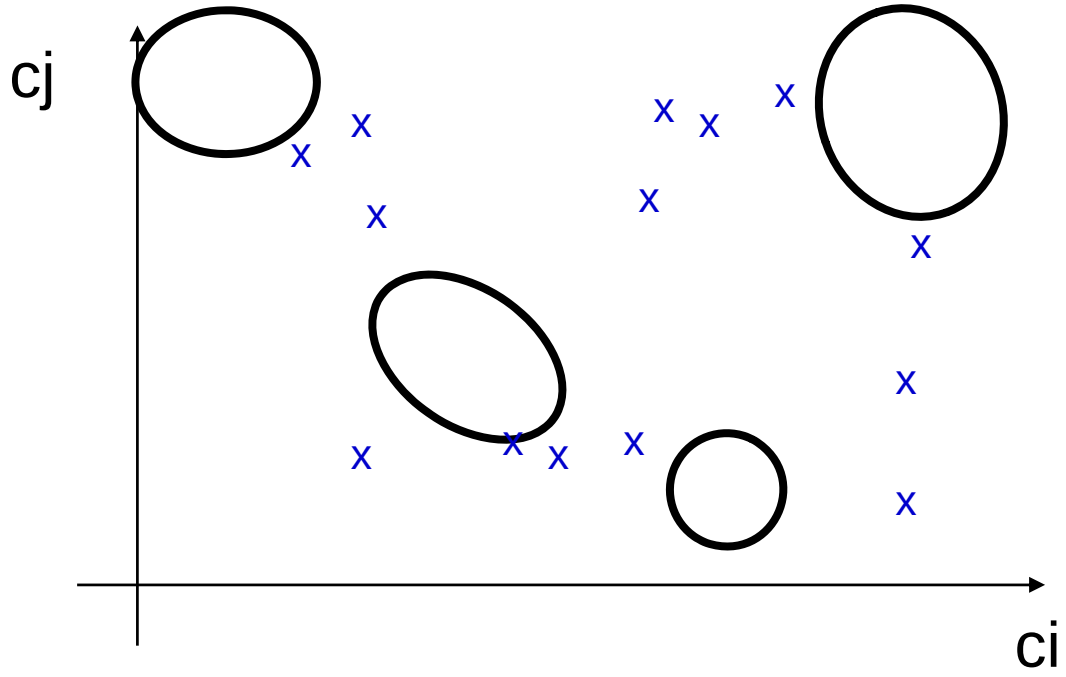
GMM training

- From training data



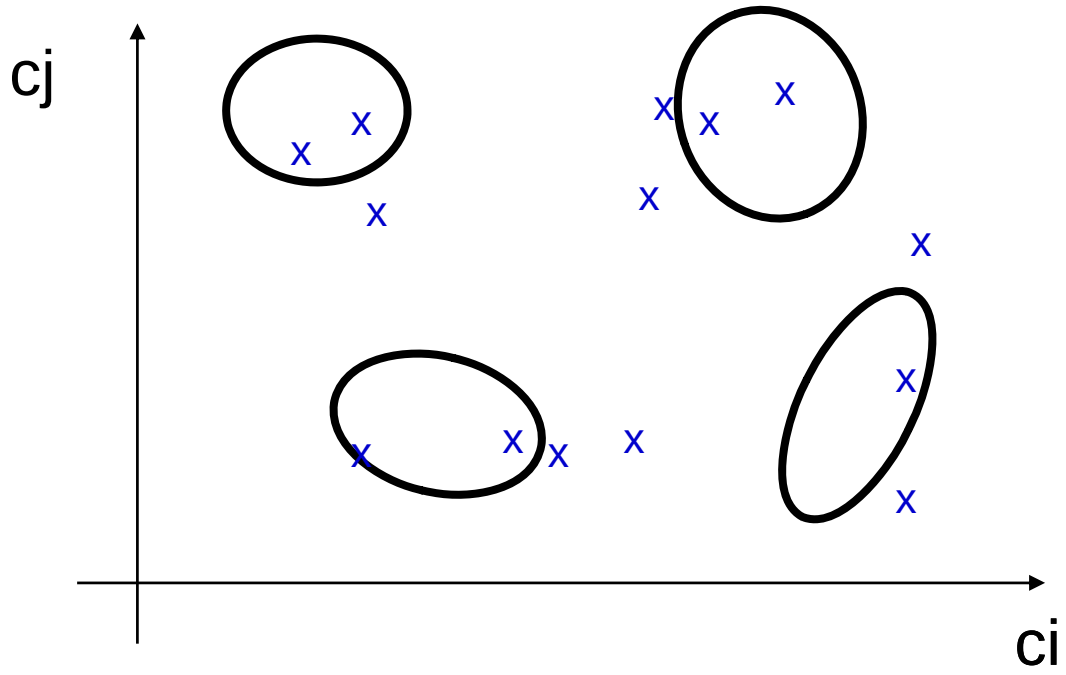
GMM training

- From training data
- Model initialization



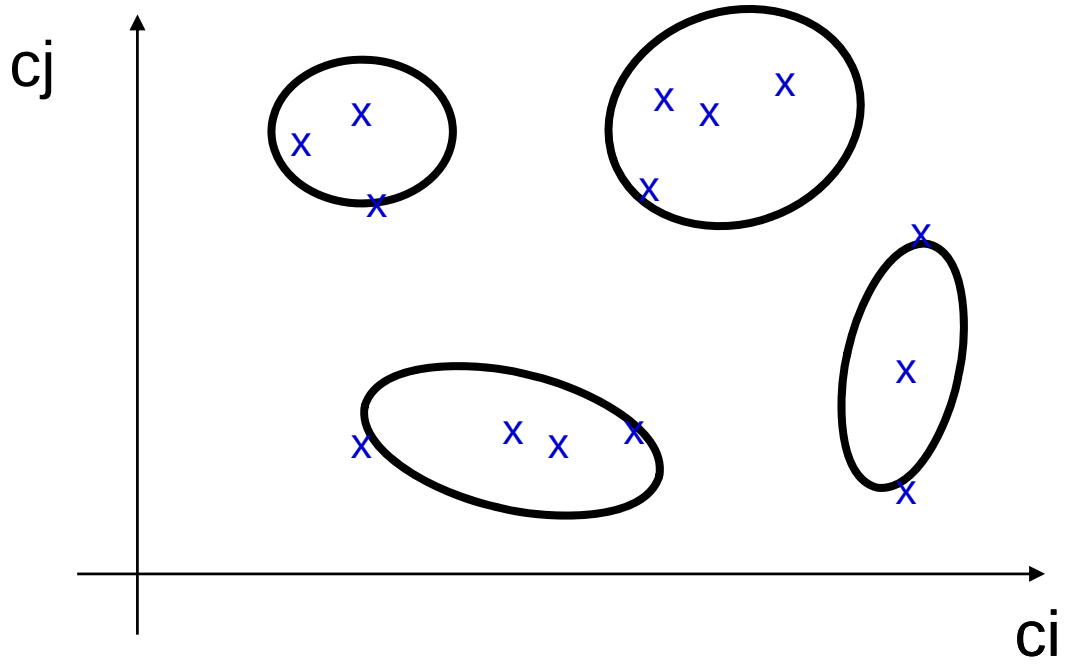
GMM training

- From training data
- Model initialization
- Maximum Likelihood (ML) fitting to the data
  - Performed iteratively
  - Expectation Maximization (EM) algorithm



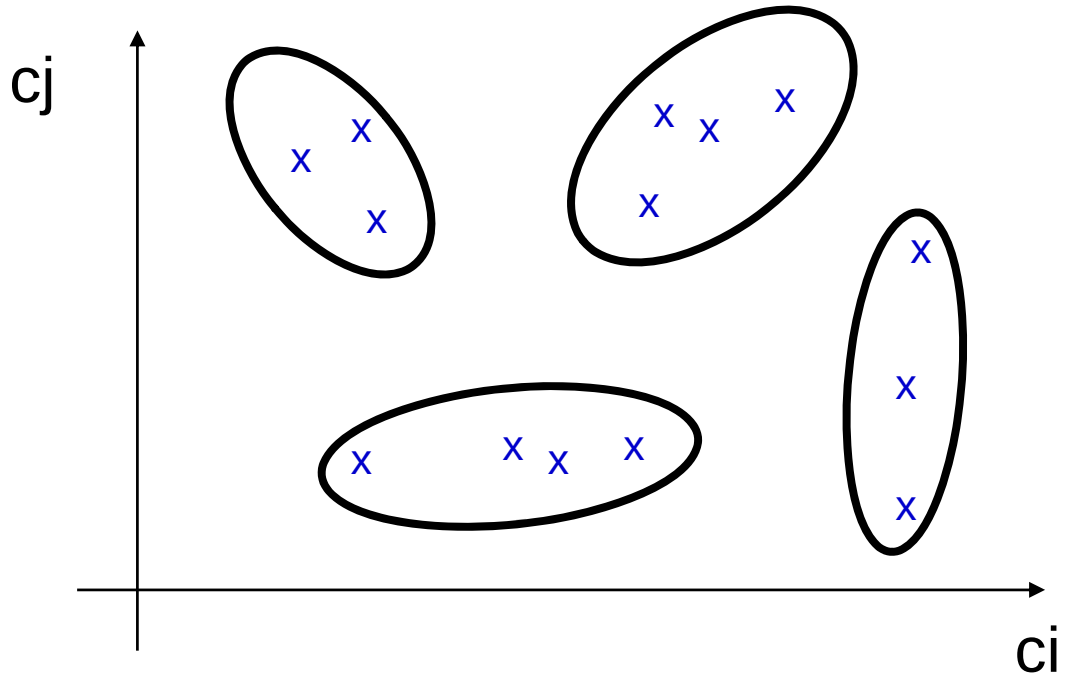
GMM training

- From training data
- Model initialization
- Maximum Likelihood (ML) fitting to the data
  - Performed iteratively
  - Expectation Maximization (EM) algorithm



GMM training

- From training data
- Model initialization
- Maximum Likelihood (ML) fitting to the data
  - Performed iteratively
  - Expectation Maximization (EM) algorithm
- Speaker GMM model A



Score computation

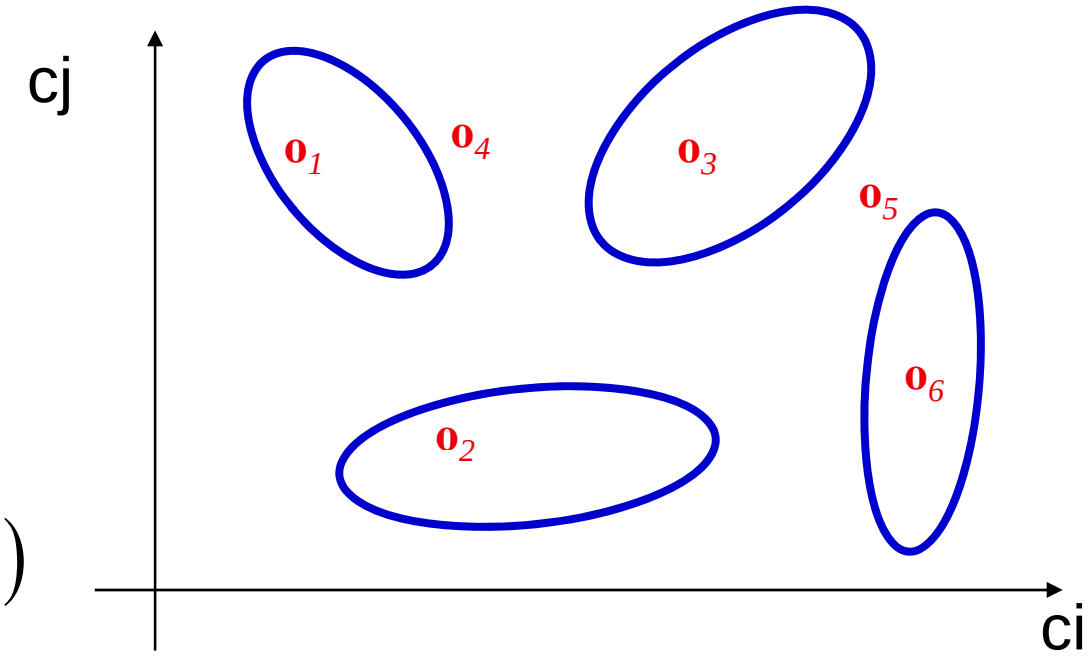
- Previously trained Speaker GMM Model A
- Features from utterance B
- Score computation:
  - Likelihood assuming independence among samples

B

Feature extraction

$$(\mathbf{o}_1,\ldots,\mathbf{o}_6)$$

$$p(\mathbf{O}|\lambda_A) = \prod_{t=1}^{T} p\left(\mathbf{o}_t|\lambda_A\right)$$

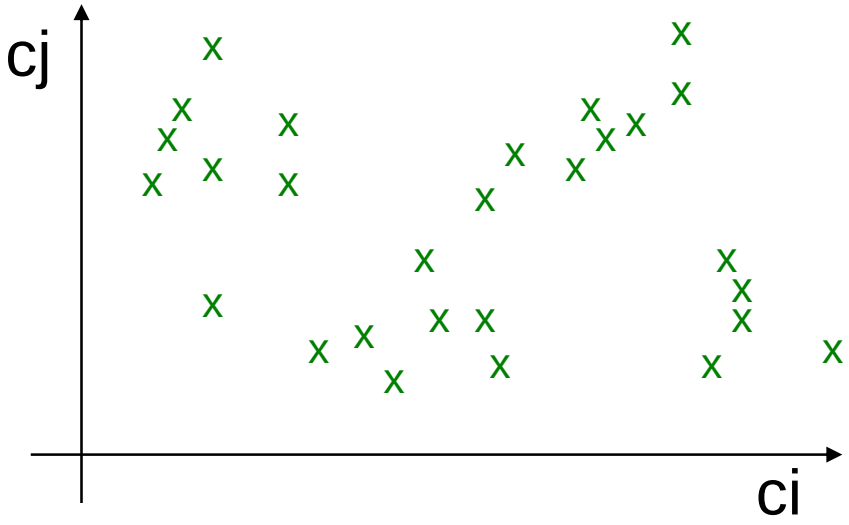


Adaptation from universal model

- Training speech is almost always limited
  - Non-general speaker models
- Universal Background Model (UBM)
  - Trained with speech from many individuals
  - Aims at modelling maximum variability for the target application
- Idea:
  - UBM represents the feature distribution common to all speakers
  - The speaker model adapts from the UBM
    - Training features: particular speaker distribution
    - Regions of the feature space not seen in training data preserves the UBM distribution
  - Robustness against lack of training data

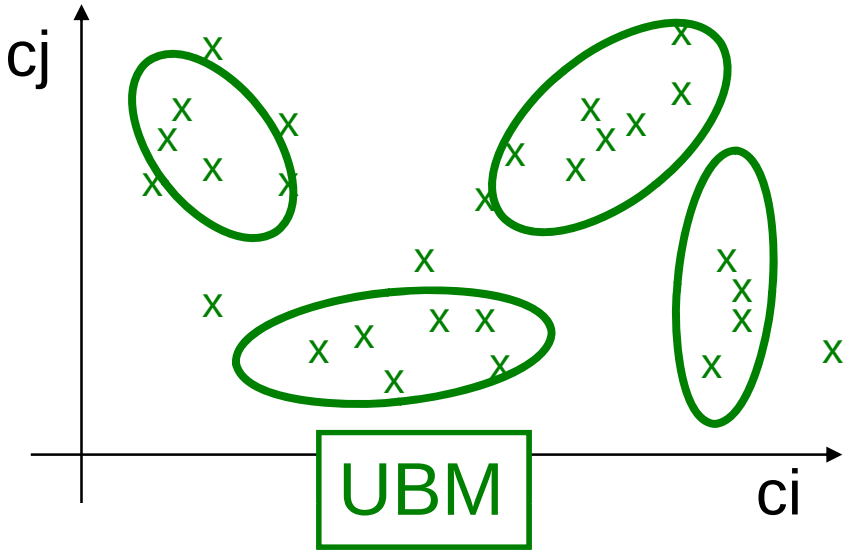
Adaptation from UBM

- Features from a universal set of speaker
  - Representing variability in the target application



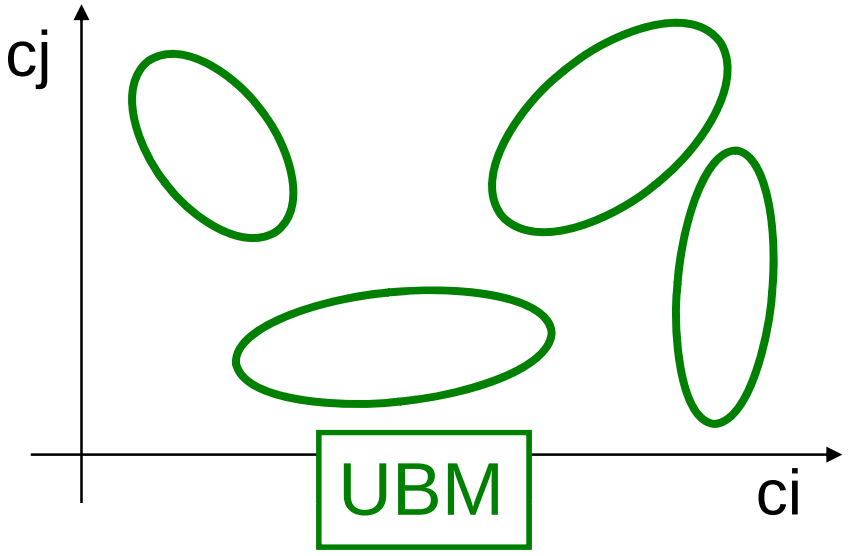
Adaptation from UBM

- UBM Training
  - Maximum Likelihood
  - Using the EM algorithm



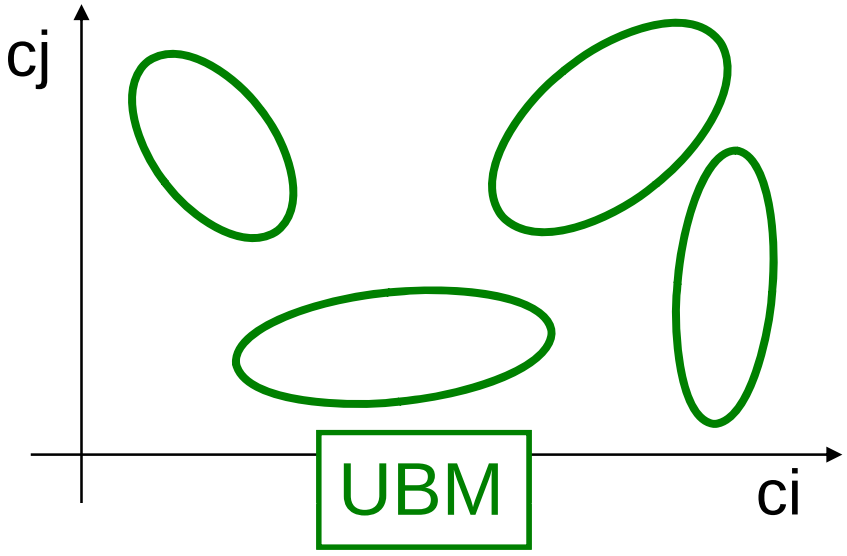
Adaptation from UBM

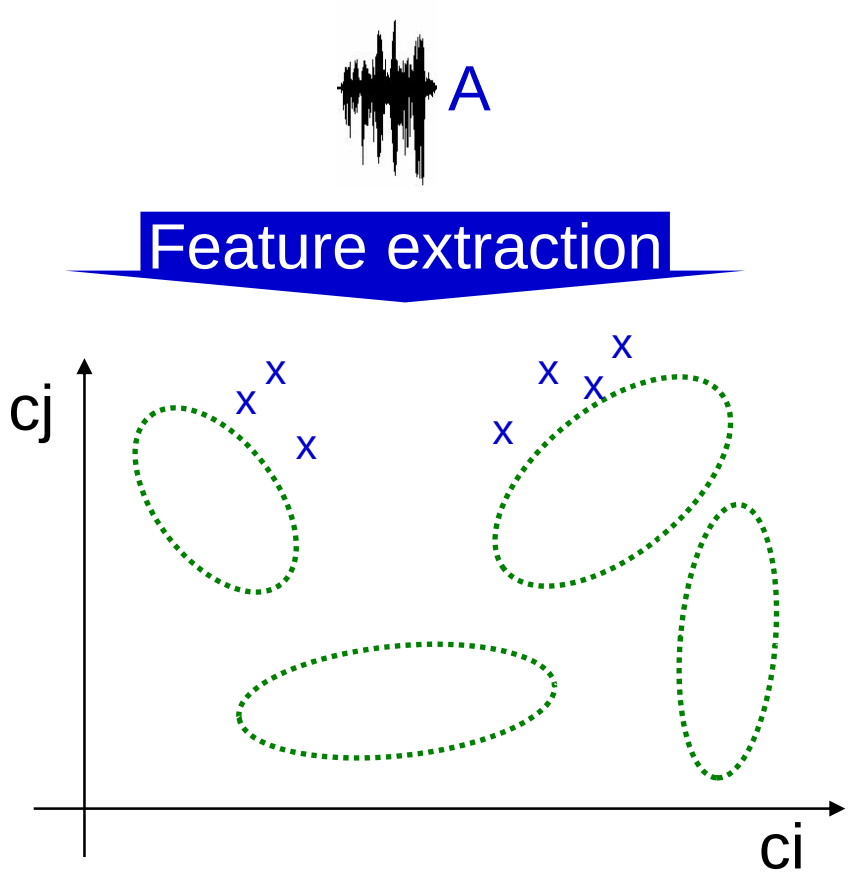
- UBM Training
  - Maximum Likelihood
  - Using the EM algorithm



Adaptation from UBM

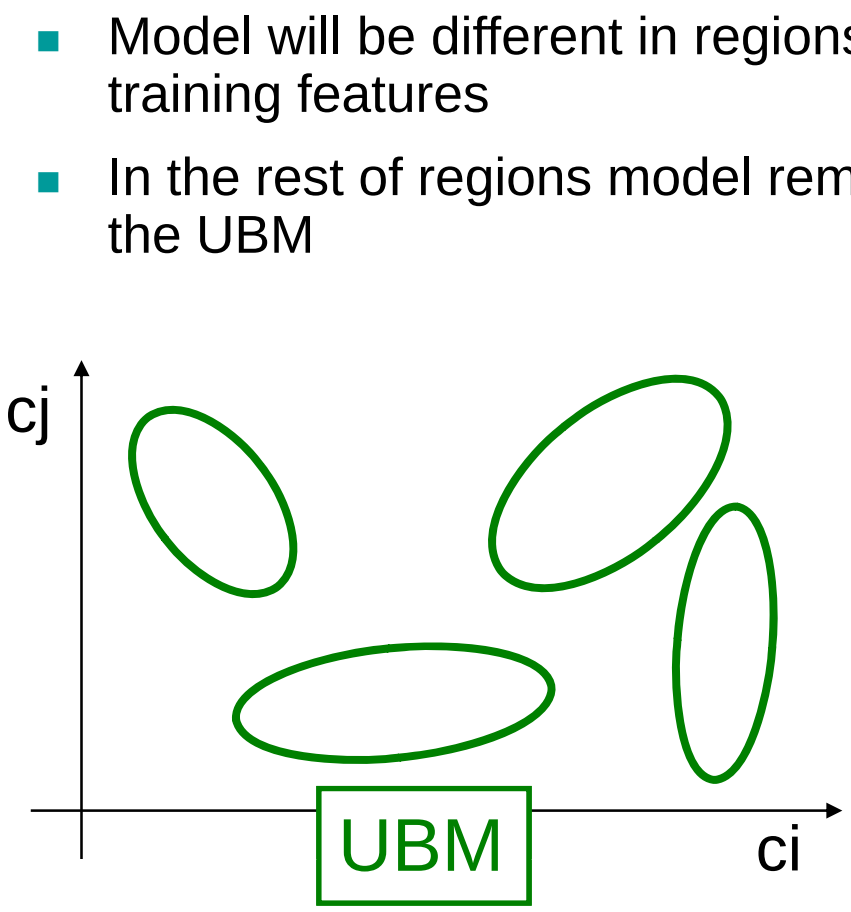
- Features extracted from the speaker training data
  - May be scarce

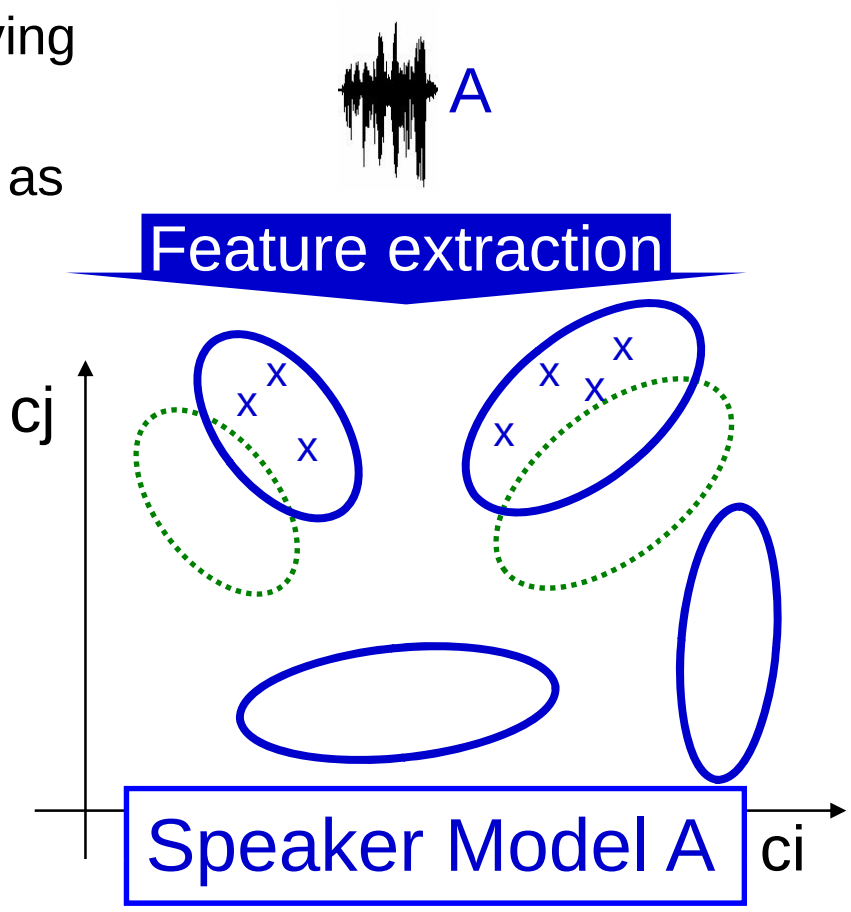




Adaptation from UBM

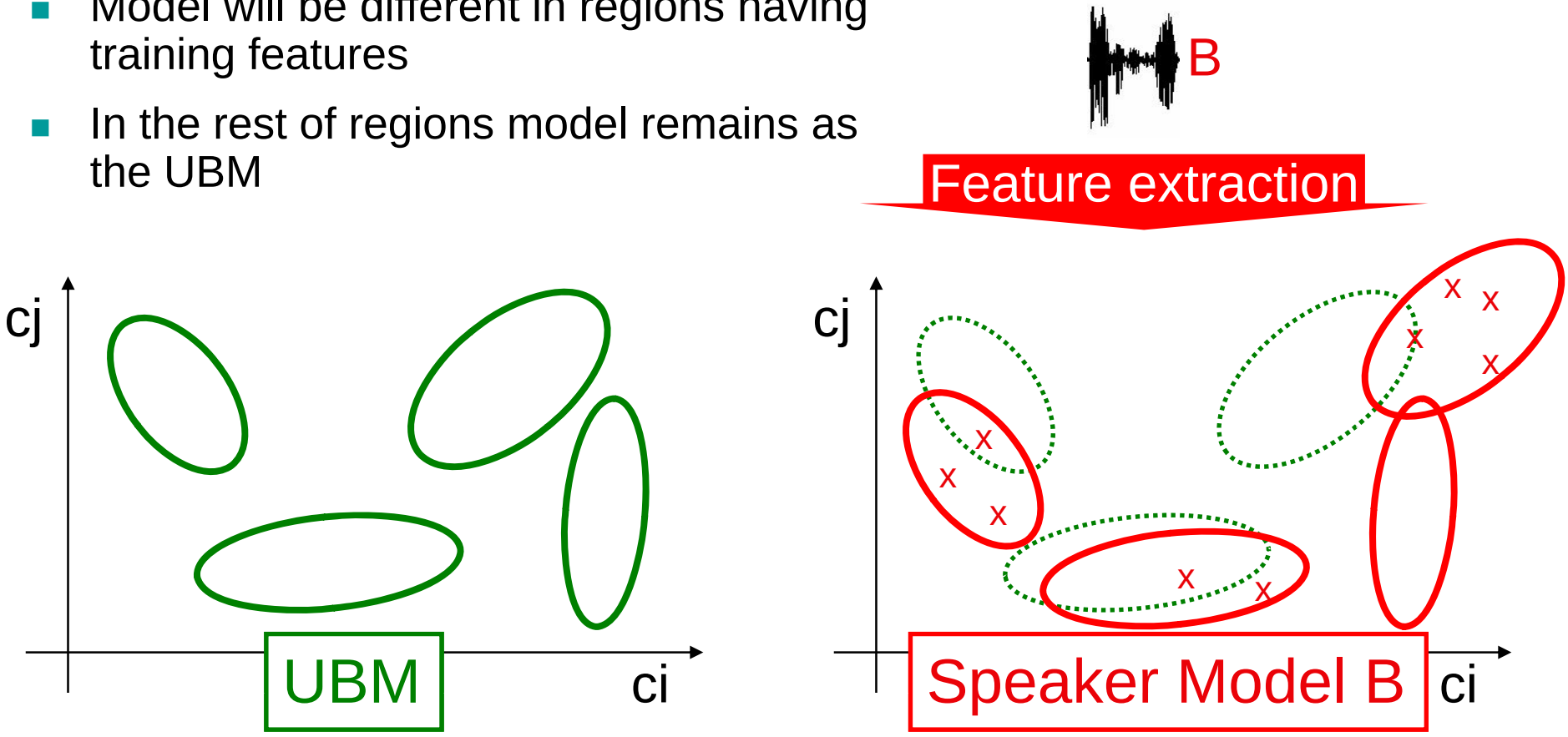
- Adaptation from the UBM
  - Maximum A Posteriori (MAP)
  - Using the EM algorithm
- Model will be different in regions having training features
- In the rest of regions model remains as the UBM





Adaptation from UBM

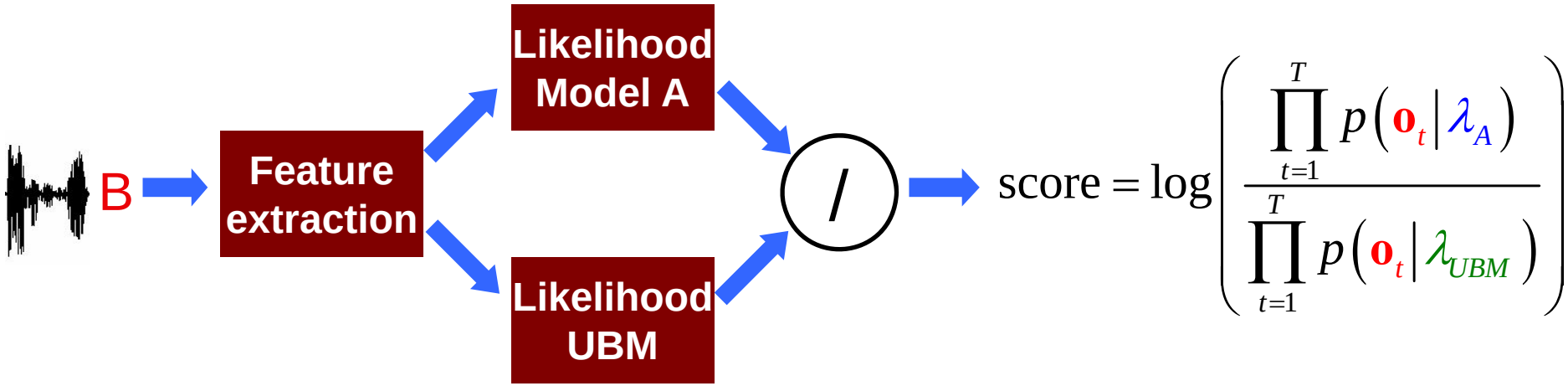
- Adaptation from the UBM
  - Maximum A Posteriori (MAP)
  - Using the EM algorithm
- Model will be different in regions having training features
- In the rest of regions model remains as the UBM



GMM-UBM Scoring

- Normalization against UBM scoring
  - Speaker specificities are highlighted w.r.t. the universe

$$p(\mathbf{O}|\lambda_A) = \prod_{t=1}^{T} p(\mathbf{o}_t|\lambda_A)$$



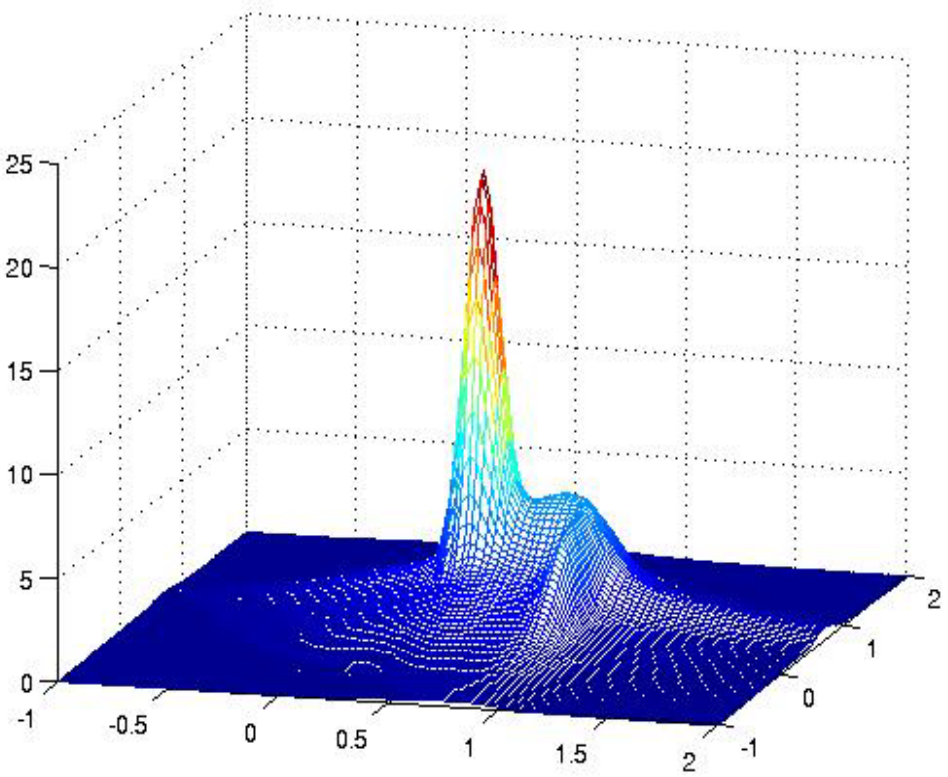
$$\log\left(\frac{\prod_{t=1}^{T} p(\mathbf{o}_t|\lambda_A)}{\prod_{t=1}^{T} p(\mathbf{o}_t|\lambda_{UBM})}\right)$$

$$p(\mathbf{O}|\lambda_{UBM}) = \prod_{t=1}^{T} p(\mathbf{o}_t|\lambda_{UBM})$$

GMM Supervectors and SVM-SM Systems

GMM Supervectors (GMM-SV)

- GMM-UBM models differ among themselves just in the values of their means
- New feature space: GMM supervectors

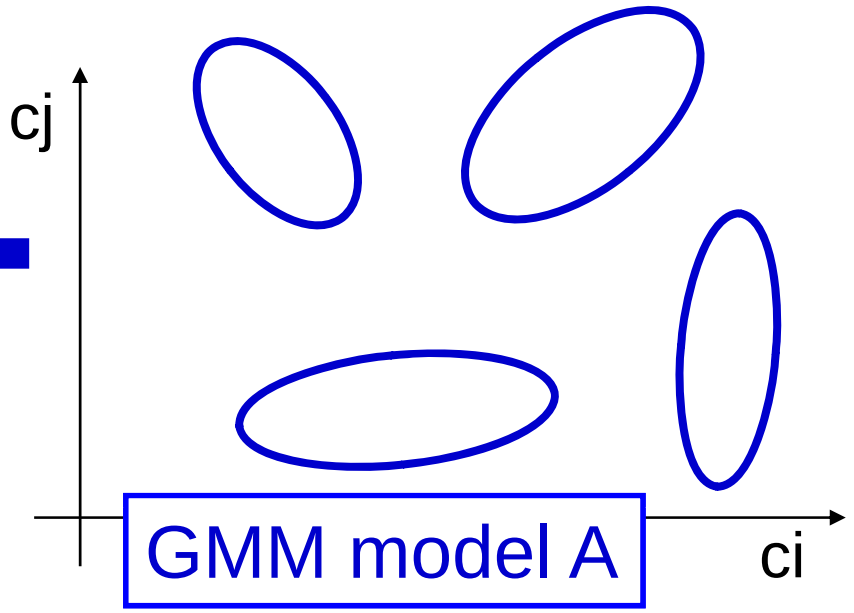


$$supervector_{\lambda} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_n \end{bmatrix}$$

Support Vector Machines (SVM) using GMM-SV

- SVM: separating hyperplane of two classes
  - Class 1: genuine user
  - Class 2: impostor
- Spectral features (*e.g.*, MFCC) difficult to separate
- Supervector feaures can be much easier separated
  - Step 1: a GMM is trained for each speech utterance
  - Step 2: build the supervector from the GMM

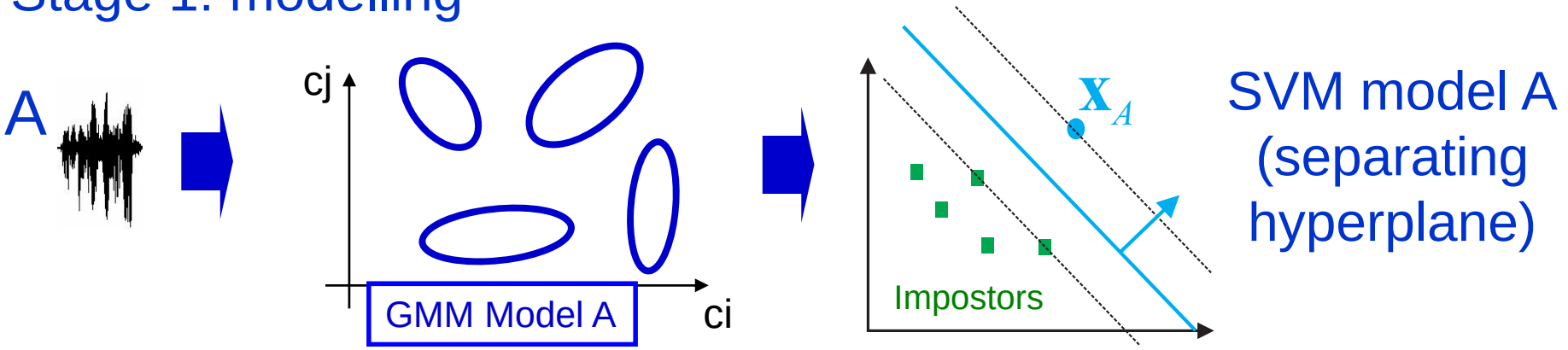
$$\mathbf{x}_A = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M)$$



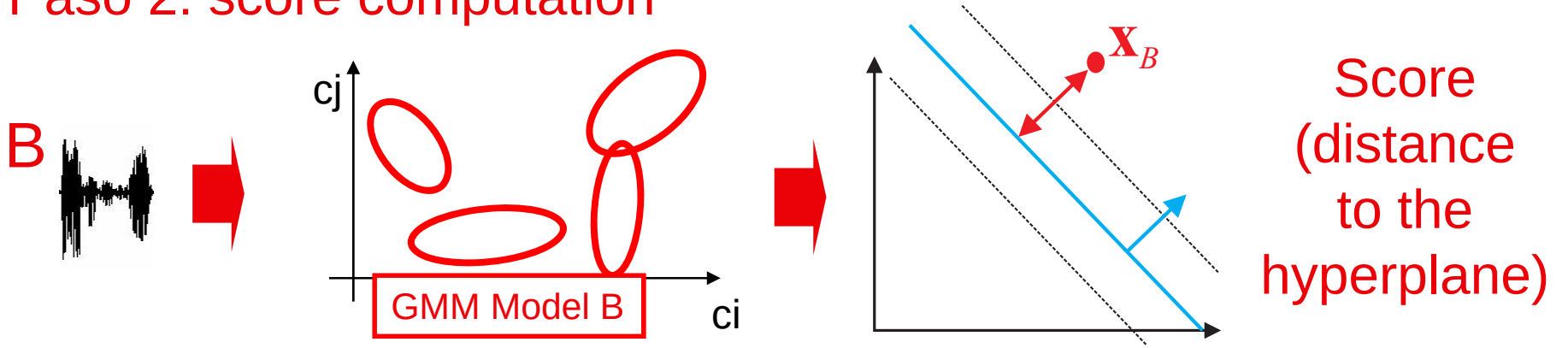
- New feature space of dimension D x M
- D: dimension of the MFCC features
- M: mixtures of the GMM

Support Vector Machines (SVM) using GMM-SV

Stage 1: modelling



Paso 2: score computation



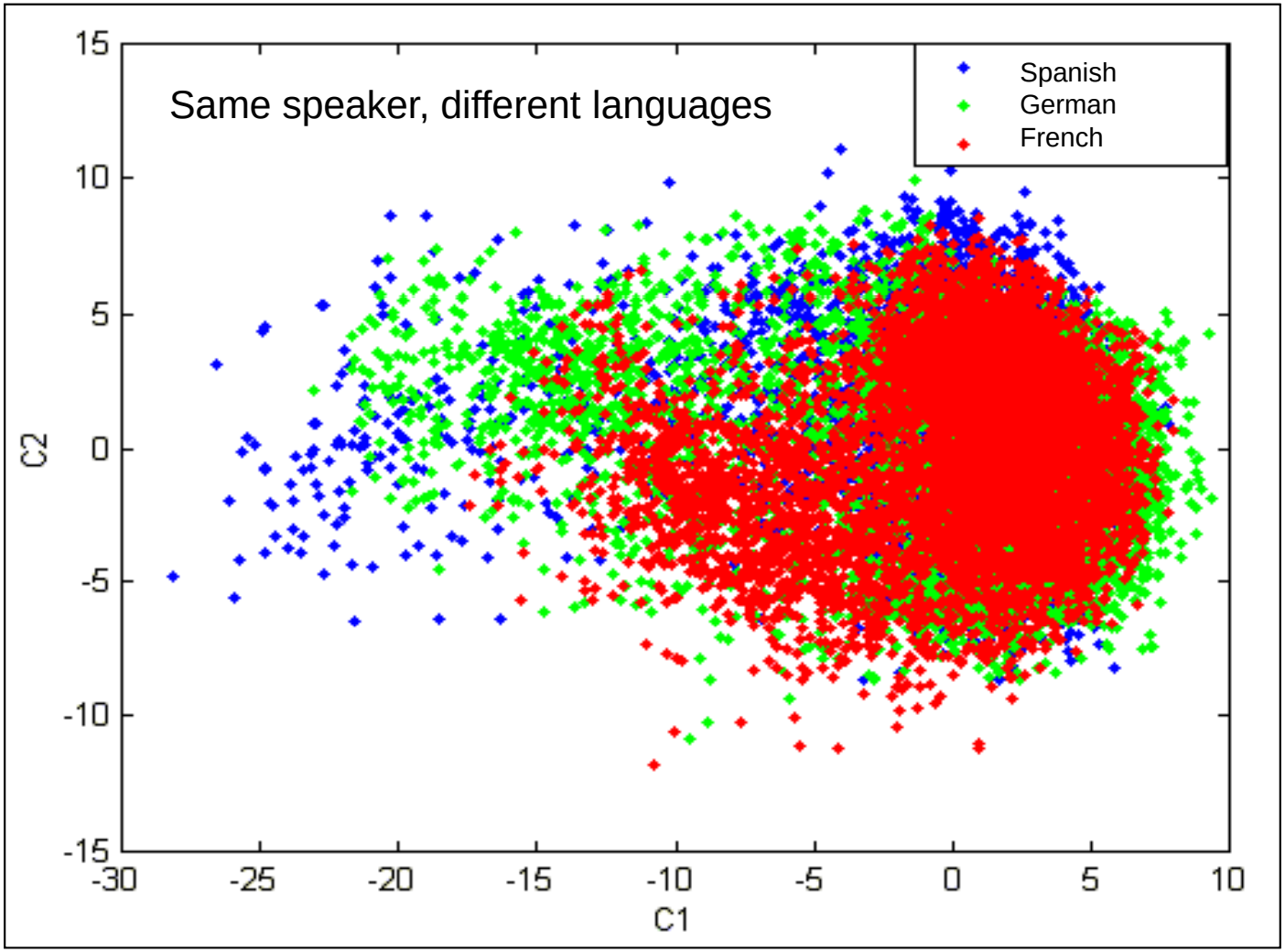
Session- and Channel-Mismatch Compensation

Inter-session variability

Variation in the feature distribution among sessions of the same speaker

- Due to many factors
  - Acoustics, channel, language, mood…
- Strongly degrades performance of systems
  - Mismatch in conditions of utterances to be compared
- Compensation of this effects
  - Fundamental research line of the community in the last decade

Example: Language Variability



Compensating Inter-session Variability

- Many techniques proposed in the literature
- Classification
  - Application domain
    - Features
    - Models (GMM models, SVM models)
  - Need of training data
  - Need of labelled data

Compensation method	Domain		Training Data?		Labeled data?	
	Models	Features	Yes	No	Yes	No
CMN		X		X		X
RASTA		X		X		X
FEATURE WARPING		X		X		X
FEATURE MAPPING		X	X		X	
FACTOR-ANALYSYS (different flavours)	X	X	X			X

Feature Warping

- Feature domain
- Does not require training data
- Does not require labeled data

Feature Warping

Feature Warping for Robust Speaker Verification

Jason Pelecanos, Sridha Sridharan

Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, AUSTRALIA

j.pelecanos@qut.edu.au s.sridharan@qut.edu.au

Abstract

We propose a novel feature mapping approach that is robust to channel mismatch, additive noise and to some extent, non-linear effects attributed to handset transducers. These adverse effects can distort the short-term distribution of the speech features. Some methods have addressed this issue by conditioning the variance of the distribution, but not to the extent of conforming the speech statistics to a target distribution. The proposed target mapping method warps the distribution of a cepstral feature stream to a standardised distribution over a specified time interval.

isation was applied over the whole utterance or over a relatively small window of one second or less. This either limited the robustness to noise variations by having a long normalisation window, or reduced the resolution and response of the channel compensation portion by use of a relatively shorter window (of approximately 250-1000ms in length). A recent approach [7] successfully examined the use of a neural network structure to perform a non-linear mapping of (mean-removed) cepstral features to establish an improved parameterisation for telephone network speaker recognition. The neural network was trained to discriminate speakers by modeling speech data from speakers recorded over different handsets. The robustness of this ap-

Feature Warping

- Hypothesis
  - Noise and channel distorsion affects mainly the distribution of cepstral features
  - But not the temporal relationship of the MFCC vectors

- Compensation solution
  - Match the MFCC distribution to a Gaussian
    - Assuming independence among MFCC components

Feature Warping

- [Pelecanos and Sridharan 2001]

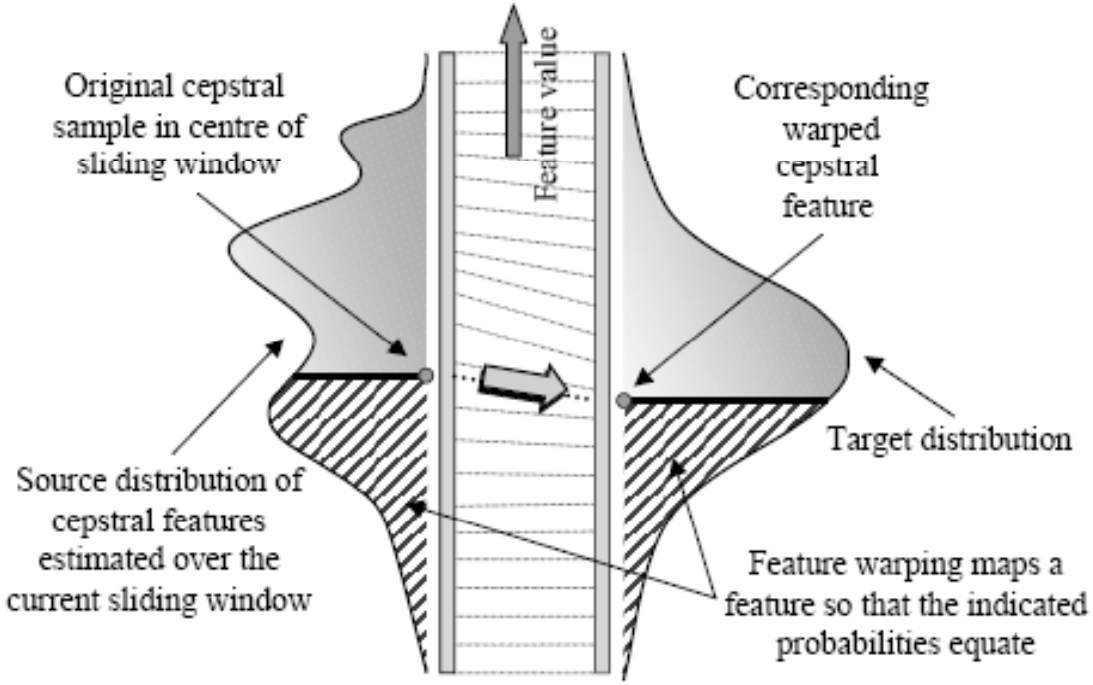


Figure 2: *Warping of features according to a target distribution shape.*

Feature Warping

- [Pelecanos and Sridharan 2001]

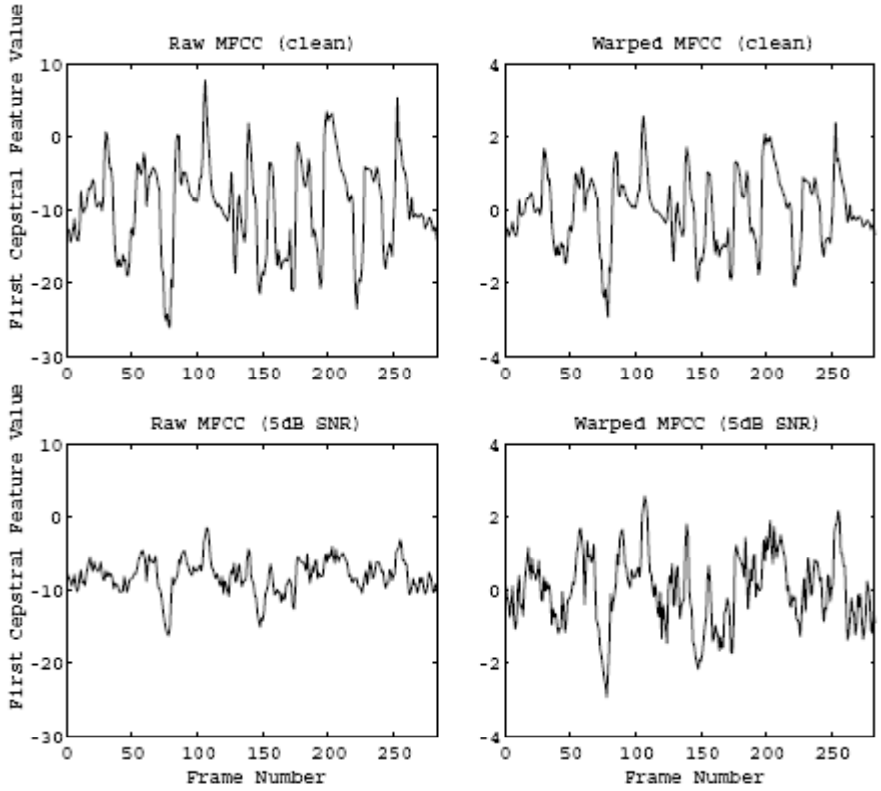


Figure 3: *Effect of additive noise on raw and warped cepstral features.*

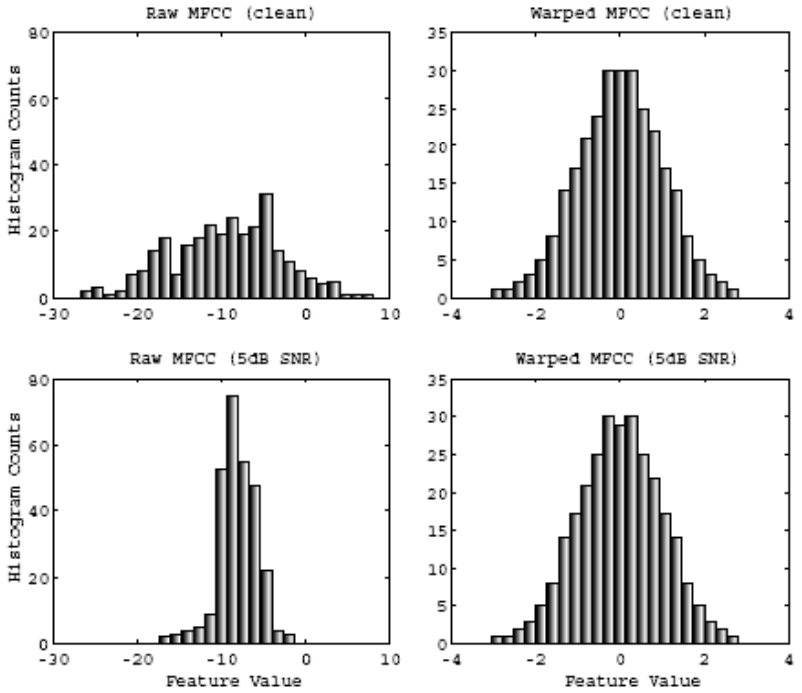


Figure 4: *Histogram of raw and warped cepstral features with and without additive noise (derived from the corresponding data used for Figure 3).*

Factor Analysis (FA)

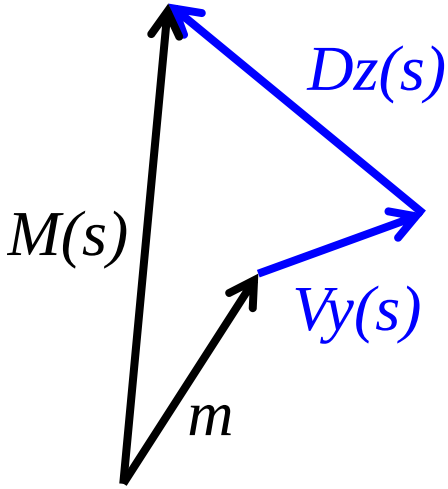
- FA linearly models the session variability of supervectors
- Training data is needed
  - Different sessions for the same speaker

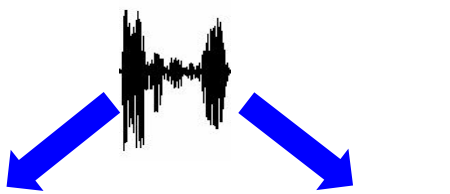
Factor Analysis: Assumptions

- Distribution of GMM supervectors depends on
  - Speaker characteristics
  - Other characteristics, varying across different sessions
- Much of the variability of supervectors for the same speaker is confined in a low-dimensional space
  - Defined by latent speaker and channel factors

Factor Analysis: Speaker Information

- Training objective
  - Acquire speaker information
  - Result: speaker-dependent model $M(s)$
    - Adapted from the UBM m
    - Confined to a speaker subspace V

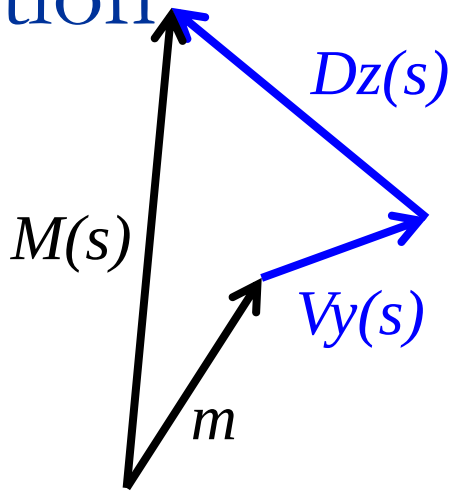




$$M(s) = m + Vy(s) + Dz(s)$$

Factor Analysis: Speaker Information

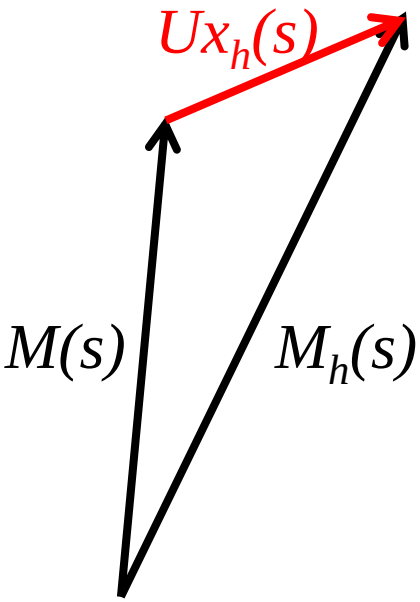
$$M(s) = m + Vy(s) + Dz(s)$$



- m = UBM (speaker-independnet)
- $Dz(s)$ = equivalent to MAP adaptation
- V = speaker subspace (columns: eigenvoices, speaker-independent)
  - Trained with information about many speakers
  - Where are the speakers confined (in which subspace)?
- $y(s)$ = speaker factors
  - Extracted form the training speaker utterance
  - Where is speaker s in the subspace defined by V?

Factor Analysis: Session Variability

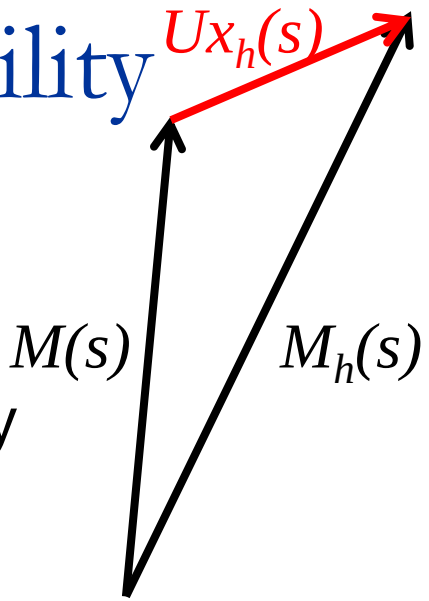
- Objective
  - Capturing session variability information
    - Explicit model of session variability
    - Can be used for
      - Eliminating it from the speaker test utterance model
      - Adapt the training model to the test data
      - …



$$M_h(s) = M(s) + Ux_h(s)$$

Factor Analysis: Session Variability

$$M_h(s) = M(s) + Ux_h(s)$$



- $M_h(s)$ = supervector of utterance h (affected by session variability)
- U = session variability subspace (columns: eigenchannels)
  - Trained with utterances of different speakers, each one in different sessions
  - Where is the session variability confined?
- $x_h(s)$ = *channel factors* (latent factors of session variability)
  - Extracted from the test segment
  - Where is the session variability of utterance h in the subspace defined by U?

Conclusion

Requirements for Forensic-Comparison Science:

- Objective
- Replicable
- Demonstrated validity and reliability
- Logically correct

Thank You

http://geoff-morrison.net
http://forensic-voice-comparison.net
http://forensic.unsw.edu.au

http://arantxa.ii.uam.es/~dramos/
http://atvs.ii.uam.es/