Pattern Recognition

Approximating class densities, Bayesian classifier, Errors in Biometric Systems

B. W. Silverman, *Density estimation for statistics and data analysis. London: Chapman and Hall, 1986.*

http://www.acsu.buffalo.edu/~tulyakov/papers/tulyakov_2009_CyberSecurity_Biometrics.pdf

Bayesian classification

• Suppose we have 2 classes and we know probability density functions of their feature vectors. How some new pattern should be classified?

• Bayes classification rule: classify x to the class w_i which has biggest posterior probability $P(w_i \mid x)$

$$P(w_1 \mid x) > P(w_2 \mid x) \ ? \quad w_1 \quad : \quad w_2$$

posterior

Using Bayes formula, we can rewrite classification rule:

$$p(x \mid w_1)P(w_1) > p(x \mid w_2)P(w_2) \ ? \quad w_1 \quad : \quad w_2$$

likelihood *prior*

Estimating probability density function.

- Parametric pdf estimation: model unknown probability density function $p(x \mid w_i)$ of class w_i by some parametric function $p_i(x;\theta)$ and determine parameters based on training samples.

  Example: Gaussian function

$$p(x;\mu) = \frac{1}{(2\pi)^{l/2}} e^{-\frac{1}{2}(x-\mu)^2}$$

- Non-parametric pdf estimation:
  1. Histogram
  2. K nearest neighbor
  3. Kernel methods (Parzen kernels or windows)

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{h}\varphi\left(\frac{x_i - x}{h}\right)\right)$$

  N is the number of training samples

  4. Other methods (estimating cumulative distribution function first, SVM density estimation, etc.)

Estimating kernel width

- Non-parametric pdf estimation:
  - Fixed kernels:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{h}\varphi\left(\frac{x_i - x}{h}\right)\right)$$

  - Adaptive kernels:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{h_i}\varphi\left(\frac{x - x_i}{h_i}\right)\right)$$

or

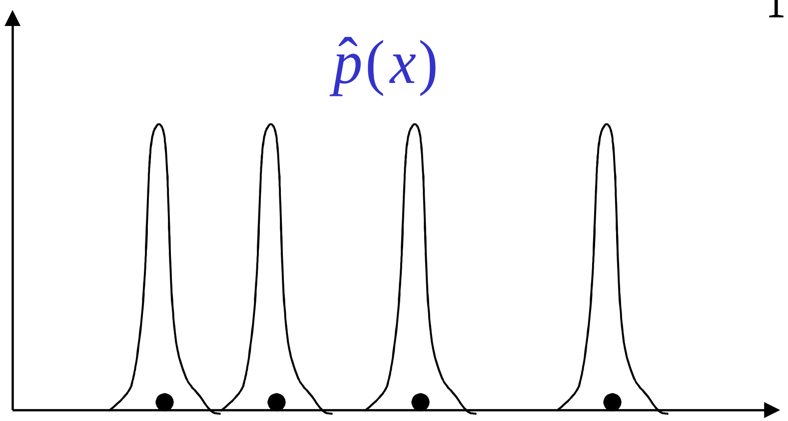
$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{h_i}\varphi_i\left(\frac{x - x_i}{h_i}\right)\right)$$

Estimating kernel width

Recall, we used maximum likelihood method for parametric pdf estimation:

$$\max_{\theta} \hat{p}(X;\theta) = \max_{\theta} \hat{p}(x_1, x_2, \ldots, x_N \mid \theta) = \max_{\theta} \prod_{k=1}^{N} \hat{p}(x_k;\theta)$$

Can we use same method for estimating the kernel width h ?



No, the max is not achievable:

$$\max_{h} \prod_{k=1}^{N} \hat{p}(x_k;h) =$$

$$\max_{h} \prod_{k=1}^{N} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} \varphi\left(\frac{x_i - x_k}{h} \right) \right) \to \infty$$

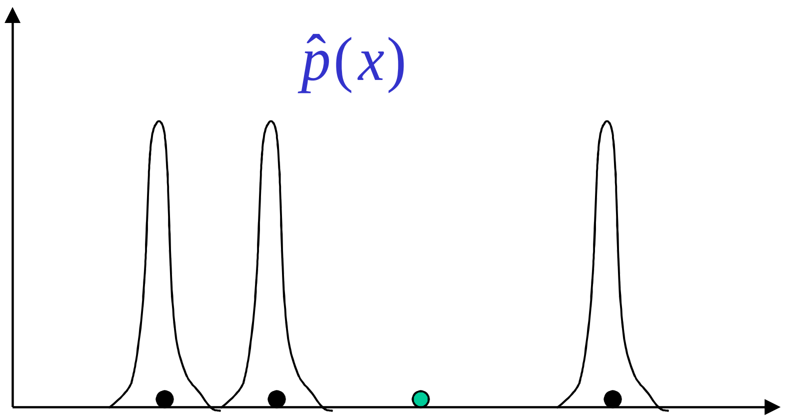
if $\quad h \to 0$

Estimating kernel width

Solution: separate model data (kernel centers) from testing data

- cross-validation technique

$$\max_h \prod_{k=1}^{N} \left(\frac{1}{N} \sum_{i \neq k} \frac{1}{h} \varphi\left(\frac{x_i - x_k}{h} \right) \right)$$

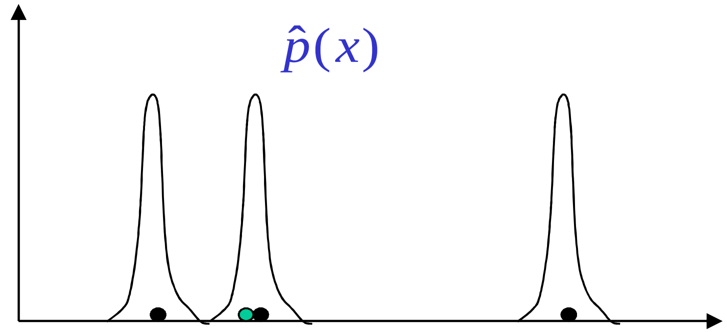


Estimating kernel width

Tried maximum likelihood cross-validation and still diverges?

$$\max_h \prod_{k=1}^{N} \left(\frac{1}{N} \sum_{i \neq k} \frac{1}{h} \varphi\left(\frac{x_i - x_k}{h} \right) \right) \to \infty$$

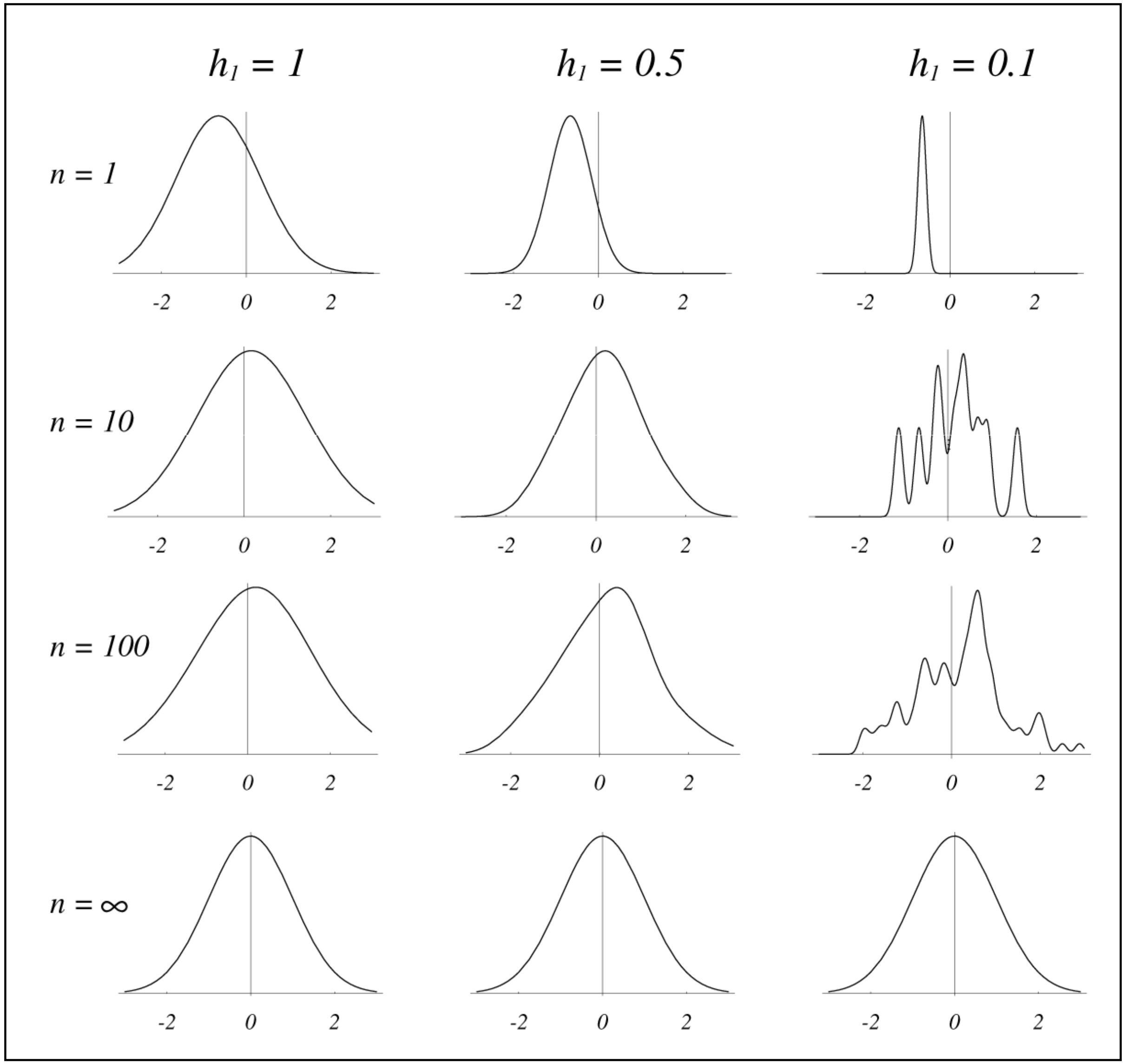
This might happen if data is somewhat discrete:



Solution - truly separate model data from testing data:

$$\max_h \prod_{k=1}^{N} \left(\frac{1}{N} \sum_{x_i \neq x_k} \frac{1}{h} \varphi\left(\frac{x_i - x_k}{h} \right) \right)$$

Examples of pdf estimation

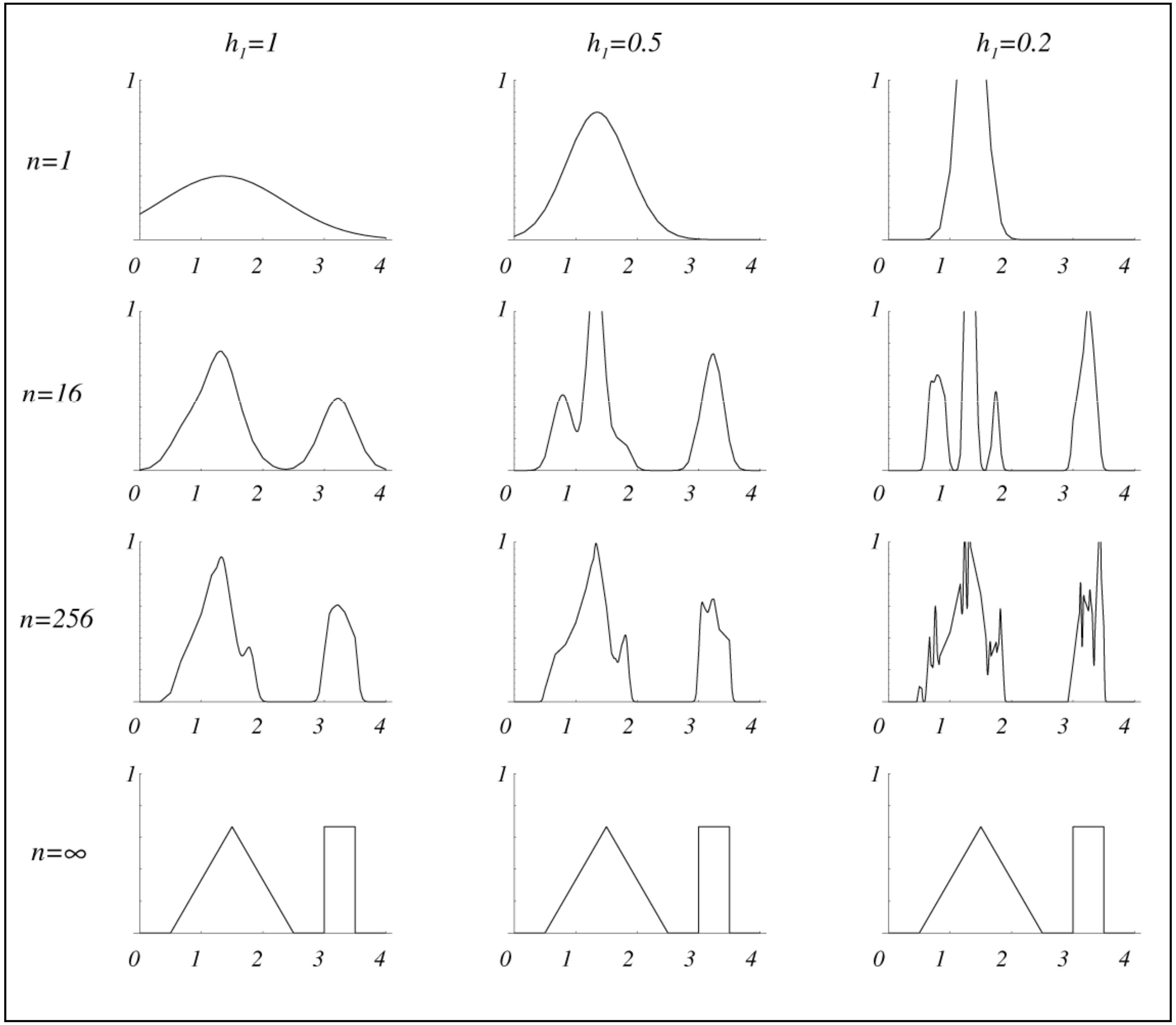


Parzen-window (kernel) estimates of a univariate normal density using different window widths and numbers of samples. (DHS)

Heuristic method of width calculation:

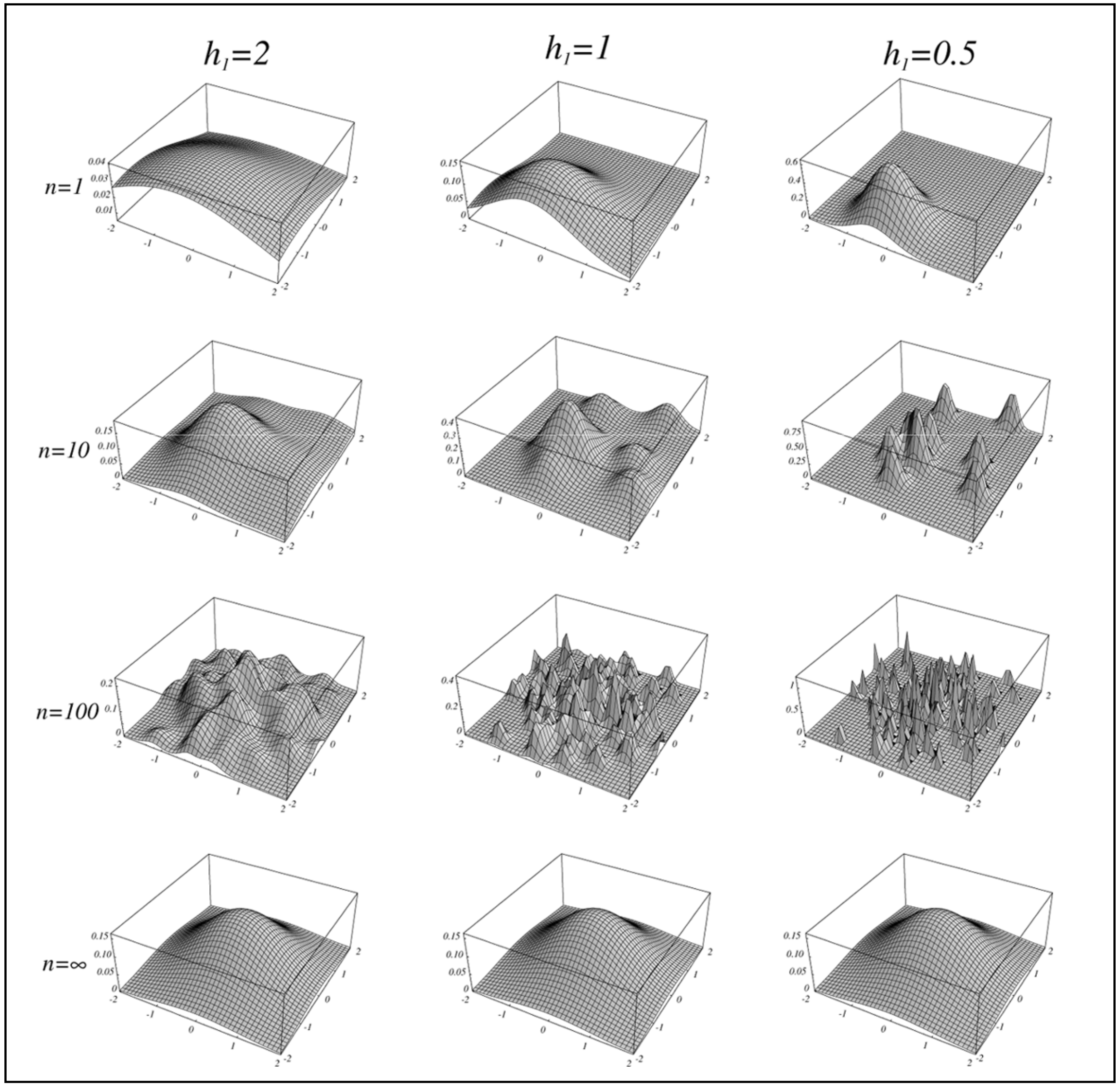
$$h_n = \frac{h_1}{\sqrt{n}}$$

Examples of pdf estimation



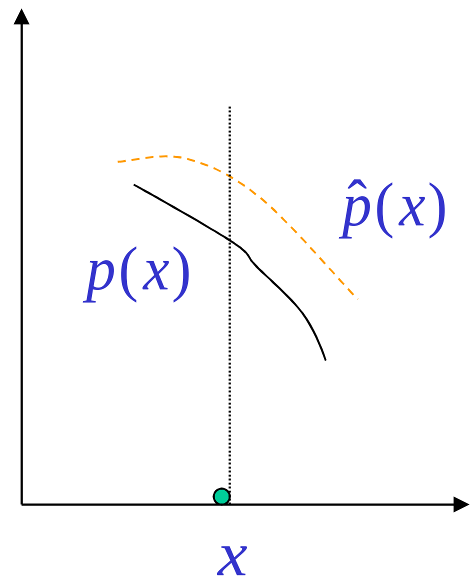
Parzen-window (kernel) estimates of a bimodal density using different window widths and numbers of samples.

Examples of pdf estimation



Parzen-window (kernel) estimates of a bivariate normal density using different window widths and numbers of samples.

Error in pdf estimation



Discrepancy between true density $p(x)$ and its estimation $\hat{p}(x)$:

$$MSE_x(\hat{p}) = E\{\hat{p}(x) - p(x)\}^2$$

- Mean Square Error

$$MISE(\hat{p}) = \int E\{\hat{p}(x) - p(x)\}^2 dx$$

- Mean Integrated Square Error

$$MSE_x(\hat{p}) = E\{\hat{p} - p\}^2 = E\{\hat{p}^2 - 2\hat{p}p + p^2\}$$

$$= E\{\hat{p}^2\} - 2E\{\hat{p}\}p + p^2$$

$$= \{E\hat{p}\}^2 - 2\{E\hat{p}\}p + p^2 + \left[E\{\hat{p}^2\} - \{E\hat{p}\}^2\right]$$

$$= [E\hat{p} - p]^2 + \left[E\{E\hat{p} - \hat{p}\}^2\right]$$

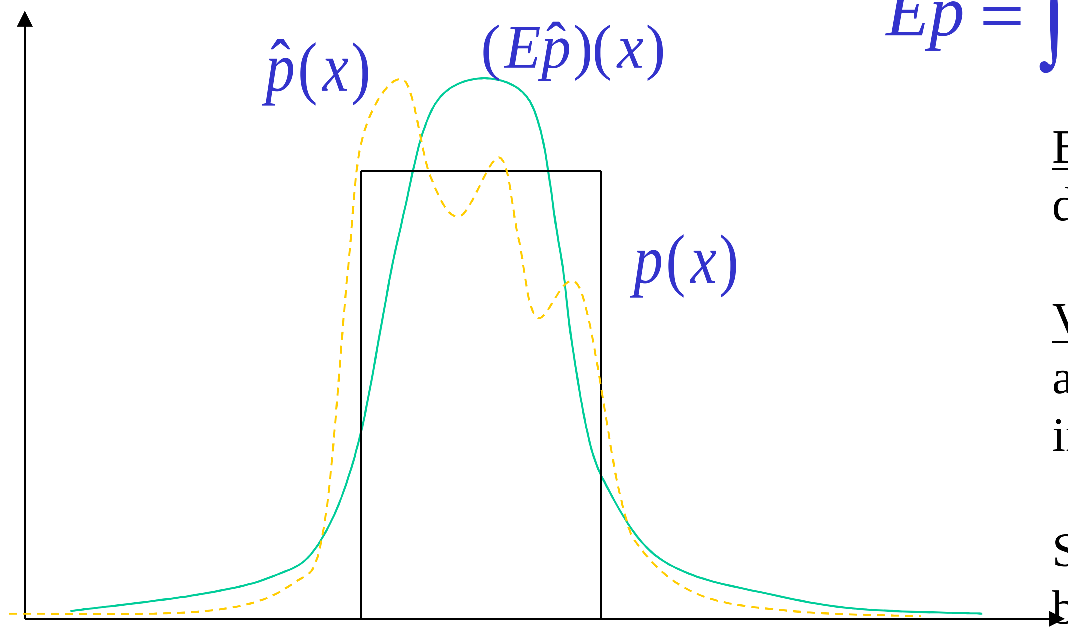
(Expectations are taken over the set of possible approximations or over the sets of training samples)

Bias and variance of estimation error

$$MSE_x(\hat{p}) = \underbrace{\left[E\hat{p} - p\right]^2}_{\text{Bias}} + \underbrace{\left[E\{E\hat{p} - \hat{p}\}^2\right]}_{\text{Variance}}$$

$E\hat{p}$ - Average approximation

$$E\hat{p} = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) p(y)\,dy$$



Bias is the difference between true density and average approximation

Variance is the difference between average approximation and individual approximations

Smaller kernel width reduces bias, but increases variance.

Bias and variance of estimation error

If some assumptions on the true density are made (e.g. $\int (p''(x))^2 dx < \infty$) then it is possible to analytically find the kernel width which gives smallest $MISE(\hat{p})$

Silverman (Parzen):

$$h_{opt} = k_2^{-2/5} \left\{ \int \varphi(t)^2 dt \right\}^{1/5} \left\{ \int p''(x)^2 dx \right\}^{-1/5} n^{-1/5}$$

Optimal kernel width gets smaller when the number of training samples n increases. For optimal kernel width $MISE(\hat{p})$ also decreases:

$$MISE \sim C(\varphi) \left\{ \int p''(x)^2 dx \right\}^{1/5} n^{-4/5}$$

Note, that $p(x)$ is unknown. Above formulas are useful for theory, but not for practical applications.

For multivariate pdf approximation: $MISE \sim n^{-4/(4+d)}$

The performance decreases exponentially when the number of dimensions increases

Bayesian classification

• Bayes classification rule: classify x to the class w_i which has biggest posterior probability $P(w_i \mid x)$

$$P(w_1 \mid x) > P(w_2 \mid x) \ ? \quad w_1 \quad : \quad w_2$$

• Bayes classification rule minimizes the total probability of misclassification.

Cost of errors.

• Errors happen when samples of class 1 are incorrectly classified to belong to class 2, and samples of class 2 are classified to belong to class 1.
• The cost of making these errors can be different :

λ_1 - the cost of misclassifying samples of class 1

λ_2 - the cost of misclassifying samples of class 2

Total cost (or risk) of classification

Classification algorithm splits feature space into two decision regions:

R_1 - samples in this region are classified as being in class 1

R_2 - samples in this region are classified as being in class 2

$\int_{R_2} p(x \mid w_1) dx$ - the proportion of samples of class 1 being classified as class 2

$\int_{R_1} p(x \mid w_2) dx$ - the proportion of samples of class 2 being classified as class 1

$P(w_1) \int_{R_2} p(x \mid w_1) dx$ - the proportion of all input samples being class 1 but classified as being in class 2

$P(w_2) \int_{R_1} p(x \mid w_2) dx$ - the proportion of all input samples being class 2 but classified as being in class 1

$$Cost = \lambda_1 P(w_1) \int_{R_2} p(x \mid w_1) dx + \lambda_2 P(w_2) \int_{R_1} p(x \mid w_2) dx$$

- total cost

Minimizing total cost of classification

Since R_1 and R_2 cover whole feature space

$$\int_{R_1} p(x \mid w_1)dx + \int_{R_2} p(x \mid w_1)dx = 1$$

Thus

$$Cost = \lambda_1 P(w_1)\{1 - \int_{R_1} p(x \mid w_1)dx\} + \lambda_2 P(w_2)\int_{R_1} p(x \mid w_2)dx$$

$$= \lambda_1 P(w_1) + \int_{R_1} (\lambda_2 P(w_2)\, p(x \mid w_2) - \lambda_1 P(w_1)\, p(x \mid w_1))dx$$

Cost is minimized if R_1 includes only points where

$$\lambda_2 P(w_2)\, p(x \mid w_2) - \lambda_1 P(w_1)\, p(x \mid w_1) < 0$$

Bayesian classification

Bayesian classifier is an optimal classifier minimizing total classification cost. Such classifier is possible only if we have full knowledge about class distributions.

If $\lambda_1 P(w_1)\, p(x\,|\,w_1) > \lambda_2 P(w_2)\, p(x\,|\,w_2)$ then classify x as class 1.

If $\lambda_1 P(w_1)\, p(x\,|\,w_1) \le \lambda_2 P(w_2)\, p(x\,|\,w_2)$ then classify x as class 2.

Alternatively, assuming non-zero terms, the class assignment is based on

testing whether $\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)} > \frac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$ or $\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)} \le \frac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$

Decision surface $\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)} = \frac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$ separates two decision regions.

$\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)}$ - likelihood ratio

$\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)} > (<) \frac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$ - likelihood ratio test

Performance of Bayesian classification

Denote:

$$t = \frac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$$ - decision threshold

$$R_1(t) = \left\{ x \mid \frac{p(x \mid w_1)}{p(x \mid w_2)} > t \right\}$$ - decision region of class 1 for threshold t

$$R_2(t) = \left\{ x \mid \frac{p(x \mid w_1)}{p(x \mid w_2)} \le t \right\}$$ - decision region of class 2 for threshold t

$$MR_1(t) = \int_{R_2(t)} p(x \mid w_1) dx$$ - misclassification rate for class 1 and threshold t

$$MR_2(t) = \int_{R_1(t)} p(x \mid w_2) dx$$ - misclassification rate for class 2 and threshold t

Performance of Bayesian classification

$MR_1(t)$ and $MR_2(t)$ completely characterize the performance of a Bayesian classifier

For a given misclassification costs λ_1, λ_2 and prior class probabilities $P(w_1), P(w_2)$ we find $t = \dfrac{\lambda_2 P(w_2)}{\lambda_1 P(w_1)}$

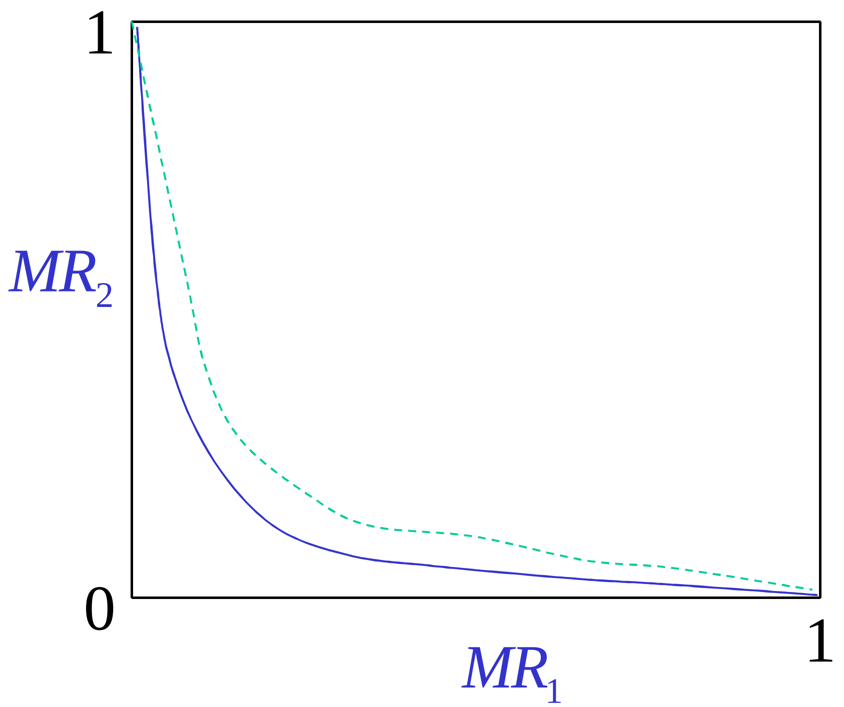
Then the (mis)classification cost is

$$Cost = \lambda_1 P(w_1) MR_1(t) + \lambda_2 P(w_2) MR_2(t)$$

ROC of a Bayesian classification

$MR_1(t)$ and $MR_2(t)$ are used only with the the same t .

Thus the parameter t is not important and the performance of a Bayesian classifier can be characterized only by the relationships between $MR_1(t)$ and $MR_2(t)$.



Example of an optimal Bayesian ROC curve (——) and some non-optimal classifier's ROC curve (-------).

For a given MR_1 the MR_2 of a non optimal classifier should be bigger; otherwise non-optimal classifier would outperform optimal.

Biometric Application Types

- **Verification System (1:1)**
  - Claim is made (enrollee identity)
  - User's biometric is matched only with stored biometric of claimed enrollee
  - The decision to accept claim is made using only one matching score

- **Identification System (1:N)**
  - No claim about identity is made
  - User's biometric is matched with stored biometrics of all enrolled persons
  - The highest matching score determines the most probable enrollee
  - The decision about accepting identification attempt is made based on the matching score for that enrollee (and optionally using other matching scores too)

- **Screening**
  - Matching against a watch list
  - Opposite of verification

Performance of Verification System

For biometric matchers (person identity verification) we distinguish two classes:

- Genuine – person's claimed identity is correct
- Impostor - person's claimed identity is in correct

The decision for genuine class is to accept, and the decision for the impostor class is to reject. The decision is usually done based on a single matching score of input biometric with the enrolled biometric template of claimed identity person.

Instead of optimal $\frac{p(x|w_1)}{p(x|w_2)} > (<) \quad \theta$ use $x > (<) \ \theta$

If $\frac{p(x|w_1)}{p(x|w_2)}$ is monotonous, these decisions are equivalent.

Instead of $MR_1(t)$ and $MR_2(t)$ use

$$FAR(t) = \int_{x>t} p(x|imp)dx$$ - false accept rate for threshold t

$$FRR(t) = \int_{x<t} p(x|gen)dx$$ - false reject rate for threshold t

Errors in Verification Systems

Each verification attempt has two possibilities:

1. Genuine event - input biometrics and stored biometrics from claimed identity belong to the same person.
2. Impostor event - input biometrics is different from claimed identity biometrics.

The scores produced by matching algorithm will have distributions:

$$p_{gen}(s) = p(s \mid \text{genuine event})$$

$$p_{imp}(s) = p(s \mid \text{impostor event})$$

Errors in Verification Systems

FAR and FRR are determined by the decision rule – accept or reject results of recognition.

Usually FAR and FRR are defined using some threshold:

$$FAR(\theta) = \int_{\theta}^{\infty} p_{imp}(s)ds = P(s > \theta \mid \text{impostor event})$$

Also called: False Match Rate (FMR)

$$FRR(\theta) = \int_{-\infty}^{\theta} p_{gen}(s)ds = P(s < \theta \mid \text{genuine event})$$

Also called: False Non-Match Rate (FNMR)

Errors in Verification Systems

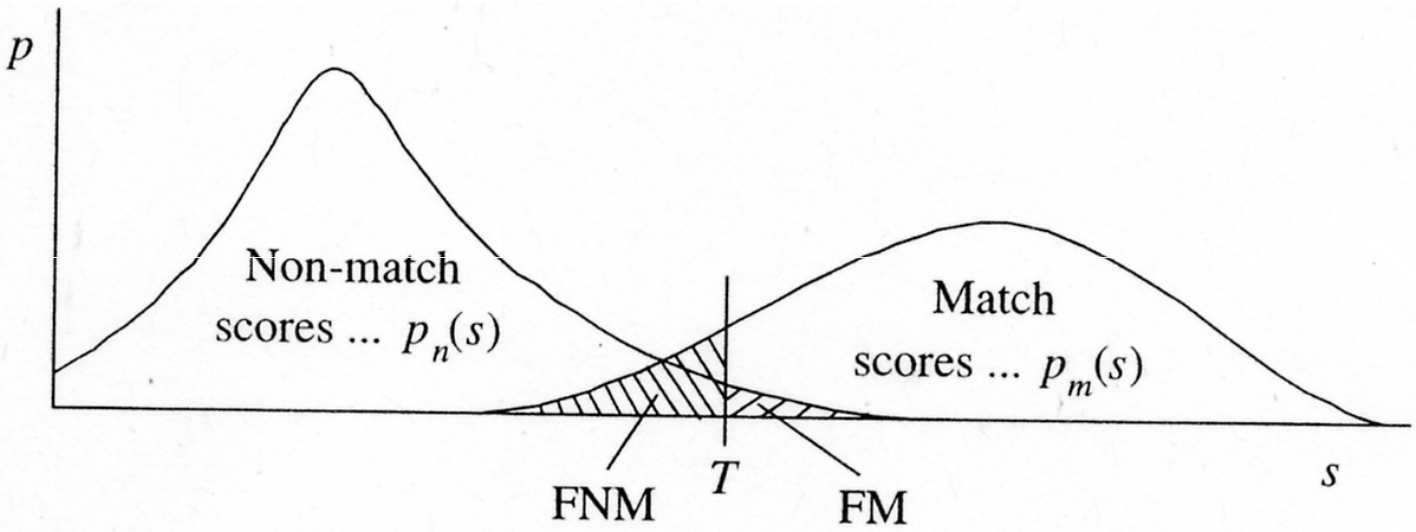
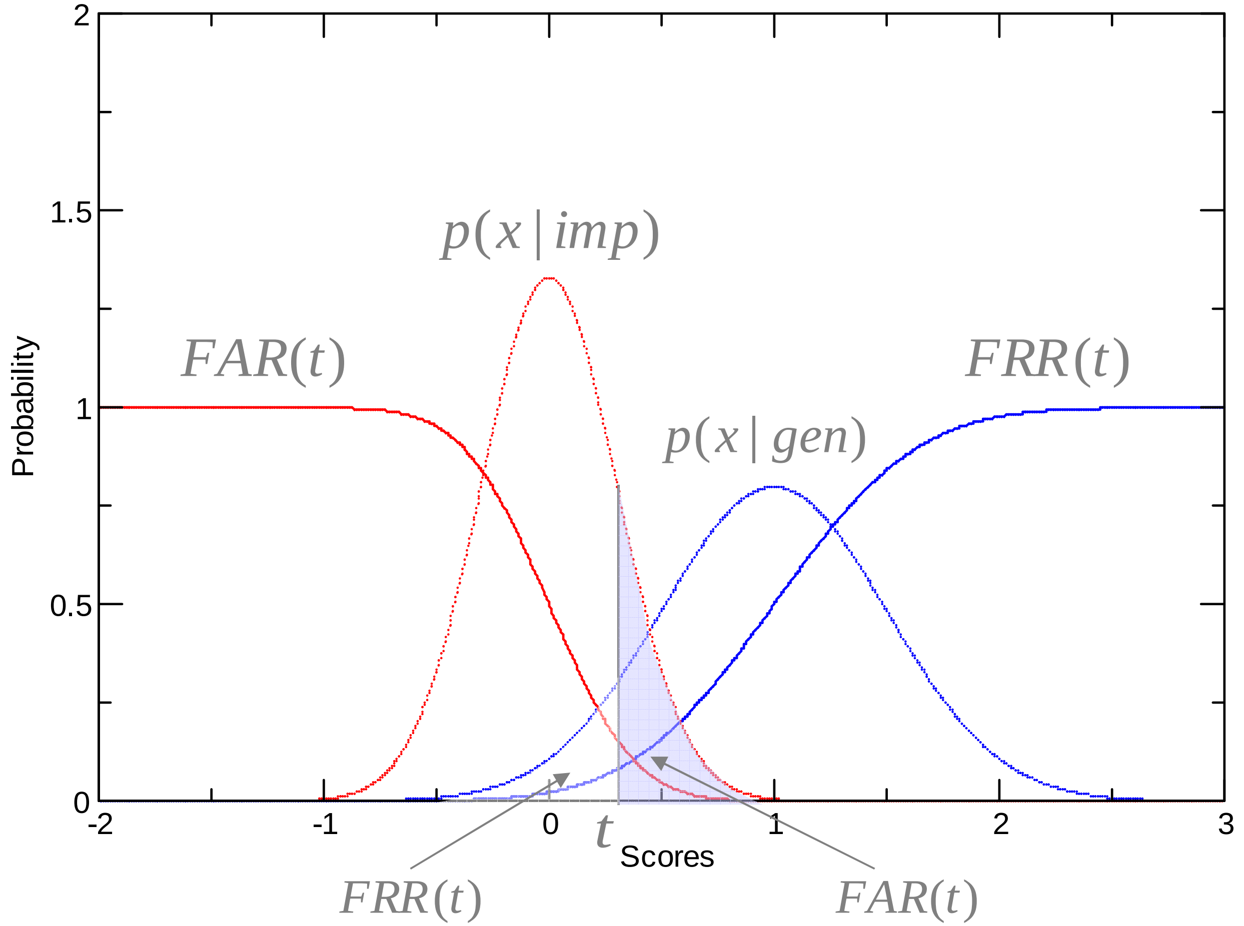


Figure 5.2: The non-match scores are on average lower than the match scores; in this case, the threshold T is set high to minimize False Accept.

Performance of Biometric Matchers



ROC Curve

ROC curve connects $FAR(\theta)$ and $FRR(\theta)$ curves.

Note that they both use same θ at the same time, so we are able to construct such plot.

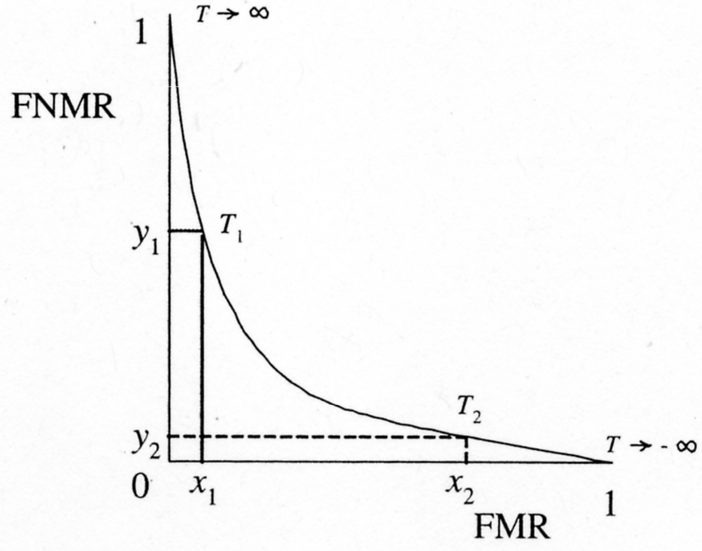


Figure 5.4: The ROC curve expresses the trade-off between FMR and FNMR.

Types of ROC Curve

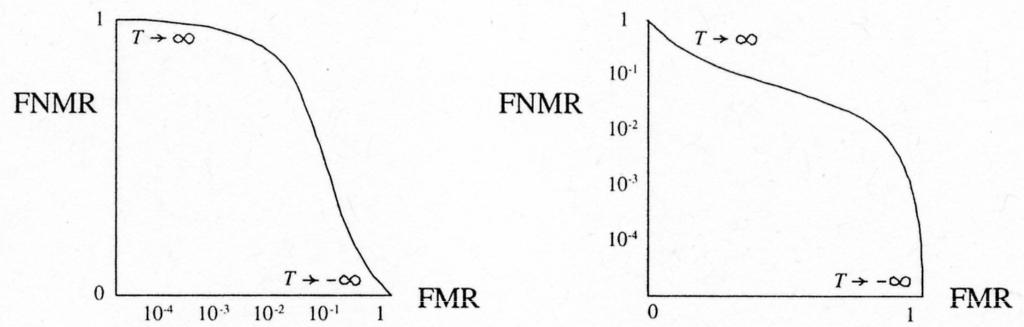


Figure 5.5: The ROC with one probability scale in logarithmic form; on the left the FMR is expressed in logarithmic form, on the right the FNMR is in logarithmic form.

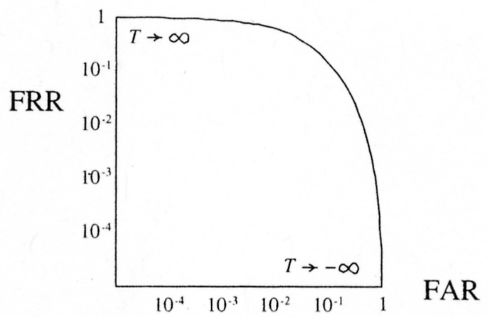


Figure 5.6: The ROC with both probability scales in logarithmic form.

Taking $\log(FAR(\theta))$ and $\log(FRR(\theta))$ instead of $FAR(\theta)$ and $FRR(\theta)$ is reasonable if they are small.

Types of ROC Curve

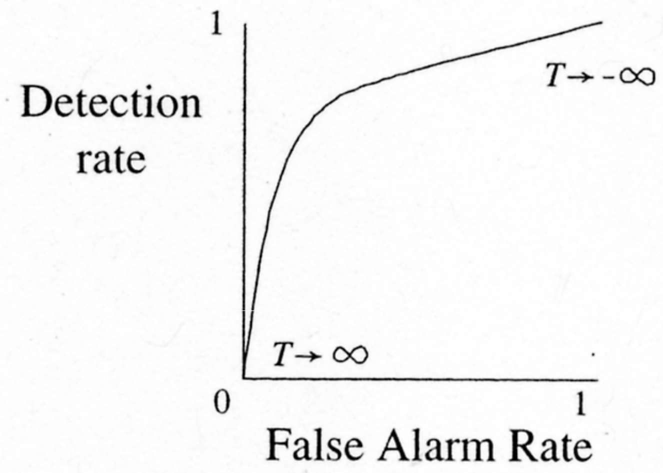


Figure 5.7: A detection error curve with the detection rate $(1 - \mathrm{FNMR})$, plotted against the False Alarm Rate.

Using ROC Curve

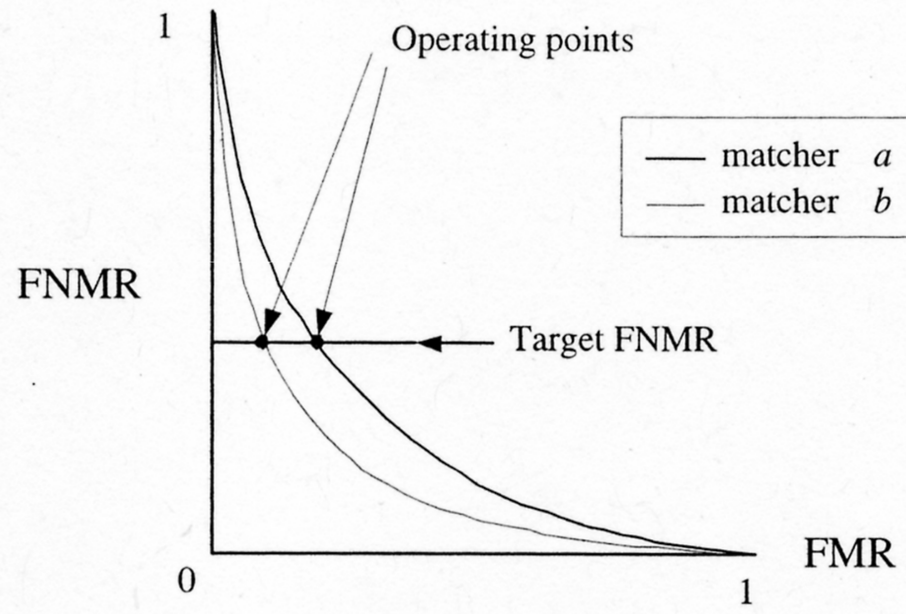


Figure 5.8: Two distinct ROC curves and a sample operating point specified with a target FMR. Matcher b is more accurate than a for all T.

Comparing ROC Curves

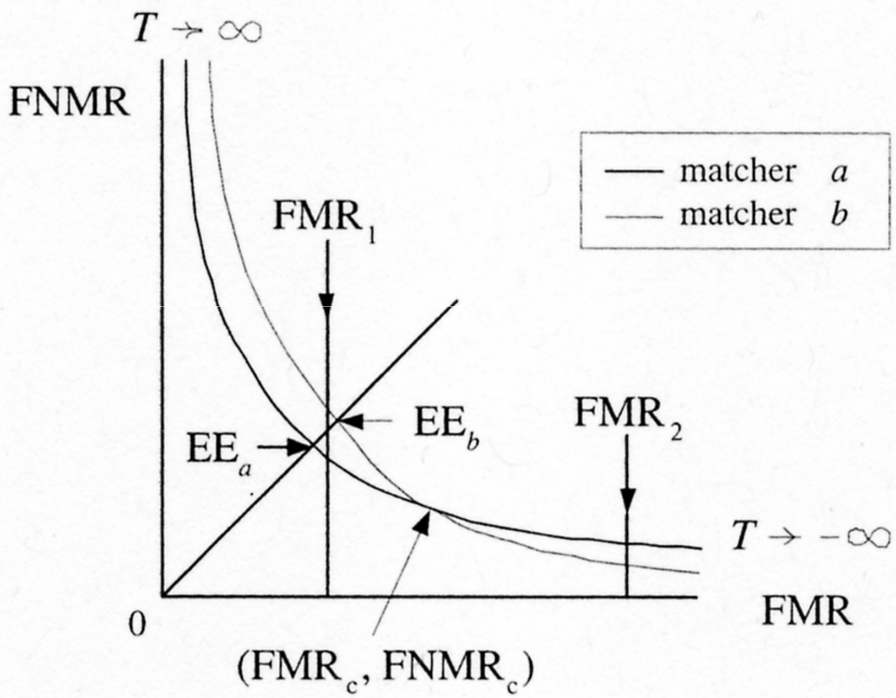
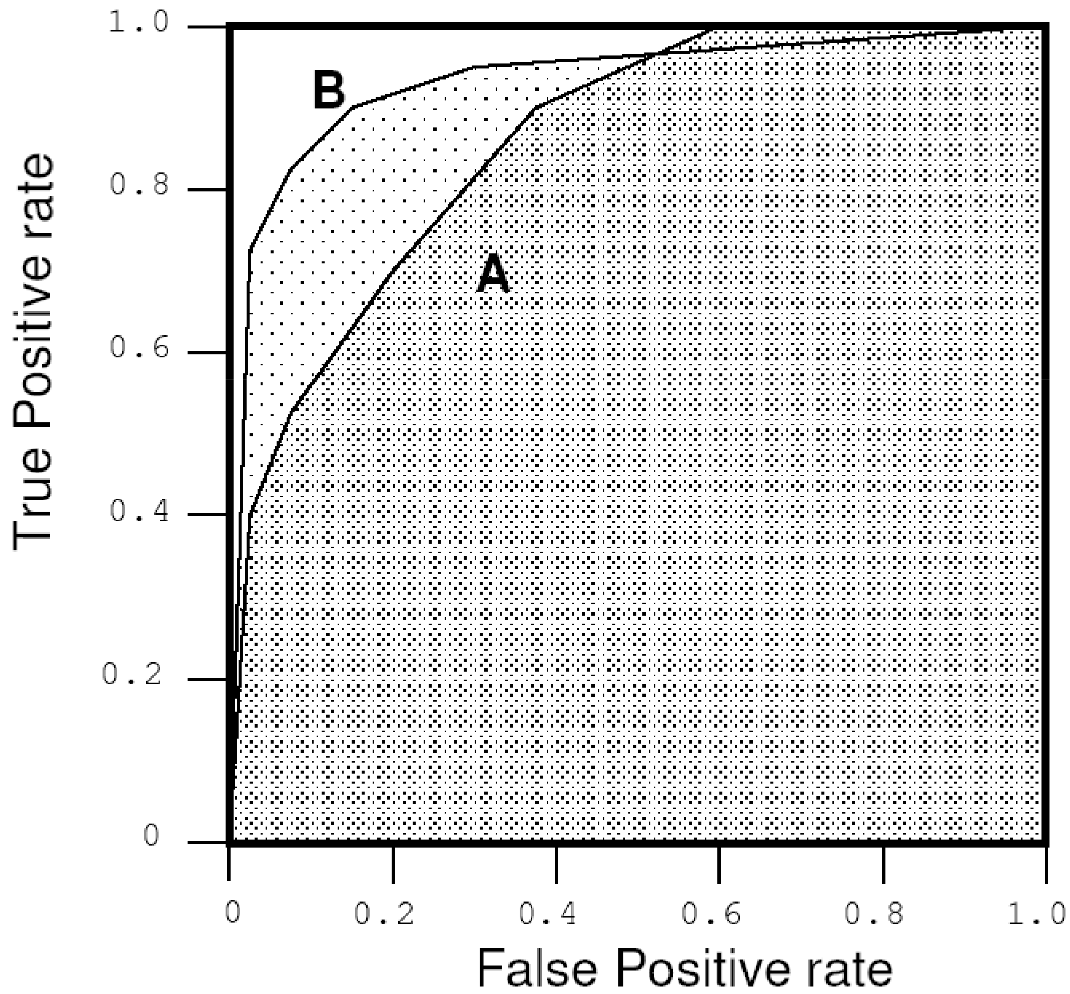


Figure 5.9: Which matcher is best can depend on the operating point chosen.

Comparing ROC Curves



Area under ROC curve (1-FRR vs FAR) represents the probability that random genuine score is higher than random impostor score.

Comparing ROC Curves

Compare match and non-match score densities by d-prime method:

$$d' = \frac{\mu_m - \mu_n}{\sqrt{\sigma_m^2 + \sigma_n^2}}$$

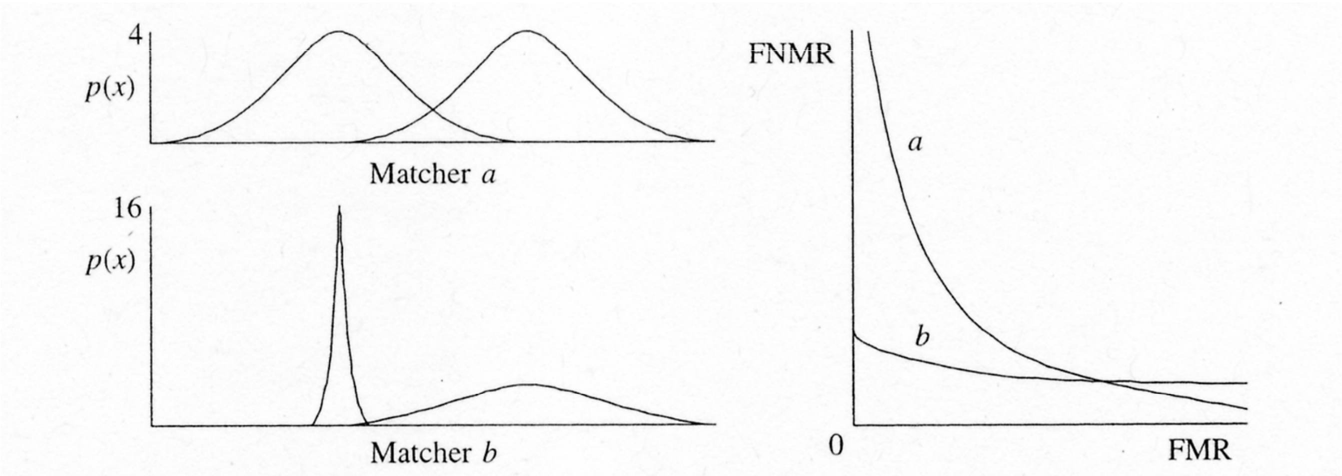


Figure 5.10: Different ROCs for two hypothetical matchers a and b with identical d'. Here Gaussian score distributions with identical means and different variances lead to the same d' but different ROCs.

Comparing ROC Curves

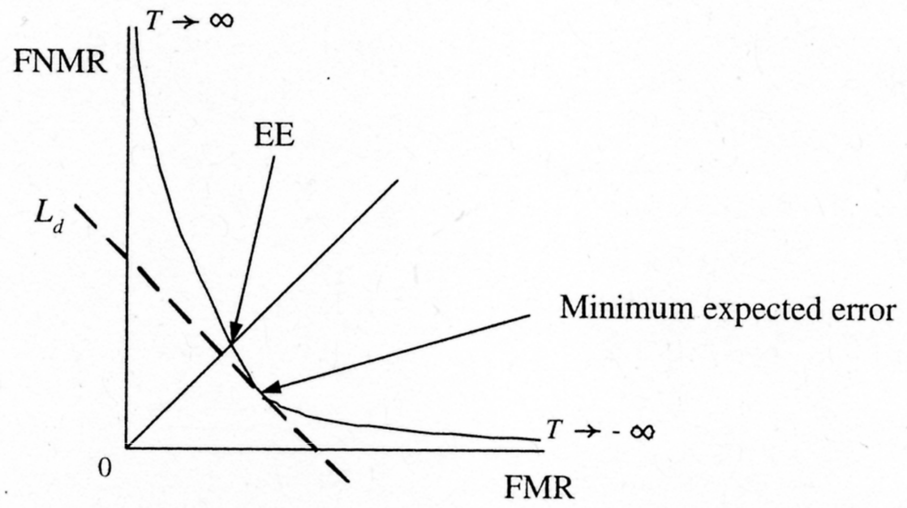


Figure 5.11: The minimum expected error will not generally be found at the same operating point as the Equal Error Rate.

Equal Error Rate (EER): $EER = FRR(\theta) = FAR(\theta)$
at θ such as $FRR(\theta) = FAR(\theta)$

Minimum Total Error Rate (TER):

$$TER = \min_{\theta} \; FRR(\theta) + FAR(\theta)$$

Trade-offs

Selection of the operating point in a particular application is a trade-off between security and convenience.

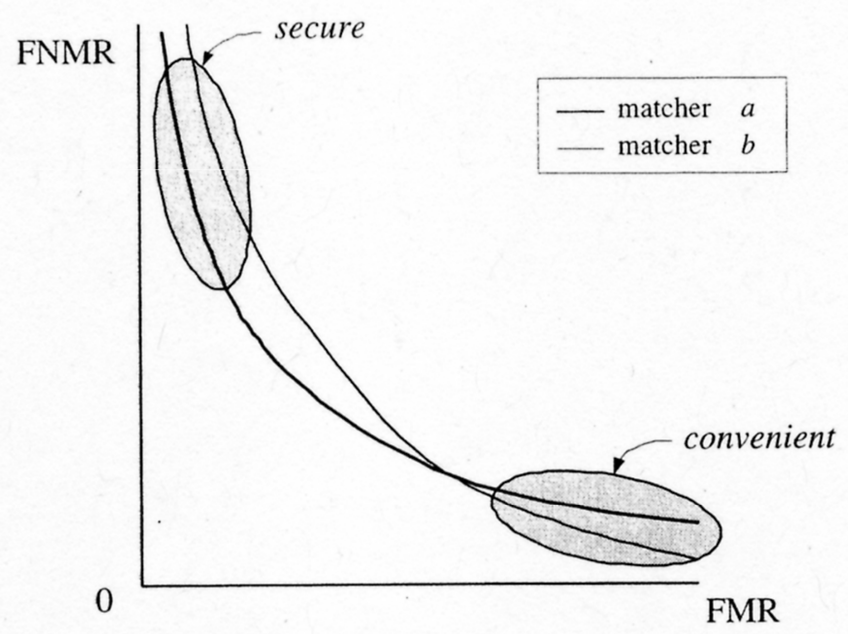


Figure 5.13: ROCs for Matcher a and b. Matcher b may be preferred for convenience and Matcher a for security..

Estimating FAR and FRR

In contrast to estimating pdf, FAR and FRR are easily estimated:

$$FAR(t) = \int_{x>t} p(x \mid imp)dx \approx \frac{\left|\{x_i \mid x_i > t, x_i \text{ is impostor }\}\right|}{\left|\{x_i \mid x_i \text{ is impostor }\}\right|}$$

$$FRR(t) = \int_{x<t} p(x \mid gen)dx \approx \frac{\left|\{x_i \mid x_i < t, x_i \text{ is genuine }\}\right|}{\left|\{x_i \mid x_i \text{ is genuine }\}\right|}$$

Types of ROC curves:

$$\{FRR(t), FAR(t)\}_{-\infty<t<\infty}$$

$$\{FAR(t),\ P(gen)(1-FRR(t)) + P(imp)FAR(t)\}_{-\infty<t<\infty}$$

$$\{\log\ FRR(t), \log\ FAR(t)\}_{-\infty<t<\infty}$$

Using FAR and FRR

In Bayesian framework we want to minimize total cost:

$$Cost = C_{FA} P(\text{impostor}) P(s > \theta \mid \text{impostor}) + C_{FR} P(\text{genuine}) P(s < \theta \mid \text{genuine})$$

$$= C_{FA} P_{imp} FAR(\theta) + C_{FR} P_{gen} FRR(\theta)$$

Correct setting of θ in verification application requires estimating $C_1, C_2, P(\text{impostor}), P(\text{genuine})$

Example

Consider the problem of deploying biometric matcher for an amusement park admission

$C_{FA} = \$20$ - cost of accepting impostor to the park

$P_{imp} = 1\%$ - probability of impostor attempts

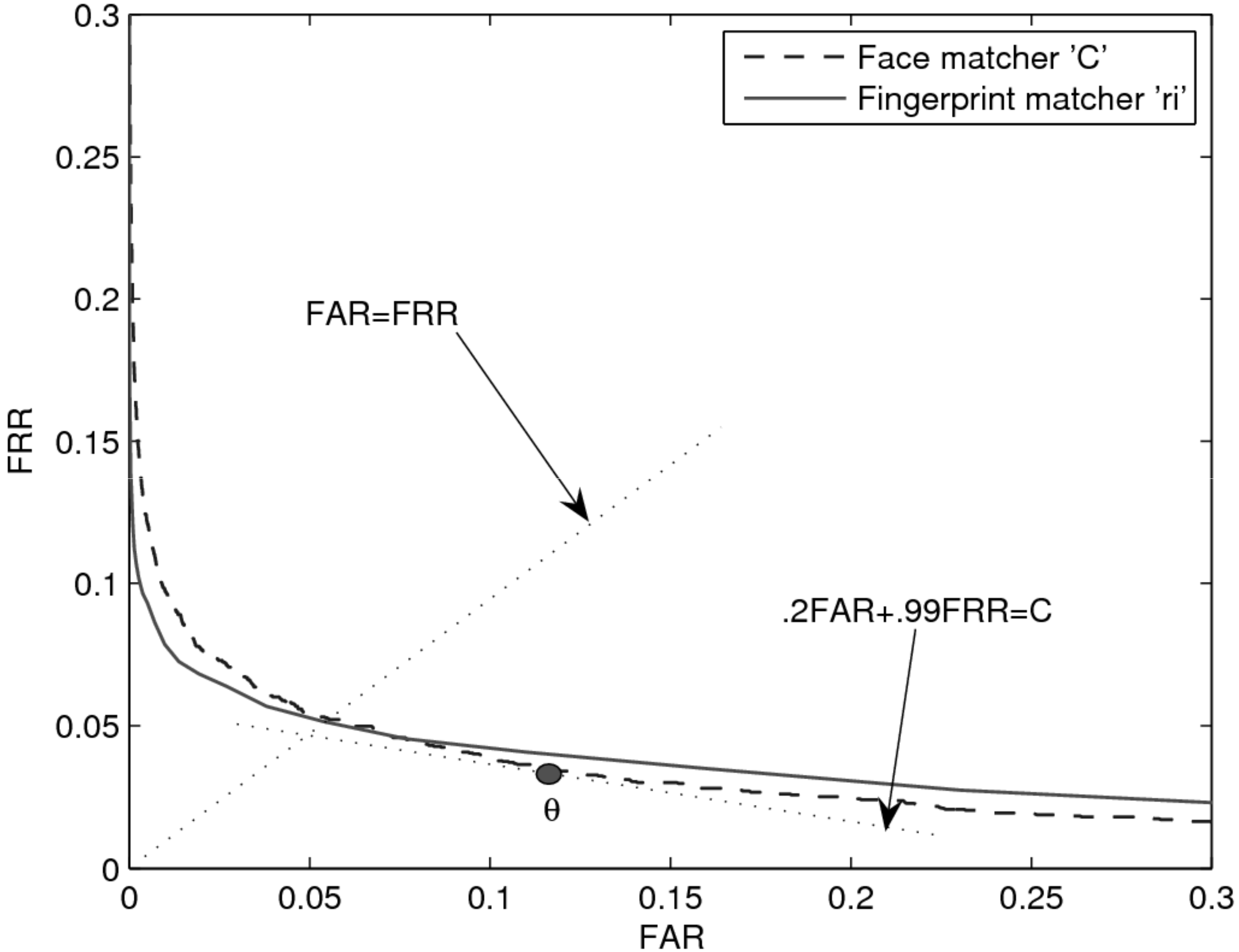
$C_{FR} = \$1$ - cost of rejecting genuine user

$P_{gen} = 99\%$ - probability of genuine attempts

$$Cost = C_{FA} P_{imp} FAR(\theta) + C_{FR} P_{gen} FRR(\theta)$$

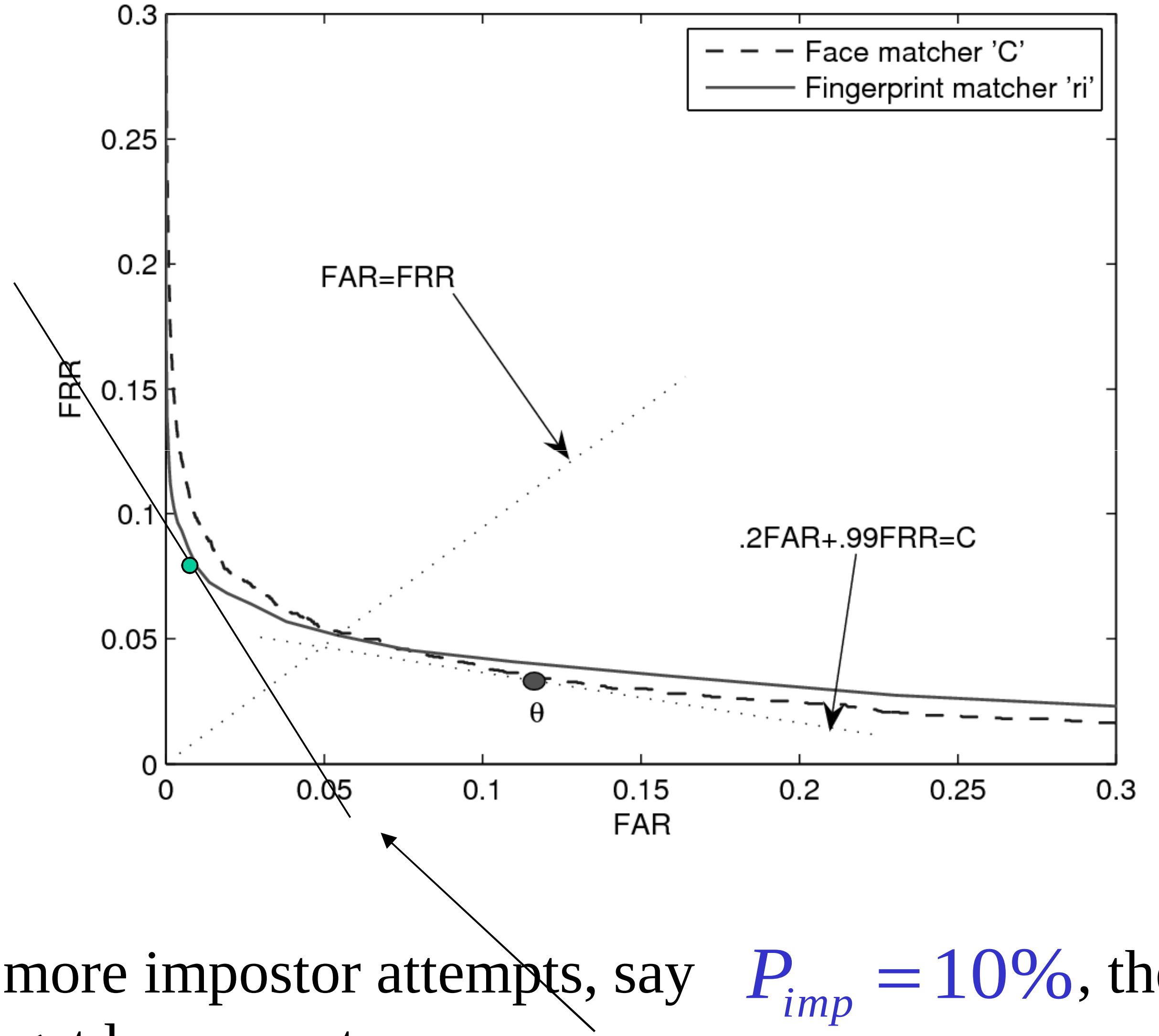
$$= 20 \times .01 \times FAR(\theta) + 1 \times .99 \times FRR(\theta)$$

$$= .2 \times FAR(\theta) + .99 \times FRR(\theta)$$



Face matcher ‘C’ better minimizes cost

$$Cost = .2 \times FAR(\theta) + .99 \times FRR(\theta)$$



If we had more impostor attempts, say $P_{imp} = 10\%$, then matcher 'ri' would get lower cost

$$Cost = 2 \times FAR(\theta) + .9 \times FRR(\theta)$$

Errors in Identification Systems

N people are enrolled in the database. The recognition algorithm performs N matchings with output scores:

$$s_1 > s_2 > \ldots > s_N$$

(the scores are ordered by magnitude, but not by people id)

The decision algorithm usually considered:

- Accept class 1 if

$$s_1 > \theta \text{ and } \theta > s_2 > \ldots > s_N$$

- Reject otherwise

Errors in Identification Systems

Other types of decisions involve selecting a subset of matched classes:

- Threshold based:

$$s_1 > s_2 > \ldots > s_k > \theta$$

-select all classes bigger than threshold

- Rank –based:

-select k classes with best scores

- Hybrid:

-select based on threshold, if not successful select k classes based on rank

FNMR and FMR in Identification Systems

FNMR – False non-match rate:

$$FNMR(\theta) = FRR(\theta) = \int_{-\infty}^{\theta} p_{gen}(s)ds = P(s < \theta \mid \text{genuine})$$

FMR – False match rate:

$$FMR(\theta) = P(\max s_i > \theta \mid i \text{ corresponds to all N-1 impostor event})$$

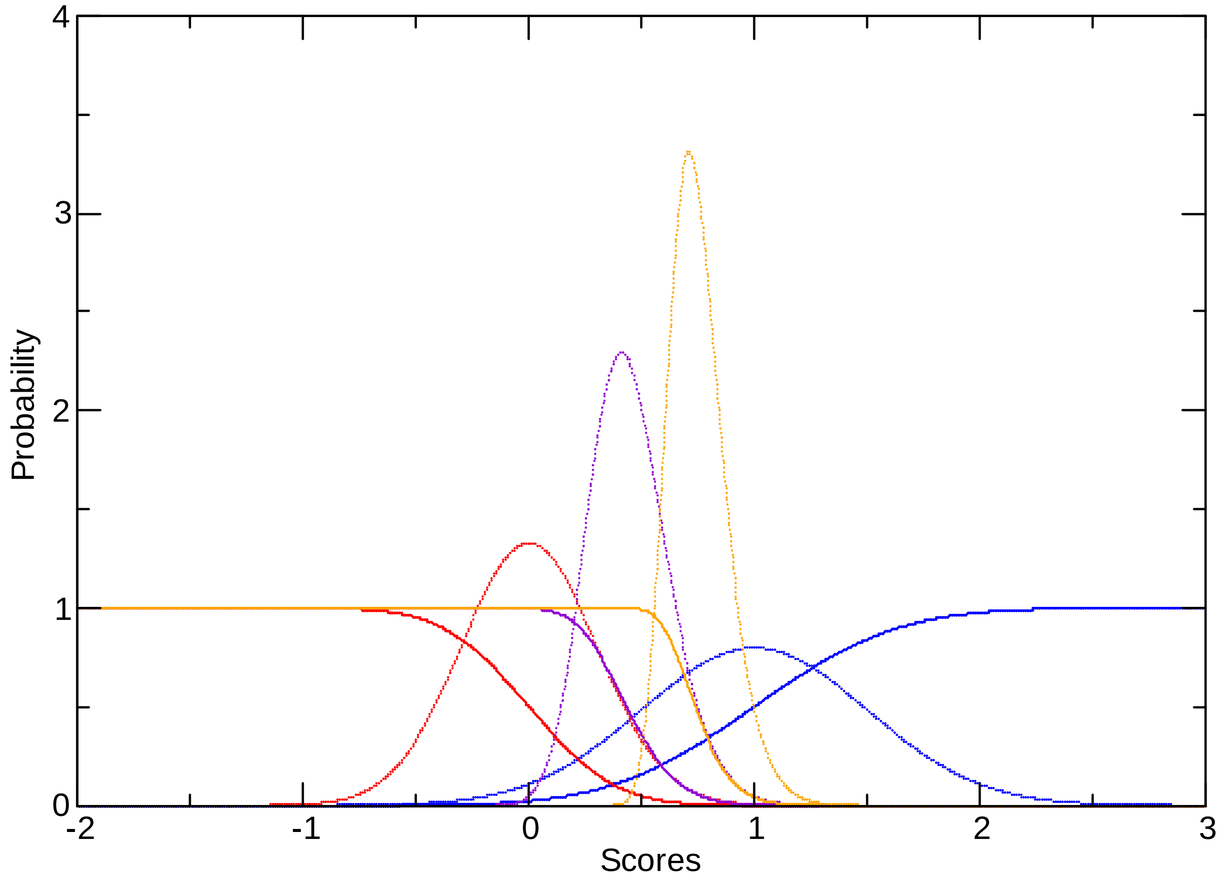
$$= 1 - P(s_i < \theta \mid i \text{ corresponds to all N-1 impostor event})$$

$$= 1 - \prod_i P(s_i < \theta \mid i \text{ corresponds to one impostor event})$$

$$= 1 - \prod_i (1 - P(s_i > \theta \mid i \text{ corresponds to one impostor event}))$$

$$= 1 - \prod_i \left(1 - \int_{\theta}^{\infty} p_{imp}(s)ds \right) = 1 - (1 - FAR(\theta))^{N-1}$$

FMR for different N



Errors in Identification Systems

• FMR and FNMR might not adequately describe the performance of identification systems

- closed set / open set identification
- rejecting all identification results might be a correct choice
- errors are connected: impostor might be a top choice, but genuine is also higher than the threshold

• Score belonging to different classes are usually dependent, so FMR can not be effectively estimated by means of FAR

• Still no good standard for measuring identification system performance exists

Investigating validity of i.i.d. assumption

Example: Identification system with 2 classes – genuine and impostor

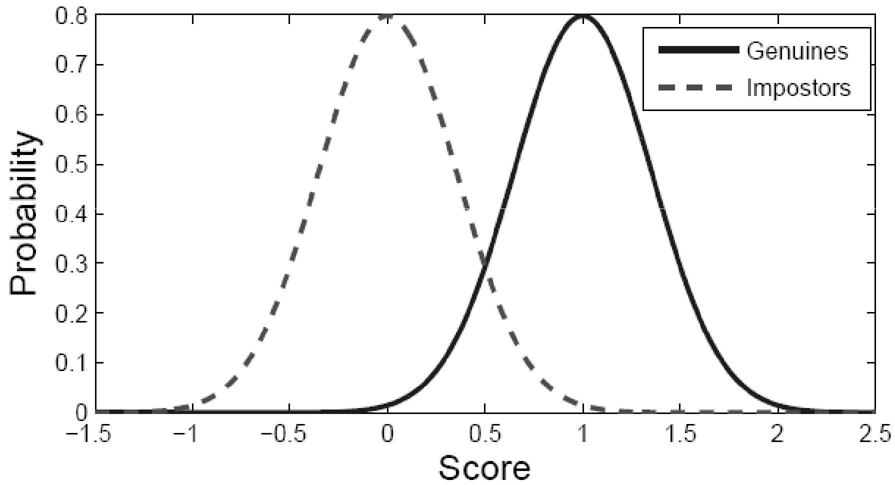


Fig. 1. Hypothetical densities of matching(genuine) and non-matching(impostors) scores.

Consider two possible scenarios on how the matching scores are generated during an identification attempt:

1) Both scores s_{gen} and s_{imp} are sampled independently from genuine and impostor distributions.
2) In every identification attempt : $s_{imp} = s_{gen} - 1$.

Scenario 1: CorrIdent<1
Scenario 2: CorrIdent =1

Dependence of scores influences performance in identification systems