Anomaly Detection

Josef Kittler

Dept. Electronic Engineering, University of Surrey
J.Kittler@surrey.ac.uk

Support by EPSRC and dstl is gratefully acknowledged

Aims

- To introduce the subject of anomaly detection, its content and relevance
- To introduce the terminology of anomaly detection
- To riview/introduce the mathematical background required

Outline

- Introduction to anomaly detection
- Problem formulation
- Statistical hypothesis testing
- One class classification (SVM)
- Critique of classical anomaly detection
- Complementary mechanisms for anomaly detection
- Anomaly detection system architecture
- Incongruence detection
- Dempster Shaffer reasoning (Prof David Parish)

Introduction to anomaly

- *Anomaly –*
  - an important notion in human understanding of the environment
  - deviation from normal order or rule
  - failure to relate sensor data to a meaning
  - manifest in weak or no support for domain specific hypotheses
- *Many synonyms signifying different nuances*
  - rarity, irregularity, incongruence, abnormality, unexpected event, novelty, innovation, outlier

- *In science/engineering*
  - prove disprove hypothesis
  - fault detection
  - outdated model requires adaptation

Diverse applications

- Many applications formulated as anomaly detection problems
  - surveillance
  - novel object detection
  - abnormal communication network activity
  - medical diagnostics
  - video segmentation
  - suspicious behaviour

Anomaly detection problem formulations

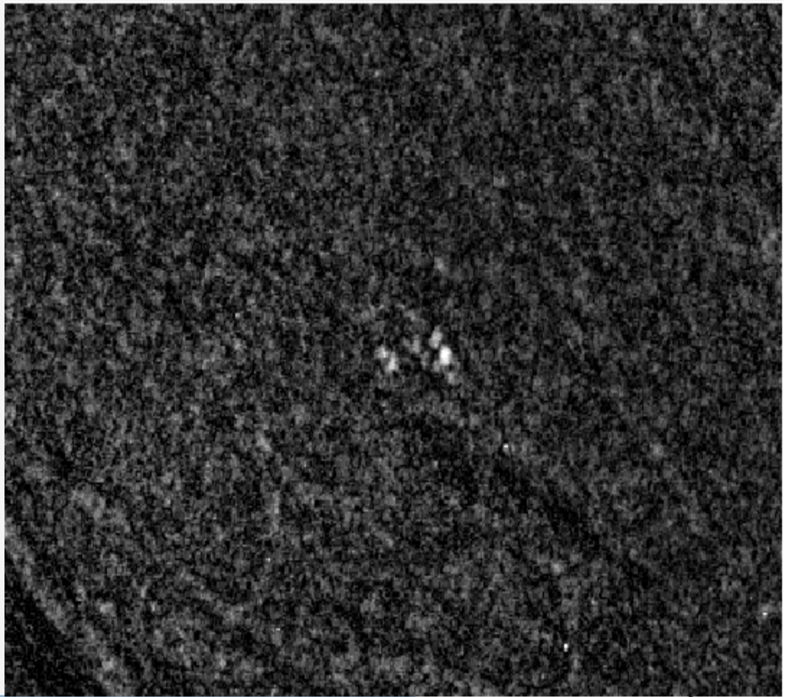
- Classification problem
  - Abnormality types known
- Detection problem
  - Samples of normal class and negative examples available
- Hypothesis testing problem
  - Only samples of normal class available
  - One class classification problem

Prior art in anomaly detection

- Edgeworth (1888)
- Hundreds of papers
- Many approaches
  - statistical, NN, classification, clustering, information theoretic, spectral
- Excellent surveys
  - Markou&Singh (SP 2003, statistical, neural)
  - Hodge&Austin (AI Review 2004)
  - Agyemang&Barker&Alhajj (Int Data Anal 2006)
  - Chandola&Banerjee&Kumar (ACM Surveys 2010)
  - Saligrama&Konrad&Vodoin (SPM2010, video)

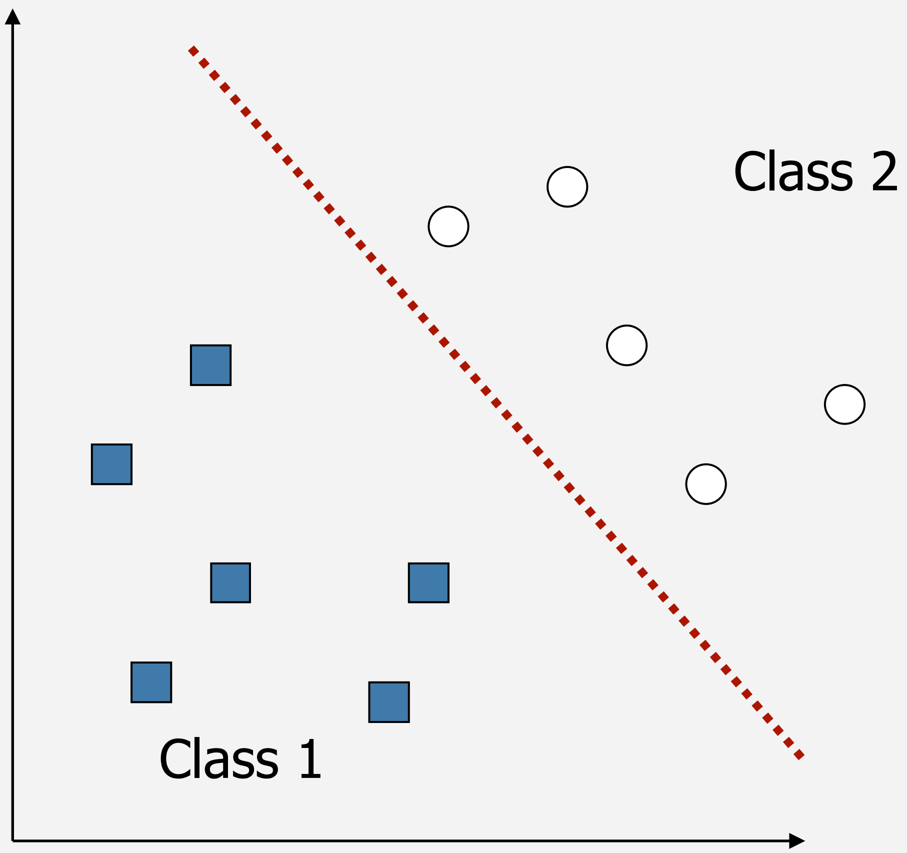
Anomaly detection as a problem in classification

- Microcalcification detection: anomaly in tissue texture
- Anomaly class known
- Anomaly detection solved as a classification problem

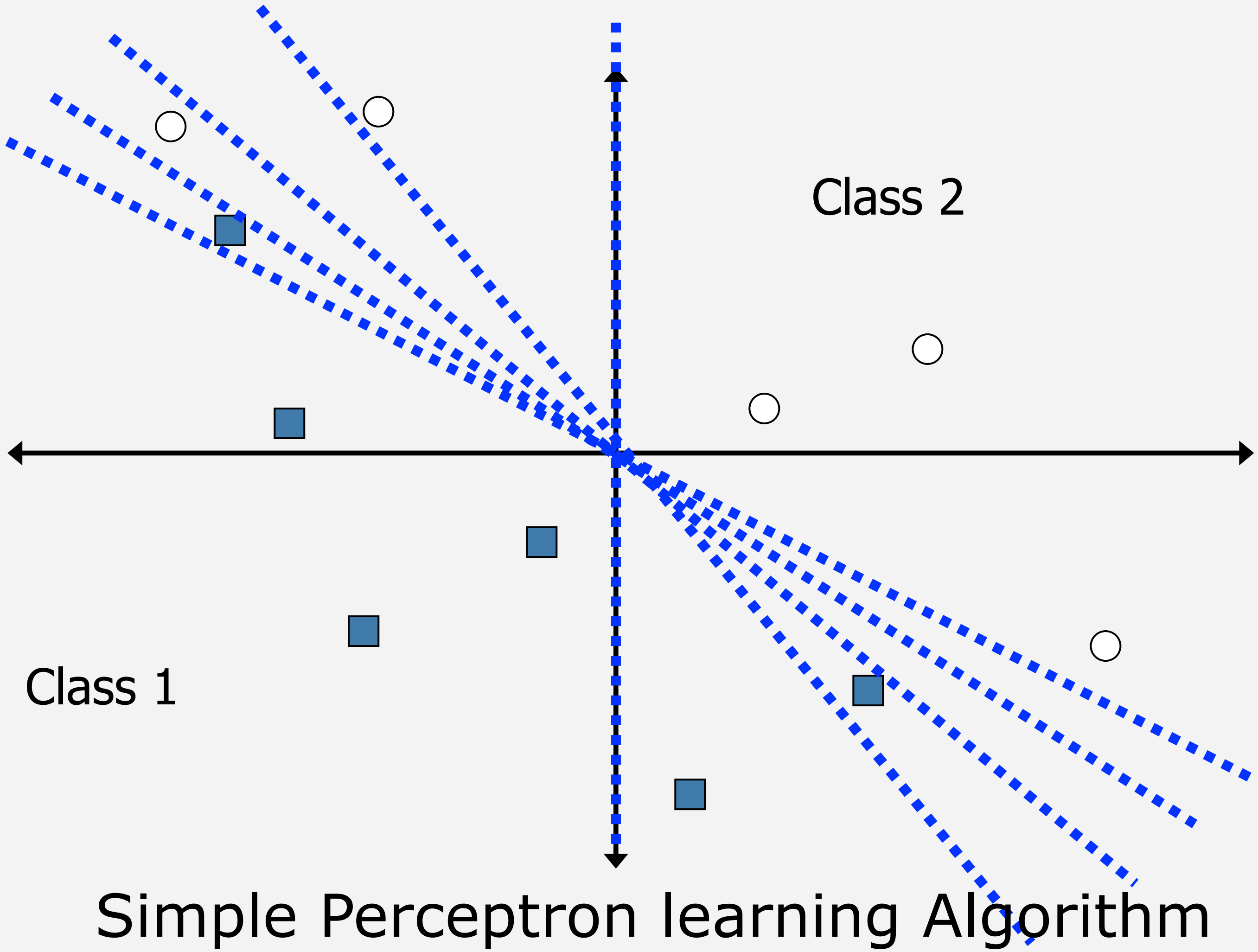


Two Class Problem

- Many decision boundaries can separate these two classes.

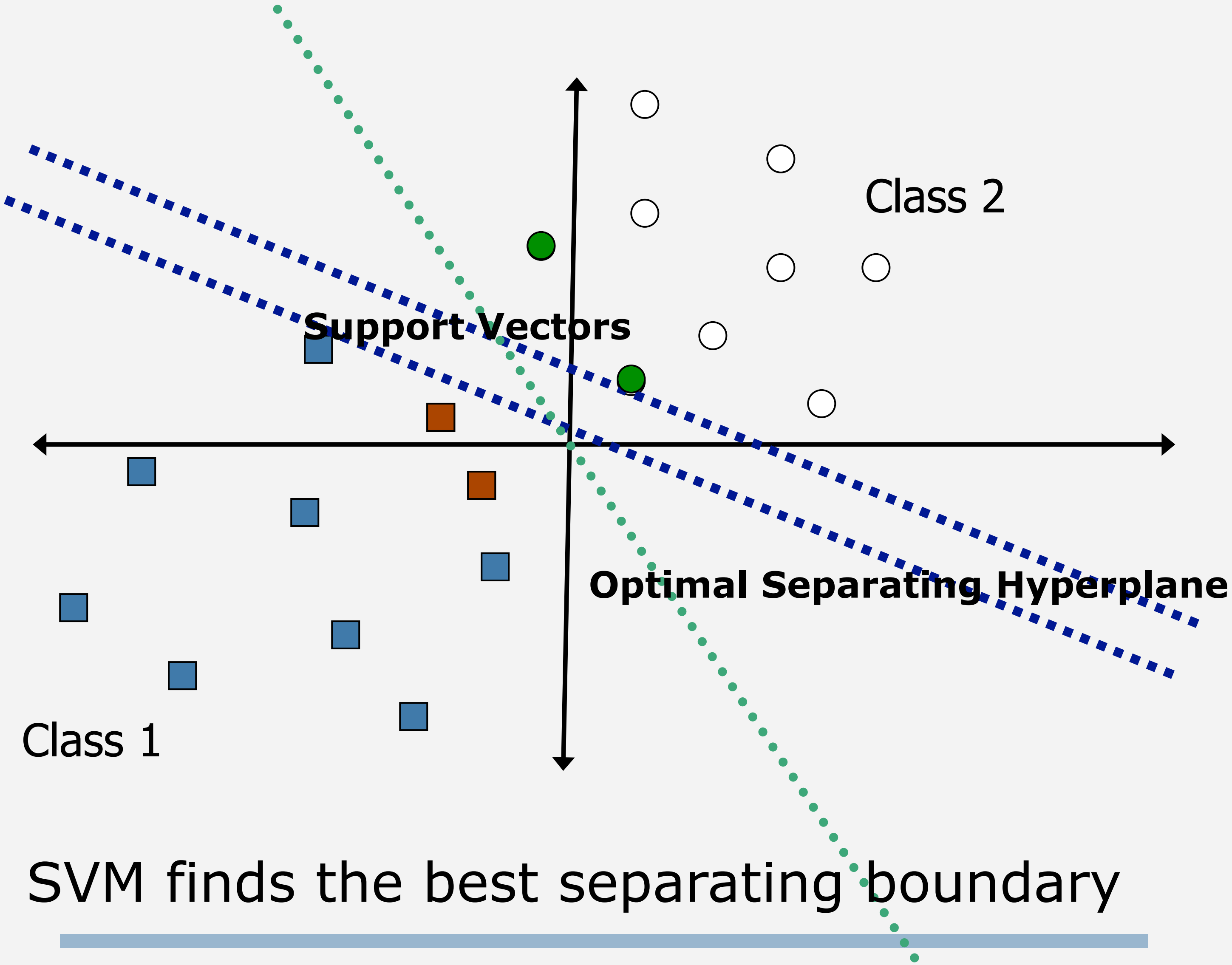


Classification



Simple Perceptron learning Algorithm

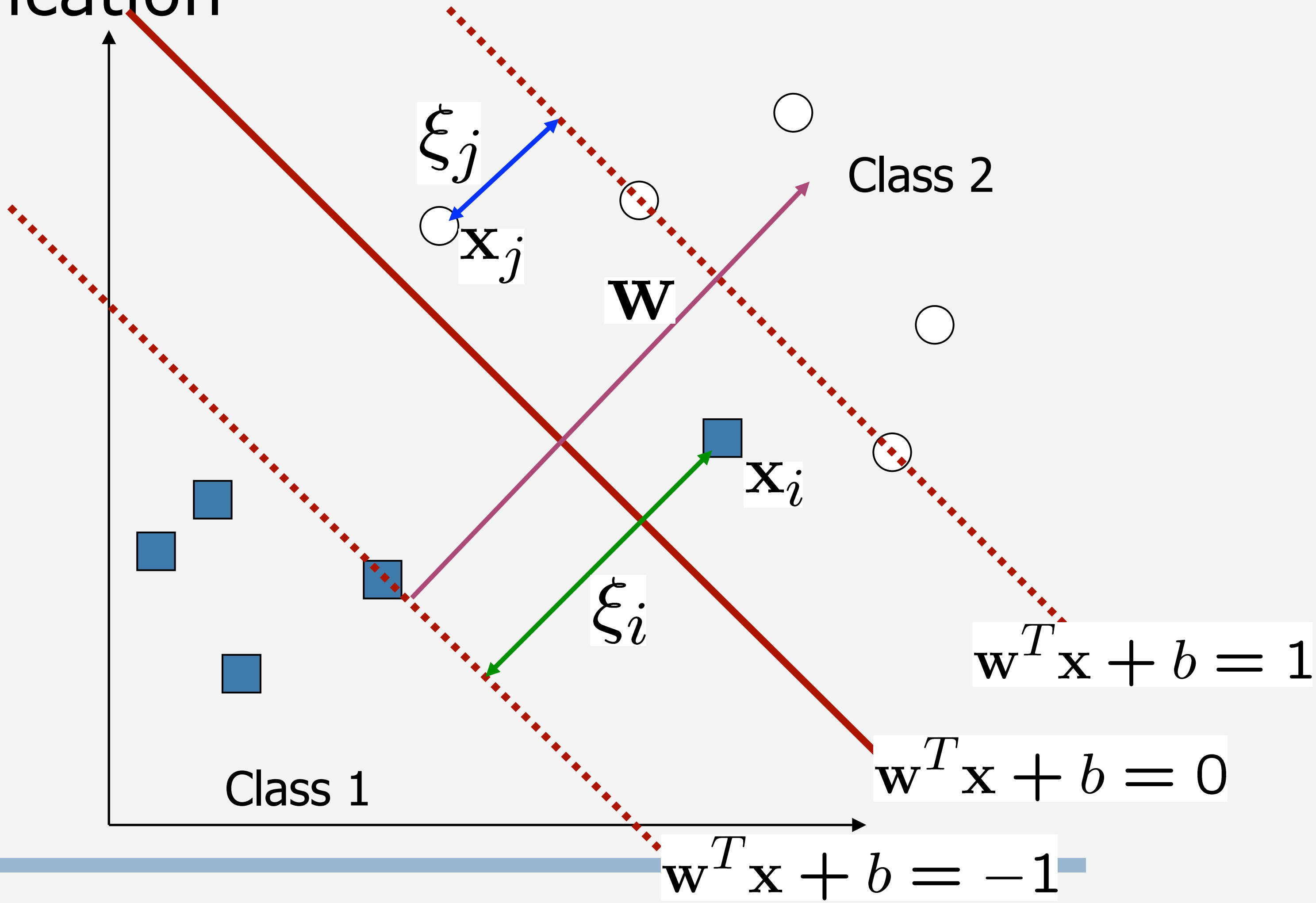
Support Vector Machine (SVM)



SVM finds the best separating boundary

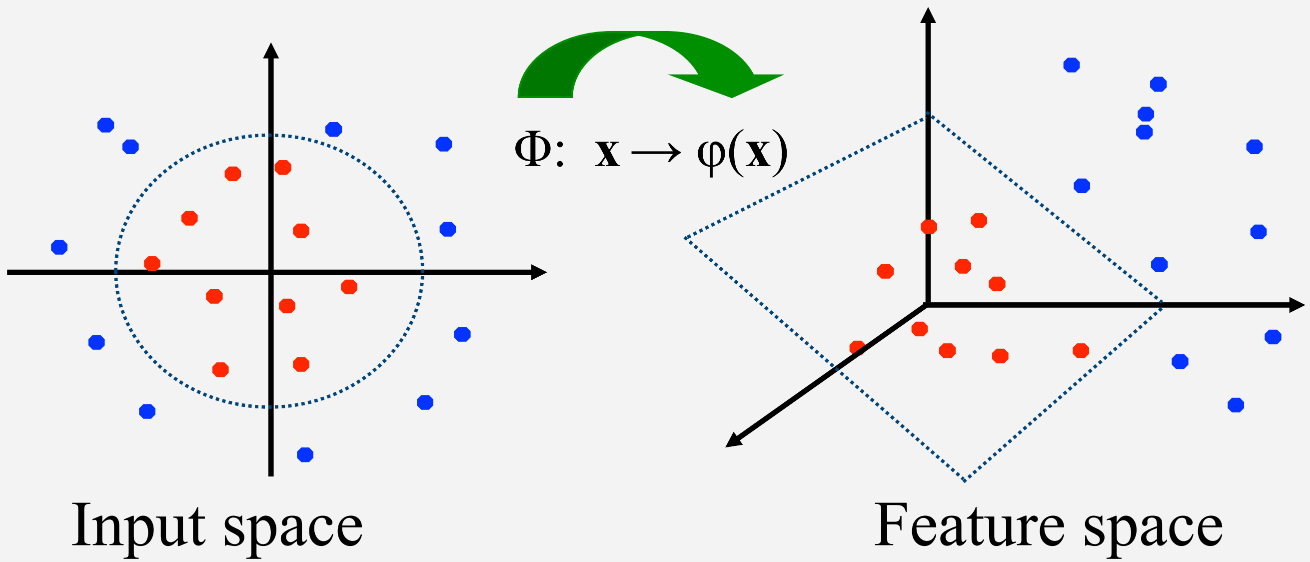
If not Linearly Separable

- Slack variable ξ_i we allow “error” in classification



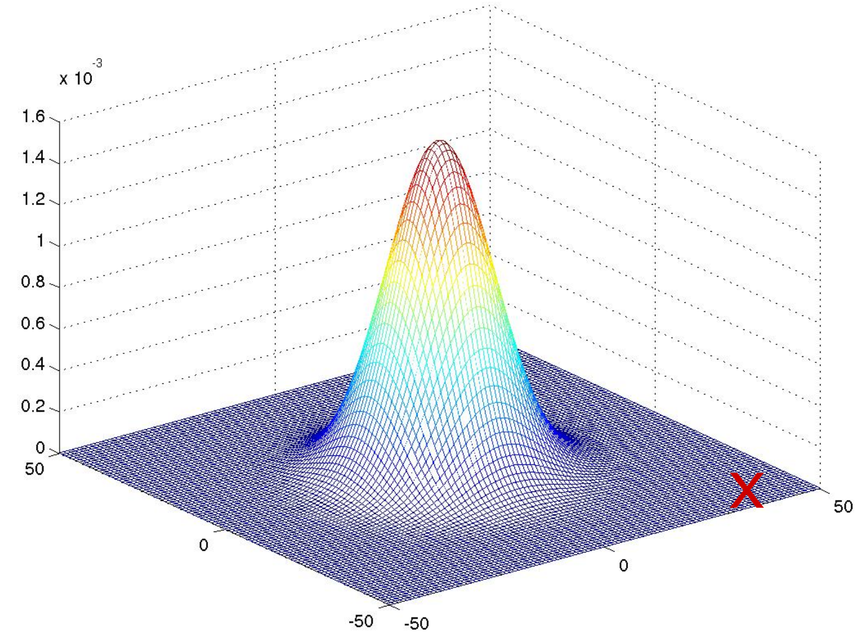
Extension to Non-linear Decision Boundary

- SVM solves this using kernel function
  - Kernel tricks for efficient computation
  - Minimizing $||\mathbf{w}||^2$ produces a "good" classifier



Classical anomaly model

- *Conventional mathematical model*
  - outlier of a distribution
  - empirical distribution deviates from the model distribution



Hypothesis testing

- This typically involves some proposition, referred to as a null hypothesis and a test statistics.
- If the outcome of the test statistics is consistent with its known distribution model $p(x)$, then the null hypothesis is accepted.
- An outlier of that distribution would lead to the hypothesis rejection.
- Example: pdf is uniform over support domain S

$$p(x) = const \quad x \epsilon S$$

- For any x outside S the hypothesis would be rejected

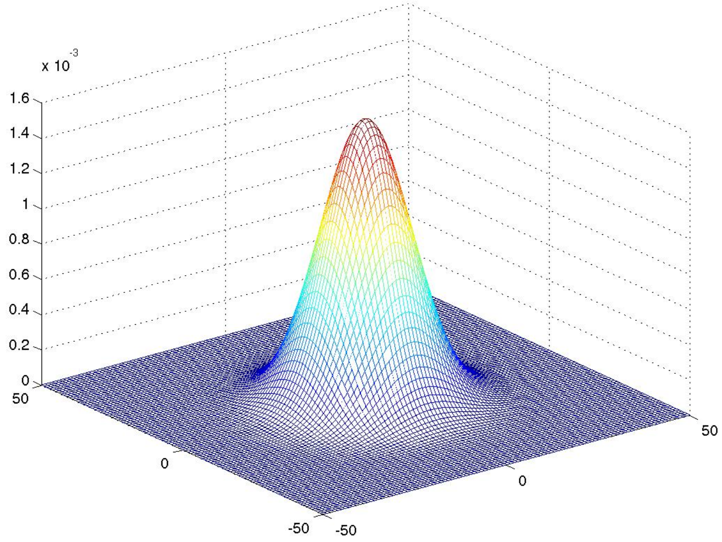
Normal (Gaussian) distribution

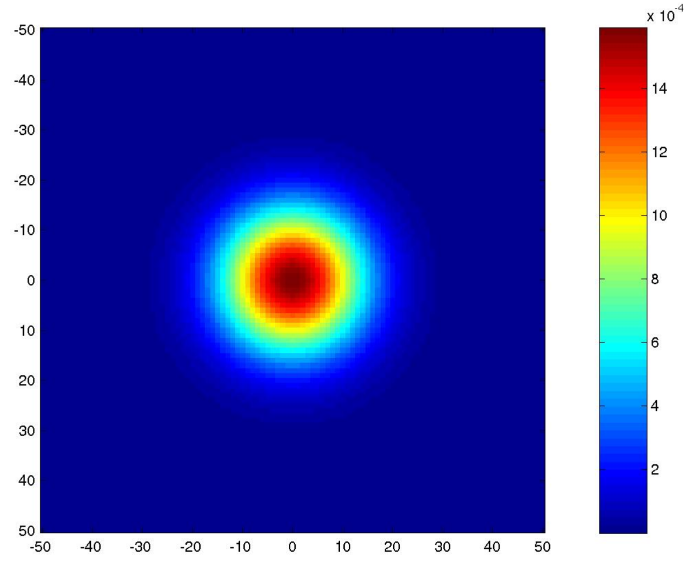
- Gaussian distribution

$$p(x) = [(2\pi)^n|\Sigma|]^{-\frac{1}{2}} exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

- where μ is its mean vector and Σ is the covariance matrix
- Gaussian extends to infinity, hence technically no observation is an outlier
- An observation is considered an outlier at a given level of significance, i.e. if the test statistics value is beyond a boundary corresponding to some vestigial probability outside it, such as 5% or 1%.

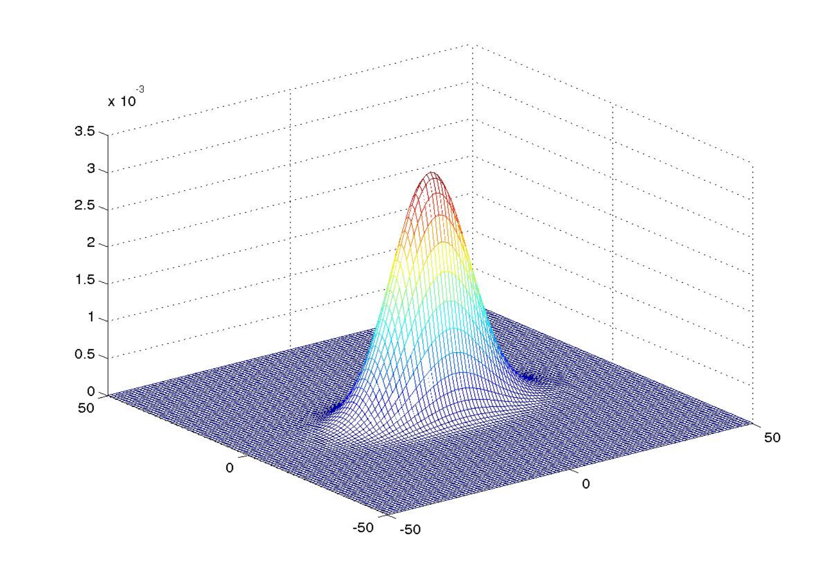
Examples of gaussians

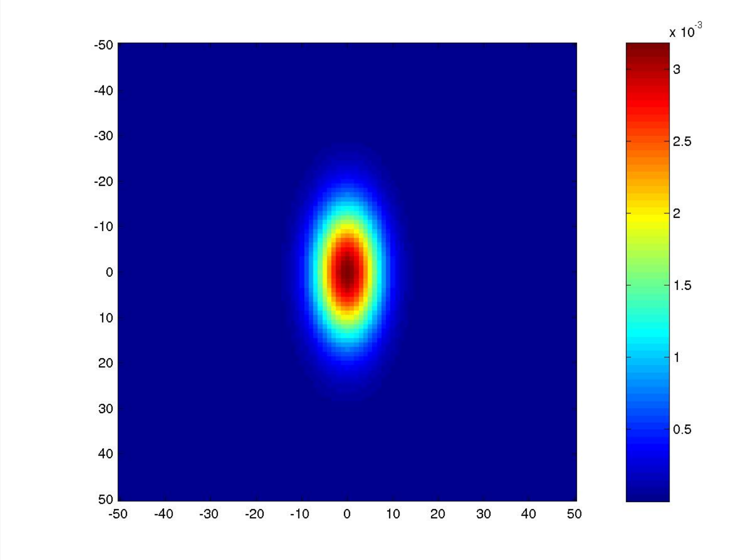




$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$$

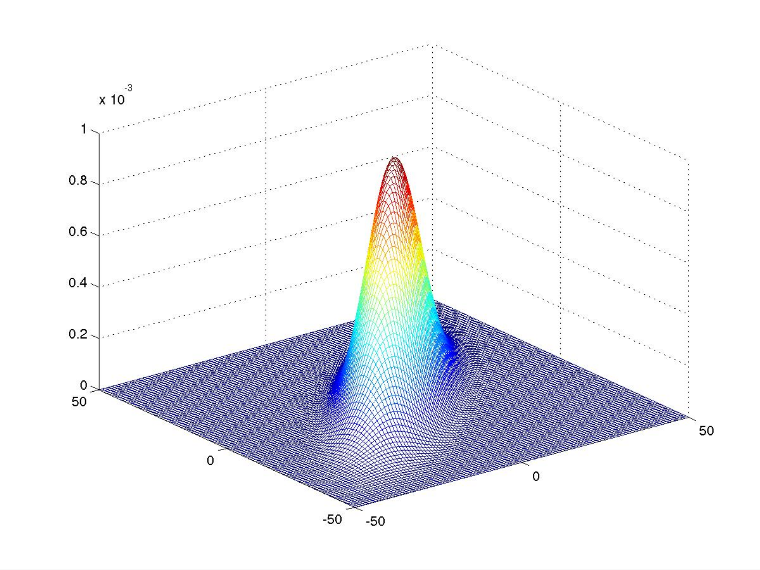
Examples of gaussians

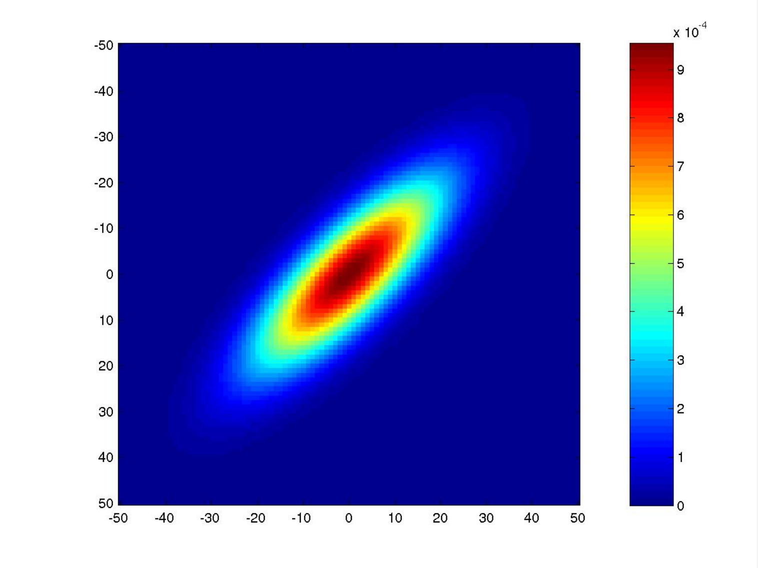




$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 100 & 0 \\ 0 & 25 \end{bmatrix}$$

Examples of gaussians

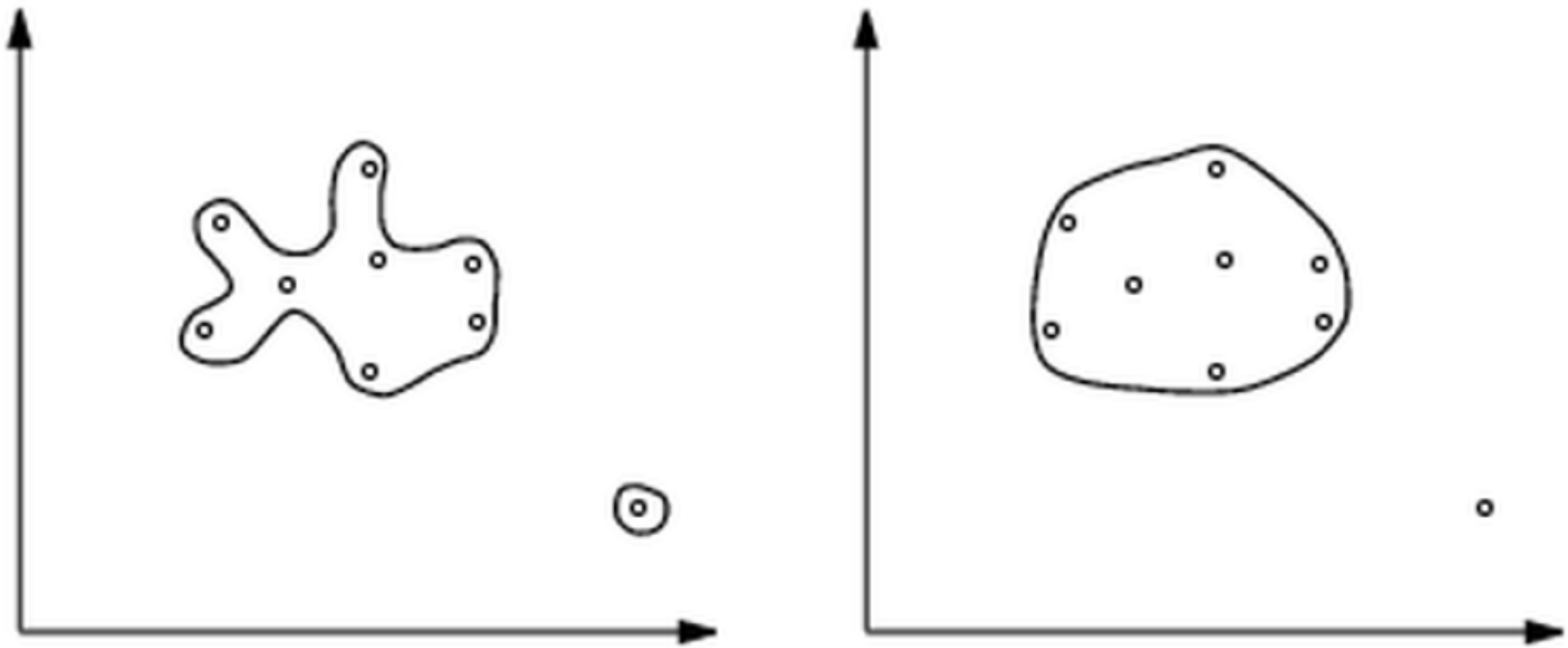




$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 60 & 40 \\ 40 & 60 \end{bmatrix}$$

Anomaly detection as one class classification

- Consider a set of points $X = \{x_1, ..., x_N\}$ where x_i is a realisation of a multivariate random variable x drawn from a probability distribution with probability density function $p(x)$.



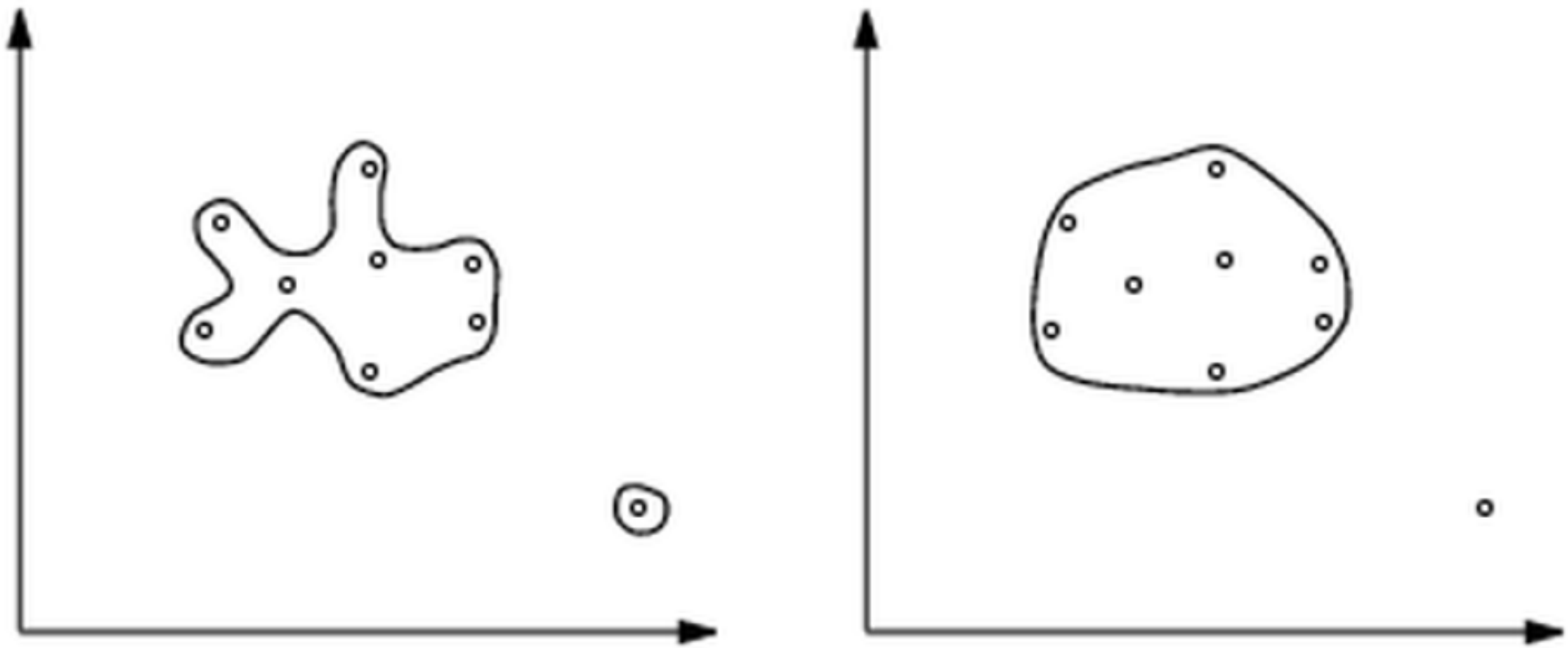
Data support domain estimation

- We would like to estimate the support domain S of x so that its future observations lie within S with probability $1-\alpha$ where parameter α is the confidence level specified by the user.

- Fundamentally different from the two class formulation
- Possible approaches
  - Parametric/nonparametric density estimation
  - Quantile function estimation
  - Convex hull enclosure
  - One class SVM

One class SVM

- The aim of one class SVM is to enclose the available one class training set
- Solution should generalise well



Kernel space

- We look for a solution in the feature space $\Phi(x)$ using the kernel representation, i.e.

$$k(x, y) = \Phi(x)^T \Phi(y) \quad (1)$$

- The kernel function, e.g. Gaussian, defines high dimensional feature space implicitly
- The solution defined in terms of a linear boundary in the feature space

$$f_{w,\rho}(x) = sgn[w^T \Phi(x) - \rho] \quad (2)$$

where w is a weight vector and ρ is an offset parametrising the hyperplane defining the boundary.

Objective function

- The function $f_{w,\rho}(x)$ takes value 1 for $x \epsilon S$ and -1 elsewhere

- It delineates the training set at a specified level of confidence

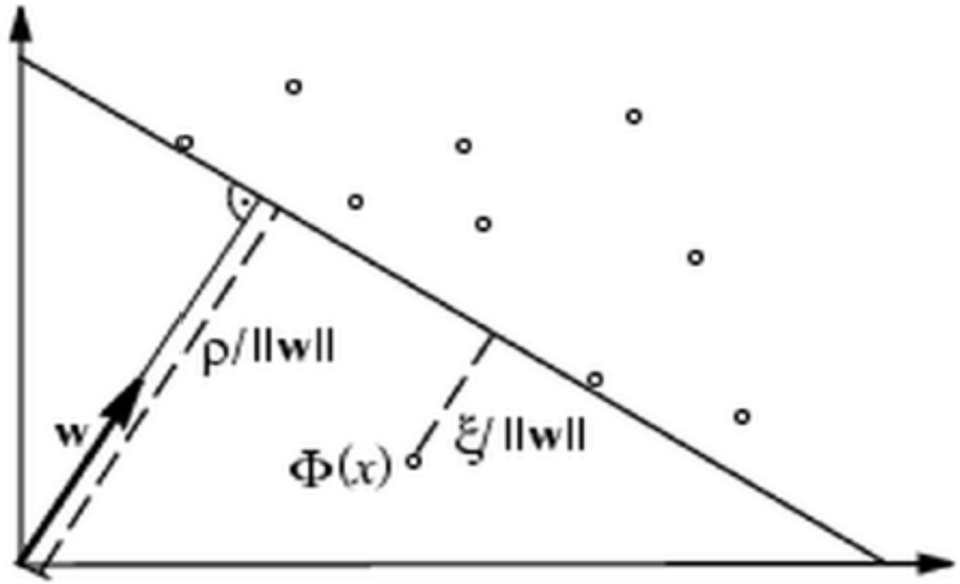
- The function can be learnt by minimising objective function

$$R[f_{w,\rho}(x)] = R^{emp}[f_{w,\rho}(x)] + w^T w \qquad (3)$$

where R^{emp} measures the empirical risk, that is misclassification of points in the training set and the term $w^T w$ regularises the solution by looking for the maximum margin between the training data and the origin.

Slack variables

- The hyperplane $w^T\Phi(x) = \rho$ separates the training set from the origin.
- Overfitting to data is minimised by allowing "outlier" training points to fall on the wrong side of the boundary. However, their number is controlled by penalising such points x_j by employing slack variable ξ_j



Constrained optimisation

- The use of slack variables and the regularisation term control the trade-off between empirical risk and overfitting
- The optimisation problems can be stated as

$$\begin{array}{l} \min_{w,\xi,\rho} \frac{1}{2}||w||^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \\ subject\ to\ w^T\Phi(x) \geq \rho - \xi_i,\ \ \xi_i \geq 0 \end{array} \quad (4)$$

where $\nu\epsilon(\alpha, 1]$ denotes the upper bound on the training data points that may be outliers

- Solve by method of Lagrange multipliers

Dual optimisation problem

- Accordingly, the two constraints are introduced into the objective function with the associated coefficients β_i and γ_i respectively
- This leads to the dual optimisation problem

$$\begin{array}{l} \min \frac{1}{2} \sum_{i,j} \beta_i \beta_j k(x_i, x_j) \\ subject\ to\ 0 \leq \beta_i \leq \frac{1}{\nu N},\ \sum_i^N \beta_i = 1 \end{array} \tag{5}$$

- The Kunt-Tucker conditions imply that if $\beta_i > 0$ and $\gamma_i > 0$ the inequality constraints become equality constraints
- ρ can be recovered as

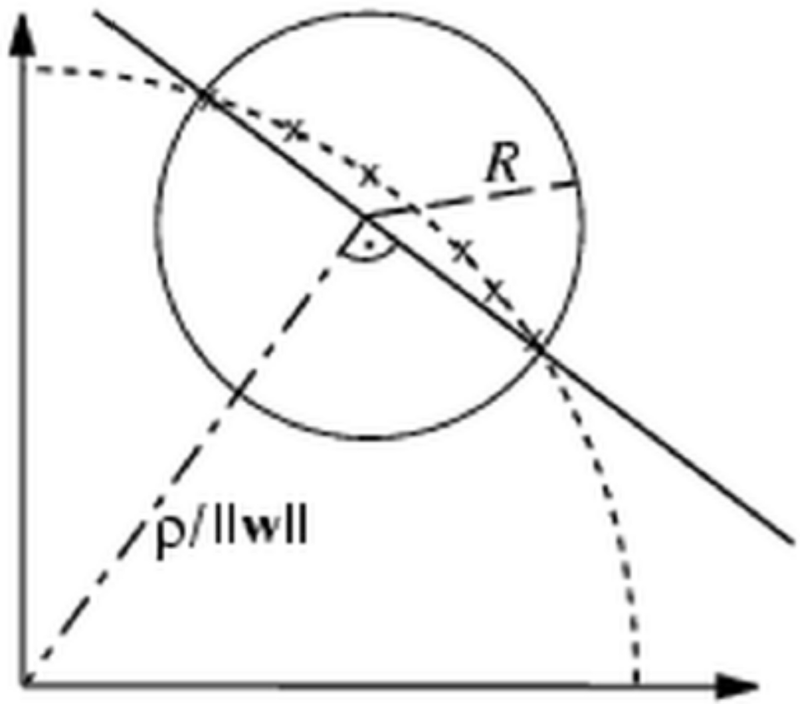
$$\rho = w^T \Phi(x_i) = \sum_{j=1}^{N} \beta_j k(x_i, x_j) \tag{6}$$

Relationship to sphere fitting

- Note, for a kernel $k(x, y)$ whose value depends only on the distance between the points $x - y$

$$k(x_i, x_i) = \; constant \; \forall i \qquad (7)$$

- Hence, all points lie on a hypersphere
- Finding the smallest hypersphere is equivalent to maximising the marging between data and the origin



Relationship with the Parzen estimator

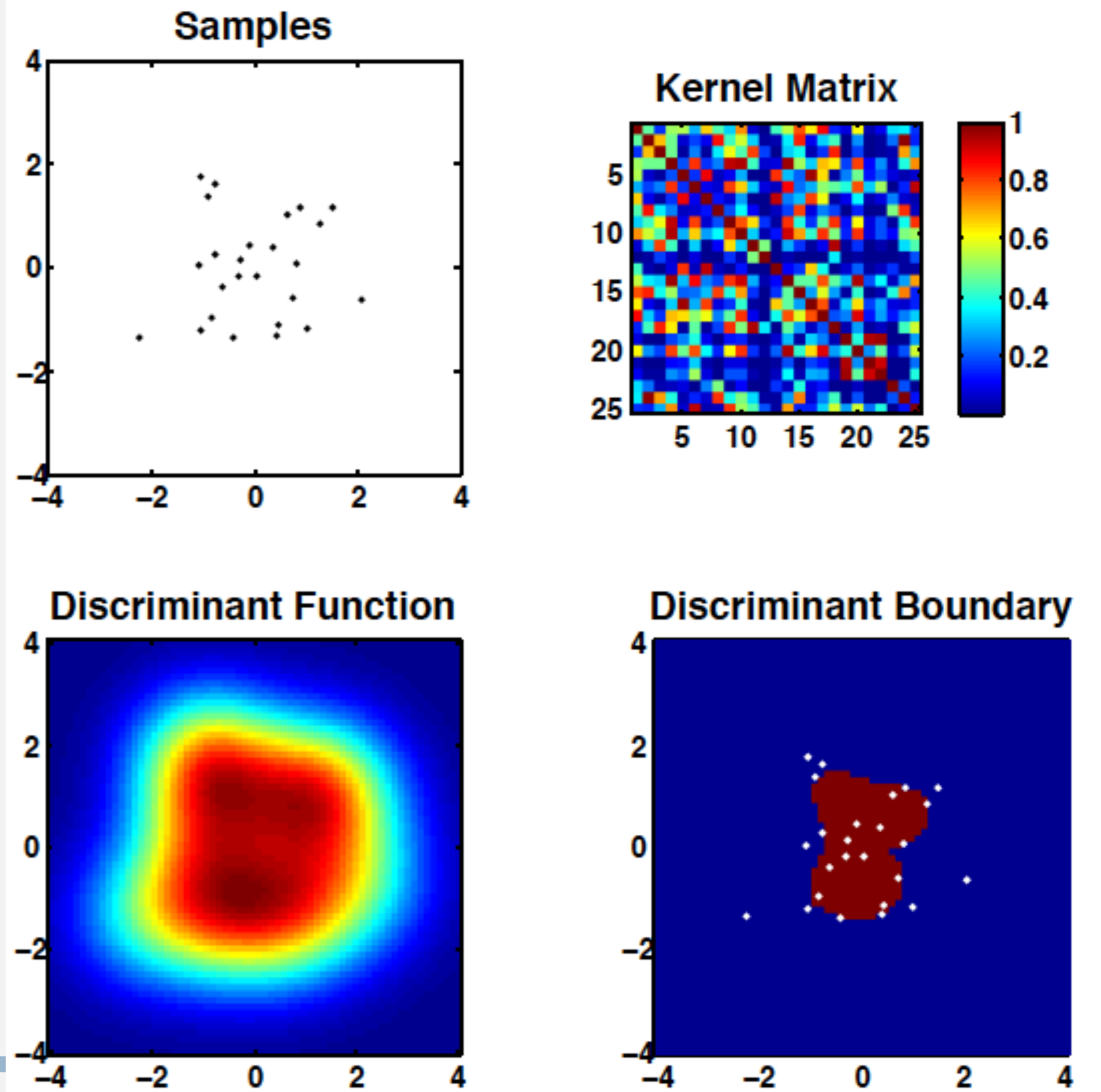
- When $\nu = 1$
  - The constraints become

$$0 \leq \beta_i \leq \frac{1}{\nu N}$$
$$\sum_{i=1}^{N} \beta_i$$

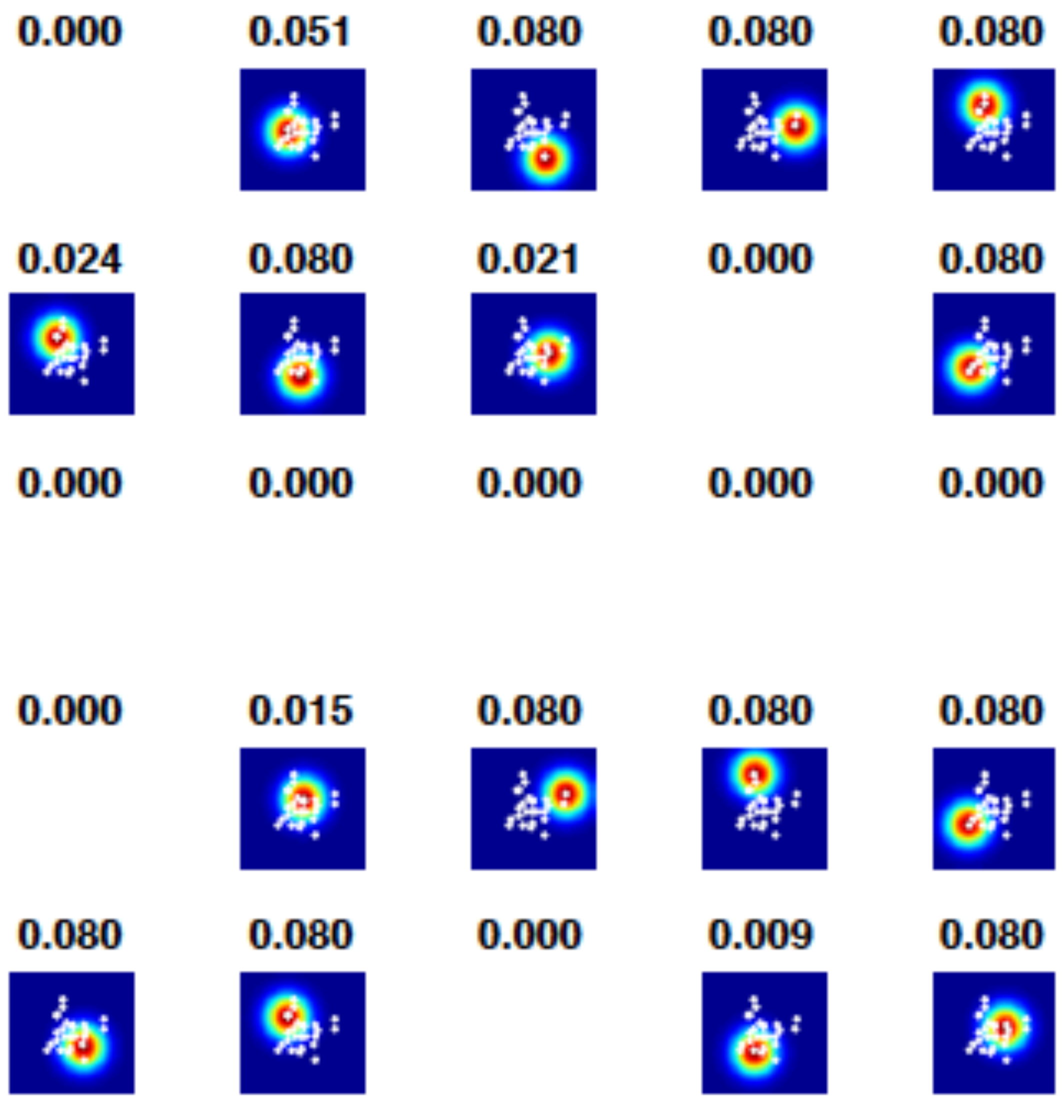
  - and they imply $\beta_i = \frac{1}{N}, \ \forall i$
  - The expansion $\sum_{i=1}^{N} \beta_i k(x_i, x)$ is a Parzen estimator

Example

- 25 samples from a Gaussian
- Parameter v=0.5



Example of coefficients



Example: A linear case

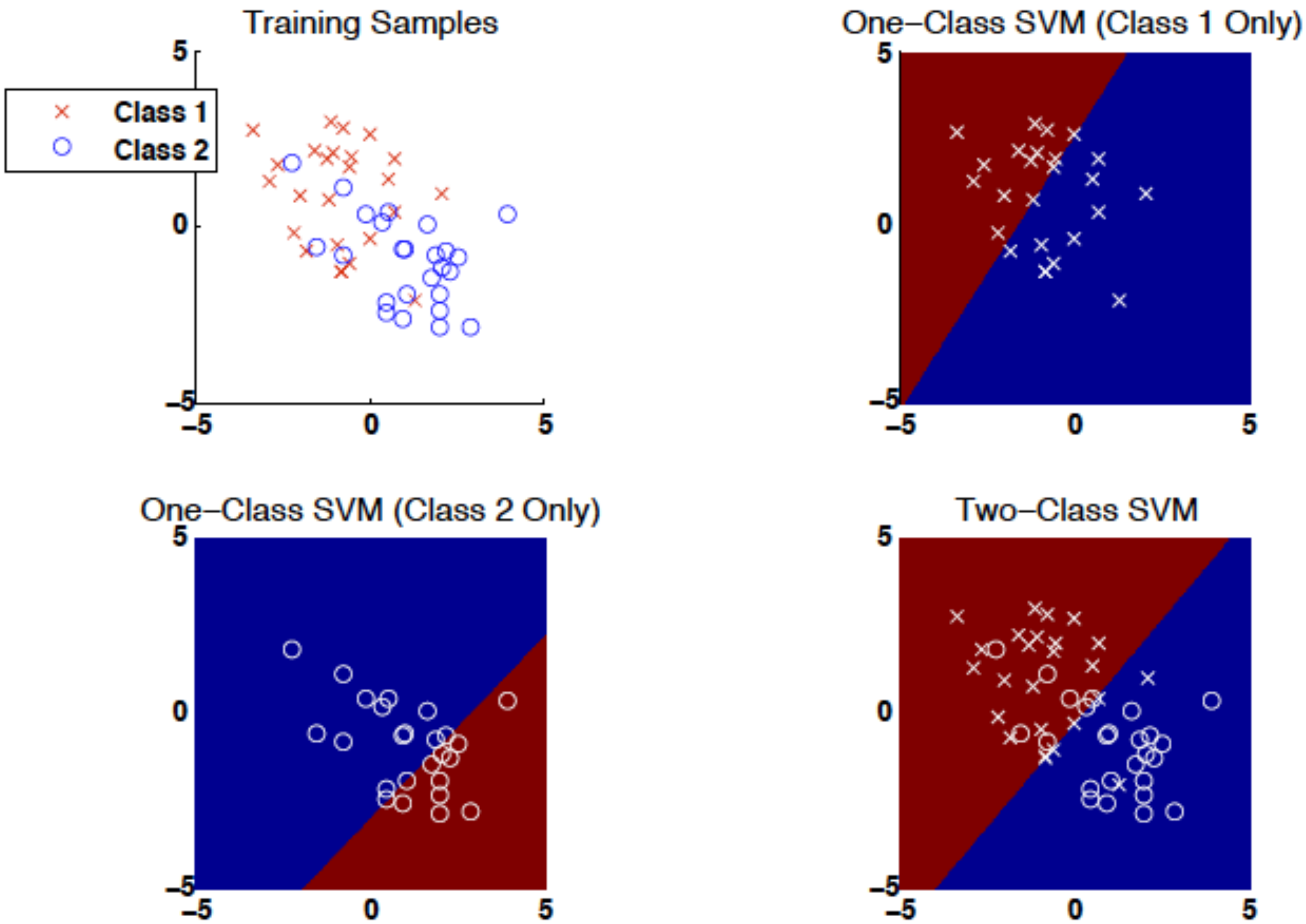
- Consider the case when $\Phi(x) = x$

- Example: Two Gaussians

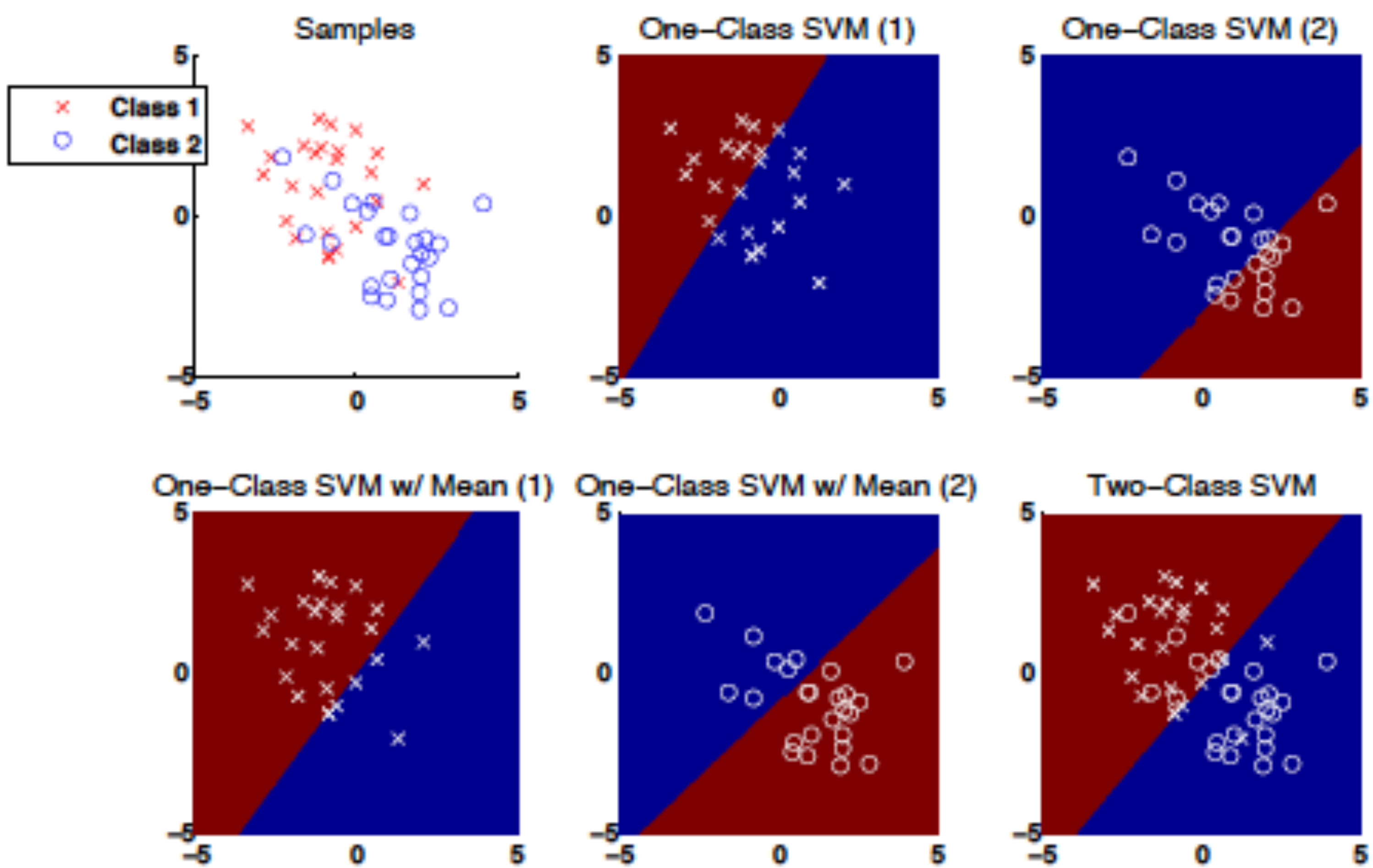
- One class SVM result in a greater Type 1 errors (leakage) and smaller Type 2 errors (false alarms).
- Instead of separating the data from the origin, we may formulate the problem so as to separate the data from the centroid as:

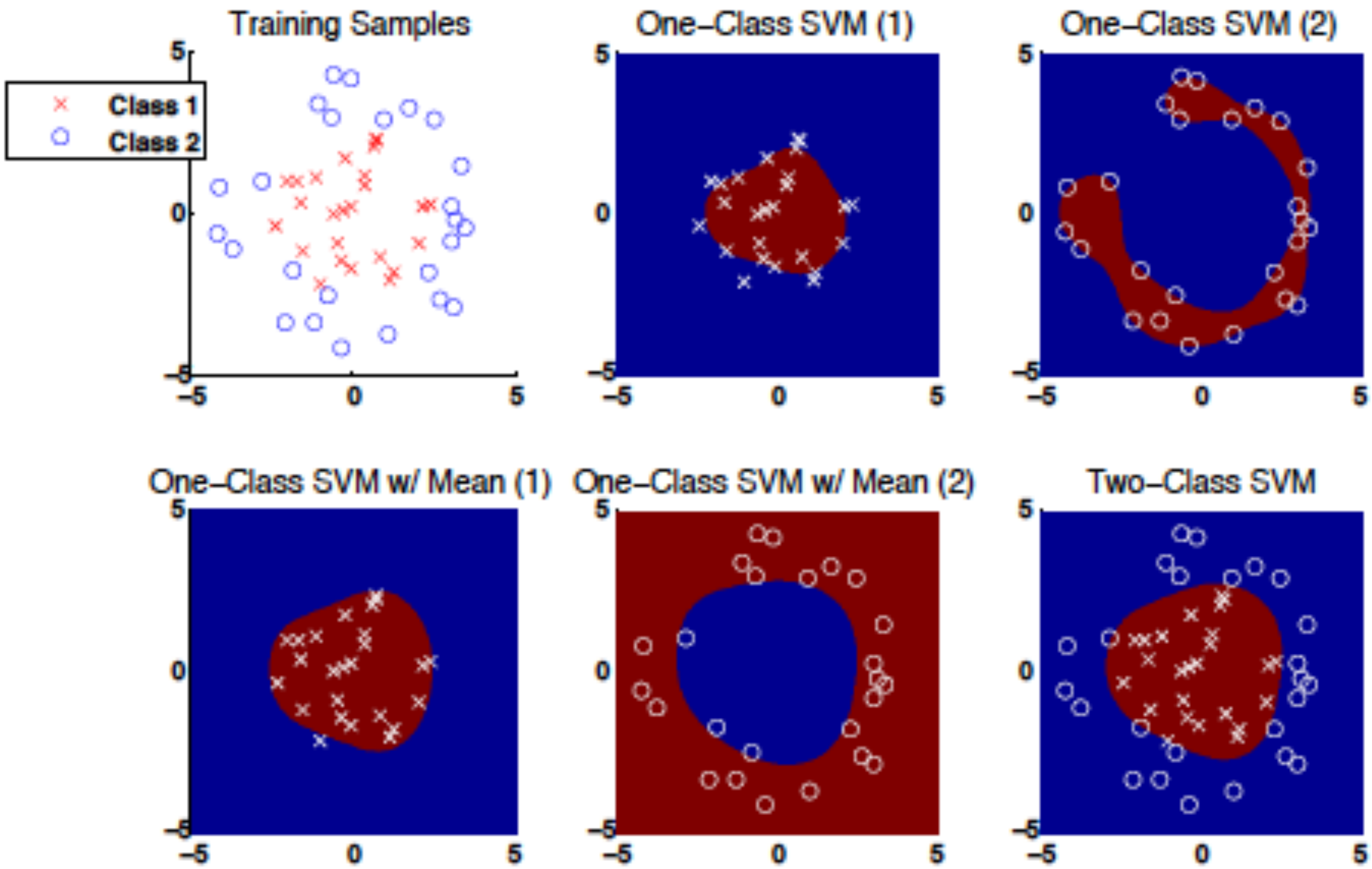
$$f_{w,\rho}(x) = sgn\{w^T[\Phi(x) - \frac{1}{N}\sum_{j=1}^{N}\Phi(x_j)] - \rho\} \quad (9)$$

1 class SVM vs 2 class SVM



Example: A linear case





Training Samples
Class 1
Class 2
One–Class SVM (1)
One–Class SVM (2)
One–Class SVM w/ Mean (1)
One–Class SVM w/ Mean (2)
Two–Class SVM
5
0
–5

Comments

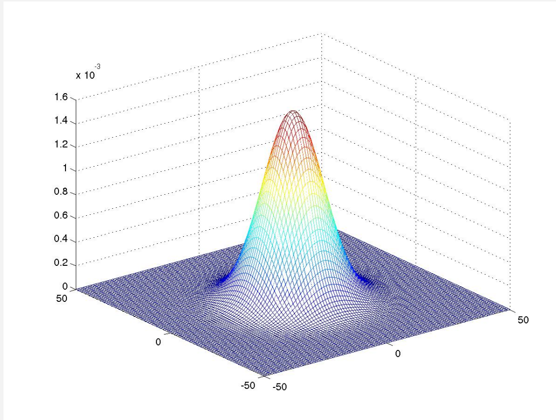
- Selection of meta parameters
  - Kernel bandwidth
  - Parameters
- Applicable to high dimensional problems
- Nonlinear boundary facilitated by the kernel trick

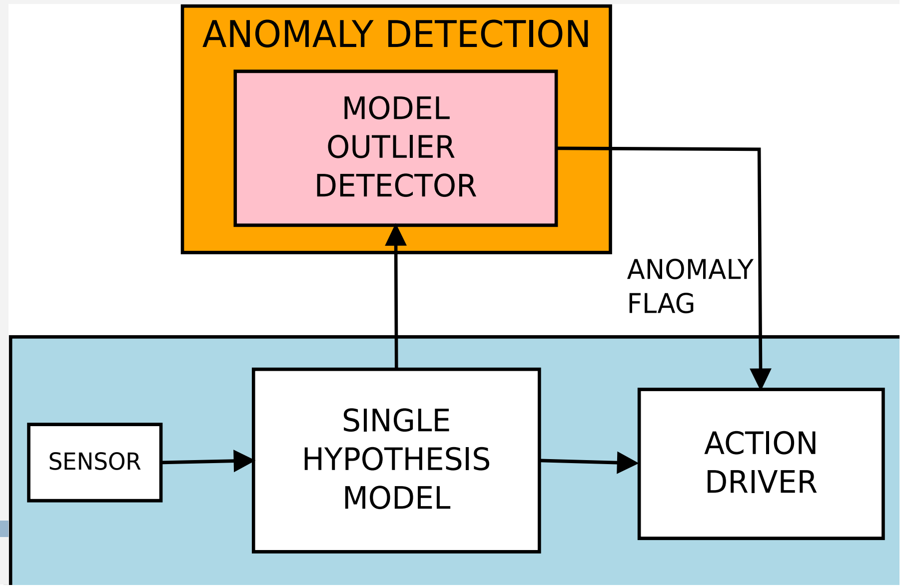
One class SVM summary

- Quantile estimation formulated as kernel machine learning
- High probability regions are estimated subject to regularisation
- One class SVM solution compared with two class SVM

Classical model and its critique

- Multiple models
- Discriminative classifiers
- Ambiguity of interpretation
- Contextual reasoning
- Hierarchical representation
- Data quality
- Model pruning





Different aspects of anomaly

Different aspects of anomaly

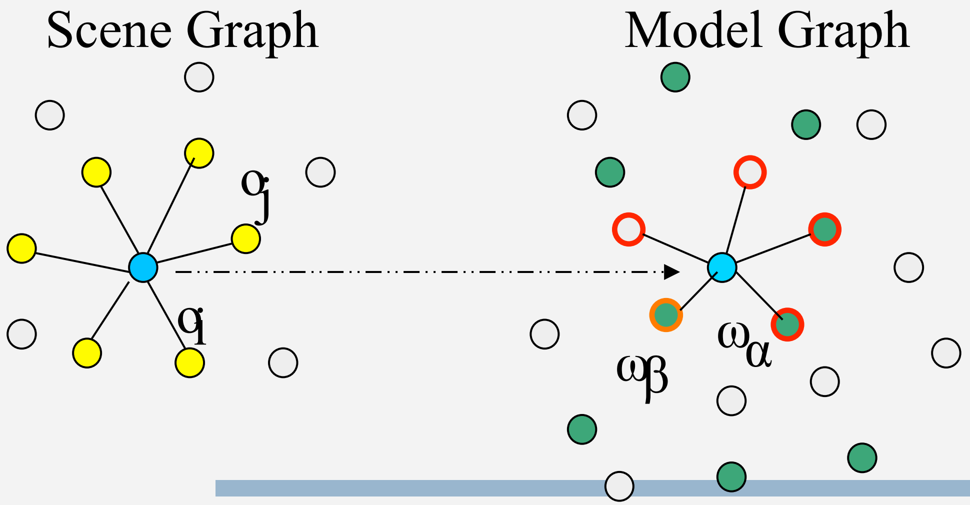
-Distribution drift
-Novelty detection

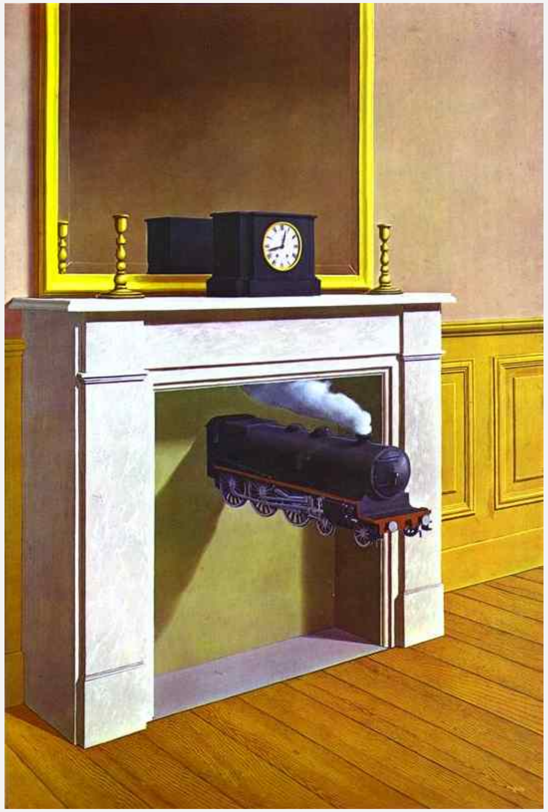
Data quality

Data quality

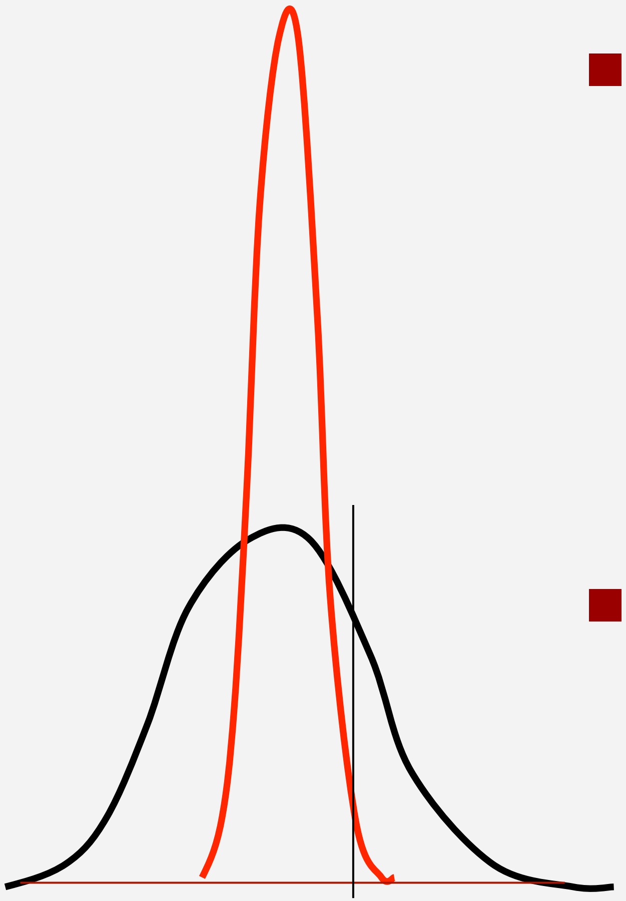
Incongruence/unexpected event

- Magritte's La duree poignard
- Model base pruning
  - Computational efficiency
- Hierarchical representation





Data quality/ decision confidence

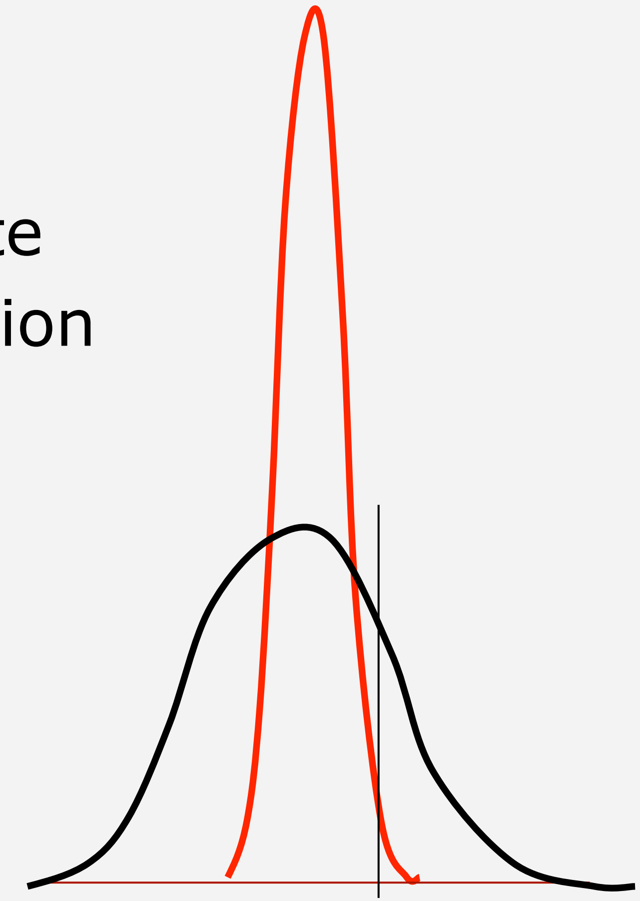


- Data quality
  - effect of noise on the notion of normality
  - need to measure data quality
  - notion of data quality and its dependence on context
- Confidence in classifier output

$$\Delta_c(x) = \frac{P(\omega_i|x) - e_i}{1 - 2e_i}$$

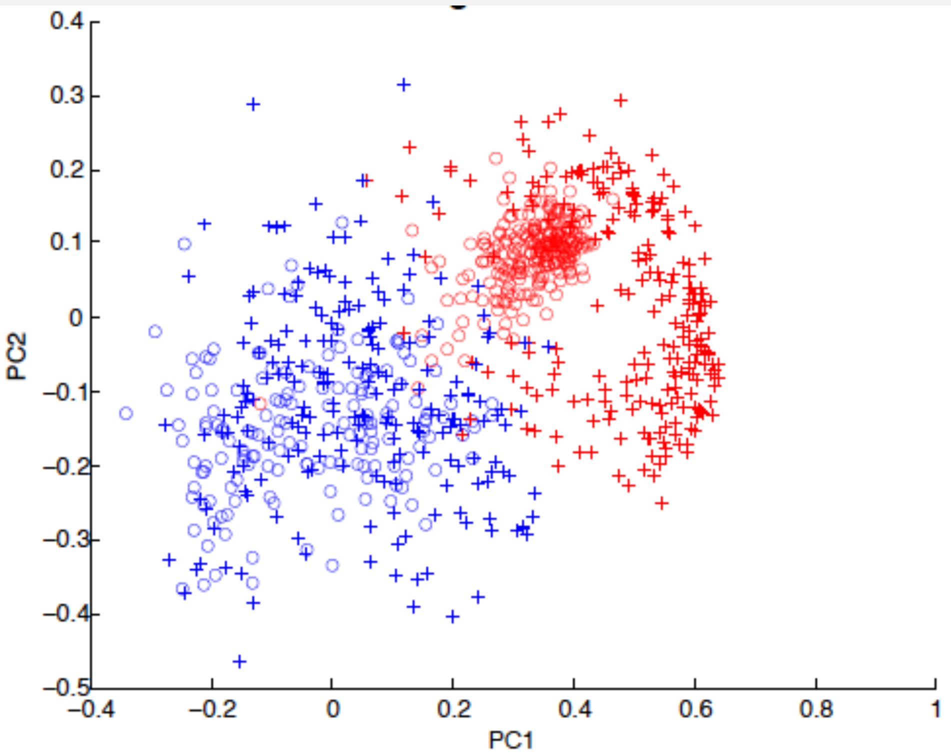
Challenges

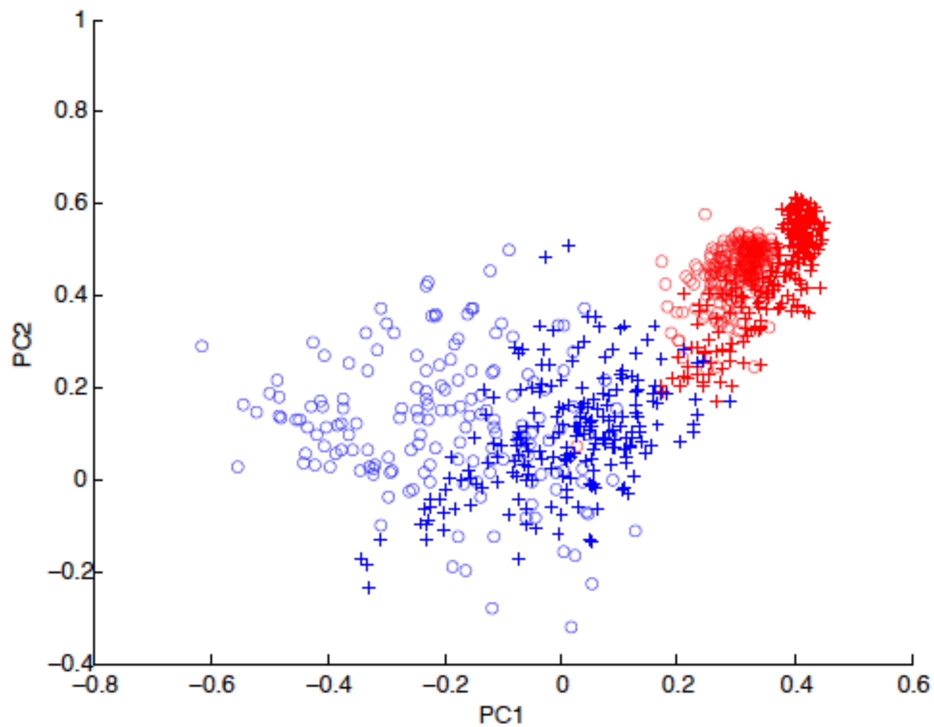
- Challenges of a more comprehensive approach
  - Meaning of data quality
    - Quality is relative, not absolute
    - Different levels of representation
  - Data quality measures
    - Multiple aspects of quality
    - Measures of quality
    - Overall quality/fusion



Distribution drift

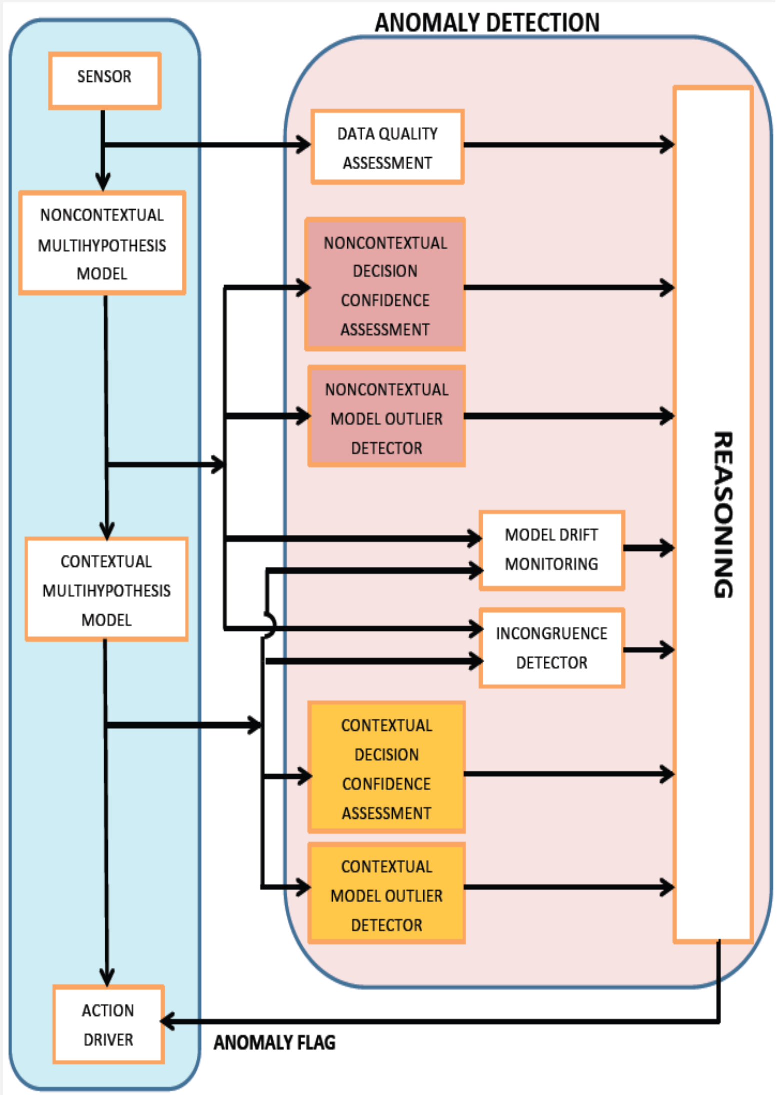
- Transfer of learning





Anomaly detection system architecture

- Classical model deficient
  - Outlier detection not enough
- Other mechanisms required
  - Data quality detection
  - Incongruence detection
  - Decision confidence estimation
  - Drift detection
  - As well as outlier detection
  - Reasoning (fusion)



Nuances of anomaly

- No anomaly
- Noisy measurement
- Unknown object
- Corrupted measurement
- Congruent labelling
- Unknown structure
- Spurious measurement errors
- Unexpected structural component
- Unexpected structural component & structure
- Measurement model drift

Context of anomaly detection

- Designing an operational system with anomaly detection capability
  - Data collection
  - System architecture
  - Representation
  - Machine learning
  - Context modelling
  - High level reasoning
  - Validation

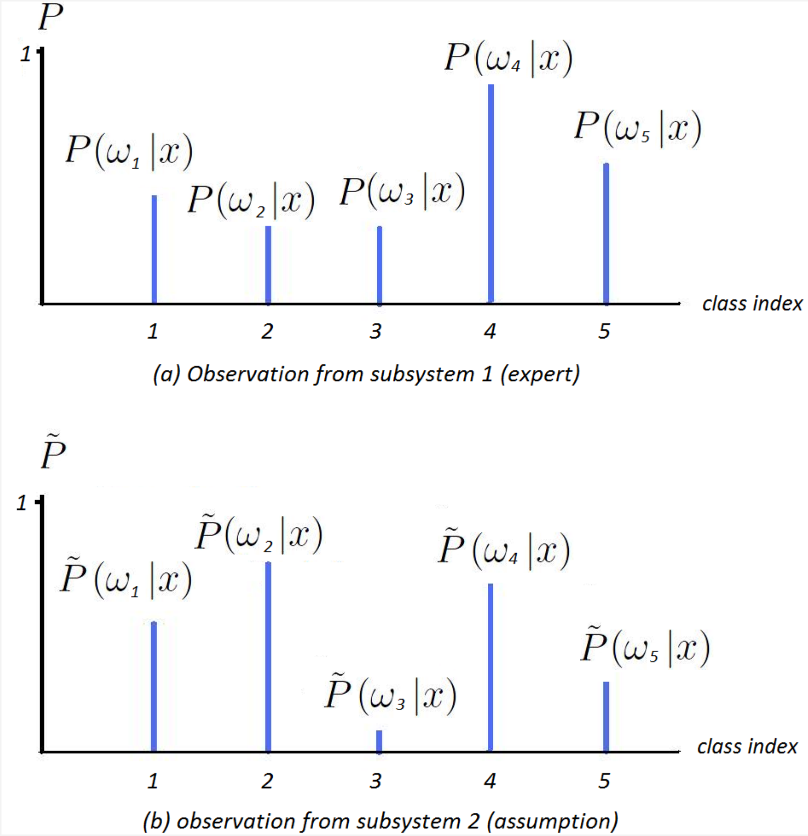
UNIVERSITY OF SURREY

Incongruence detection

- Detecting differences between observations and expectations (anomaly, rare event, incongruence)
- Basic principle – comparison of outputs of weak and strong classifiers (Ketabdar et al 2007)
- Dirac Project (Burget et al 2008, Weinshall et all [2009-2012])
- Exemplified by out-of-vocabulary word detection
  - Phoneme recognizer (weak classifier)
  - HMM speech recognizer (strong, contextual classifier)

Classifier incongruence

- Testing for incongruence
  - need an incongruence measure
  - understand its properties
  - sensitivity to noise



$P(\omega_j|x)$ classifier 1 output

$\tilde{P}(\omega_j|x)$ classifier 2 output

Incongruence measures

- Kulback-Leibler divergence: measures mutual information between the two distributions

$$\Delta_{BS} = \sum_{j=1}^{r} \tilde{P}(\omega_j|x) \log \frac{\tilde{P}(\omega_j|x)}{P(\omega_j|x)}$$

  - Known as Bayesian surprise
- Chi-square measure

$$\psi^2 = \sum_{i=1}^{m} \frac{[P(\omega_i|x) - \tilde{P}(\omega_i|\mathbf{x})]^2}{P(\omega_i|\mathbf{x}) + \tilde{P}(\omega_i|\mathbf{x})}$$

  - Assumption: estimation errors Gaussian
  - Variance proportional to the sum of probabilities

Properties of Chi-square

- Errors for non dominant classes are magnified (scaled by small variance)
- Joint zero entries are ignored
- Even when the probabilities of the dominant hypotheses agree, the sum over all the other hypotheses could be high
- The test statistics based on the assumption that the sampling distribution of errors is a product of Gaussian with zero mean and different variance for each class posterior

Bhattacharyya distance

- Bhattacharyya (geometric) distance

$$T_B = \sqrt{\sum_{i=1}^{m} P(\omega_i|\mathbf{x}) \times \tilde{P}(\omega_i|\mathbf{x})}$$

- Properties:
  - Distance different for different distributions, even if the two classifier outputs are identical for all hypotheses
  - Works as a matched filter
  - Measure can be affected by disagreements in the probabilities of minor hypotheses
  - Using as a reference the classifier output with the lowest entropy, the measure would yield much higher value than the posterior distribution with the highest entropy

Bhattacharyya distance

- Properties (cont)

- If the class probabilities are uniformly distributed, max value of the matching distribution is $\frac{1}{m}$. For observed zero-one distribution the surprise measure will have the same output value as for an optimal match. On the other hand for zero-one distribution as a reference, the maximum possible value is 1. An observed uniform distribution would yield surprise measure equal to $\frac{1}{m}$.
- Effect of errors can be gauged from $\sqrt{\sum_i [P(\omega_i|\mathbf{x}) + \eta_{\omega_i}][\tilde{P}(\omega_i|\mathbf{x}) + \tilde{\eta}_{\omega_i}]}$. It looks robust but because of non negativity constraints, etc. there will be some bias.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is defined as follows. Let the cumulative probability values c_i and $\tilde{c}_i$ denote

$$c_i = \sum_{k=1}^{i} P(\omega_k|\mathbf{x}) \tag{4}$$

and similarly

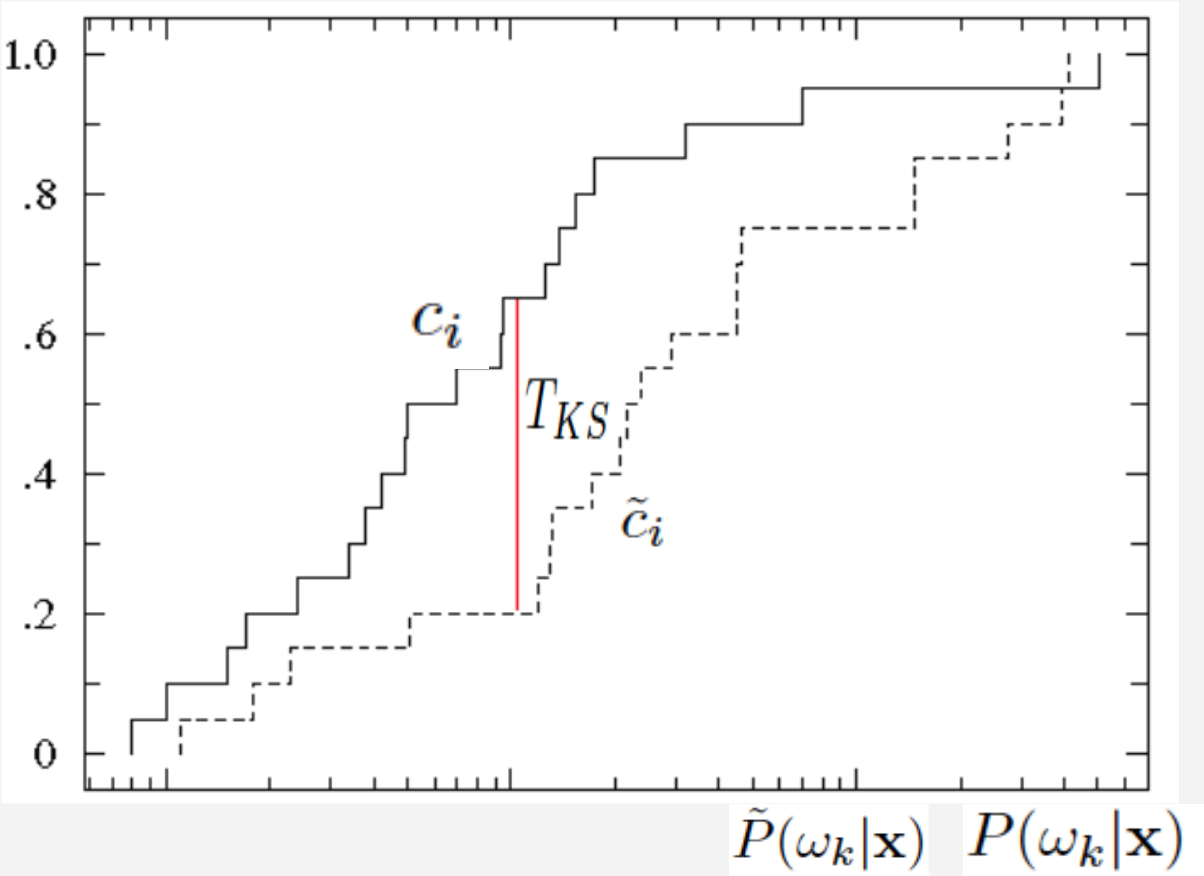
$$\tilde{c}_i = \sum_{k=1}^{i} \tilde{P}(\omega_k|\mathbf{x}) \tag{5}$$

Then Kolmogorov-Smirnov test of incongruence can be defined as

$$T_{KS} = max_i|c_i - \tilde{c}_i| \tag{6}$$

Properties of K-S test

- Resilience to estimation noise



Cramer von Mises measure

- Defined as

$$T_{CM} = \frac{1}{2}\sum_{i=1}^{m}[P(\omega_i|\mathbf{x}) + \tilde{P}(\omega_i|\mathbf{x})](c_i - \tilde{c}_i)^2$$

- Measures cumulative sum differences weighted by sum of probabilities (variance)
- All terms contribute, not only the max term
- This may impact on error robustness

Bayesian surprise measure

- Properties
- It goes to infinity for any hypothesis ω for which $P(\omega|x) \to 0$ while $\tilde{P}(\omega|x) \neq 0$. This can occur even for insignificant hypotheses and result in producing false alarms of incongruence.
- Not symmetric
- Divergence difficult to calibrate
- Classifier decision agnostic

Delta measure

- Defined as

$$\Delta_{avg} = \frac{1}{4}\{|P(\mu|x) - \tilde{P}(\mu|x)| + \delta(\mu,\tilde{\mu})|\tilde{P}(\tilde{\mu}|x) - \tilde{P}(\mu|x)| + |\tilde{P}(\tilde{\mu}|x) - P(\tilde{\mu}|x)| + \delta(\mu,\tilde{\mu})|P(\mu|x) - P(\tilde{\mu}|x)|\}$$

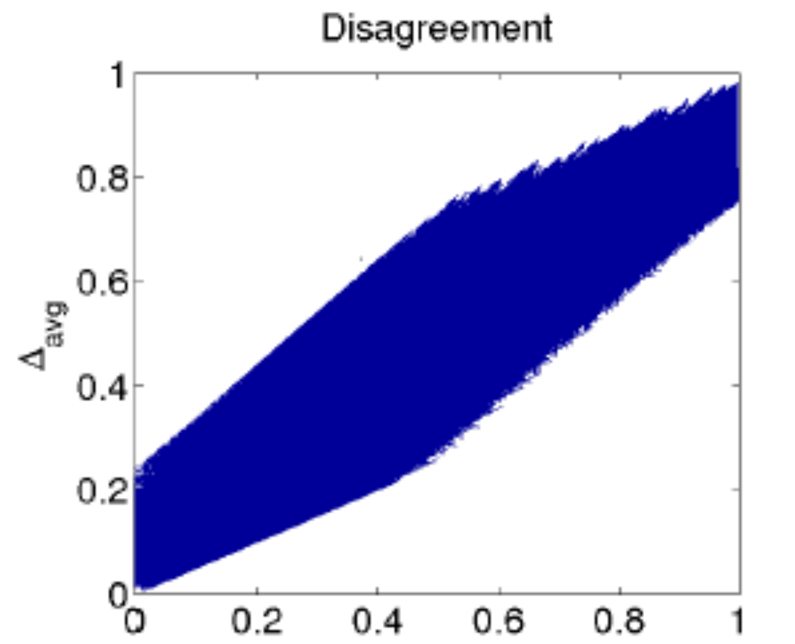
where $\delta(\mu,\tilde{\mu})$ function is defined as

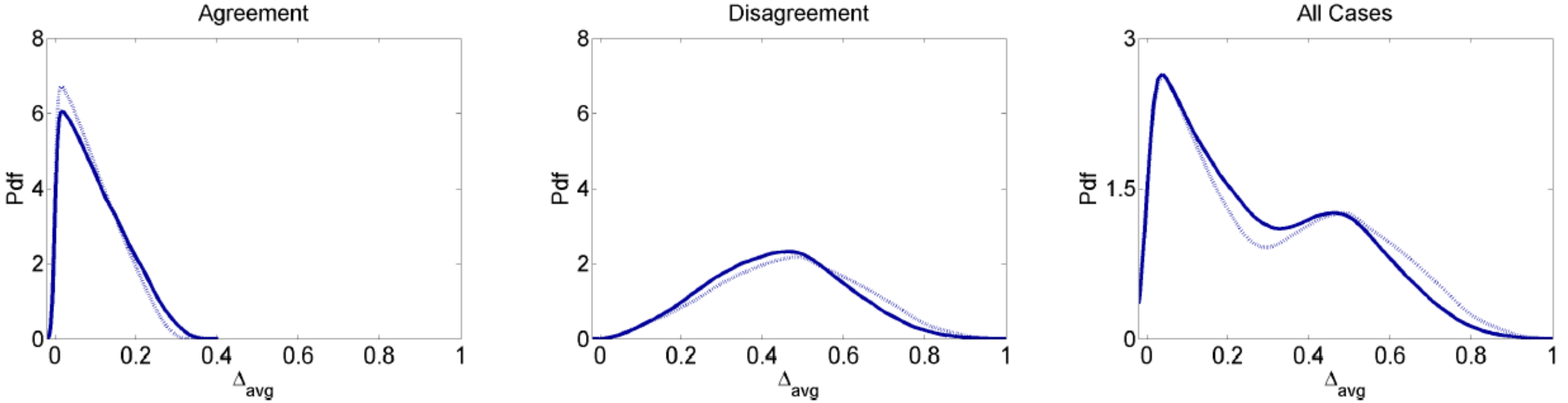
$$\delta(\mu,\tilde{\mu}) = \begin{cases} 0 & if \quad \mu = \tilde{\mu} \\ 1 & if \quad \mu \neq \tilde{\mu} \end{cases}$$

- Dominant hypotheses taken into account, non dominant ignored

Distribution of Δ_{avg}

- Distribution of noise free Δ_{avg} values





Estimation errors

- Class probabilities corrupted by noise

$$\hat{P}(\omega|\mathbf{x}) = P(\omega|\mathbf{x}) + \eta_\omega(\mathbf{x})$$

- satisfying

$$\sum_{i}^{m} \eta_\omega(\mathbf{x}) = 0$$

$$0 \le \eta_\omega(\mathbf{x}) + P(\omega|\mathbf{x}) \le 1$$

Error sensitivity

- Probabilities estimates affected by errors

$$P(\omega|\mathbf{x}) + \eta_\omega(\mathbf{x})$$

$$\tilde{P}(\omega|\mathbf{x}) + \tilde{\eta}_\omega(\mathbf{x})$$

- Constraints

$$\sum_i^m \eta_\omega(\mathbf{x}) = 0$$

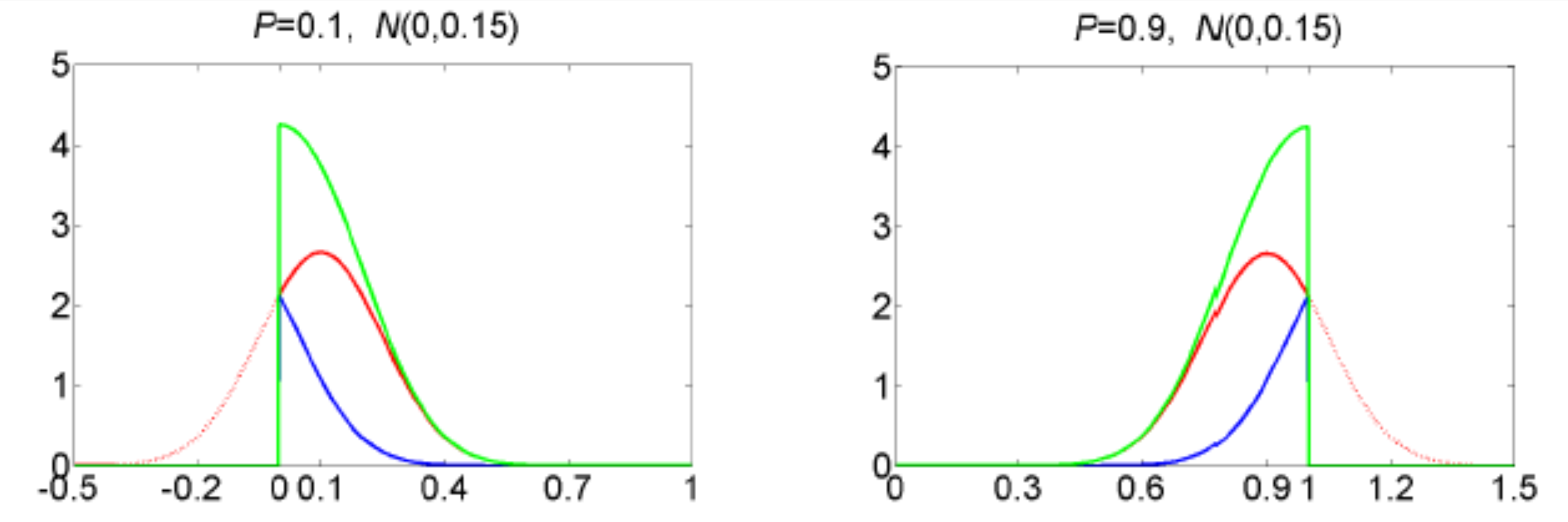
$$0 \leq \eta_\omega(\mathbf{x}) + P(\omega|\mathbf{x}) \leq 1$$

Estimation error distribution

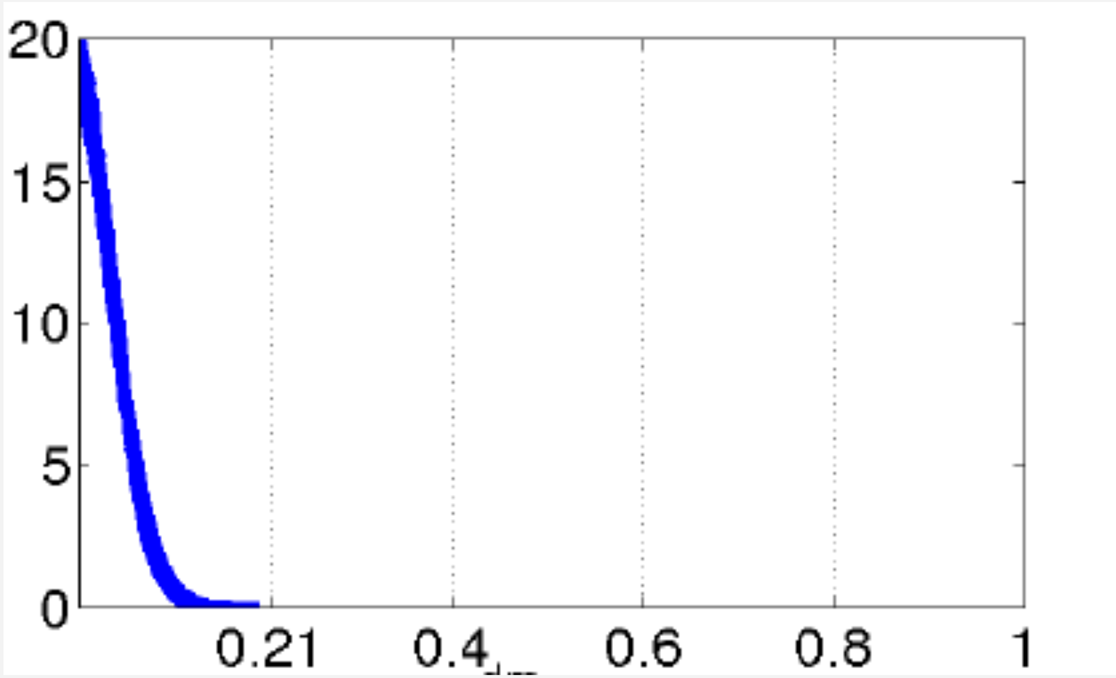
- Gaussian $q(\eta) = N(0, \sigma)$

with folded tails

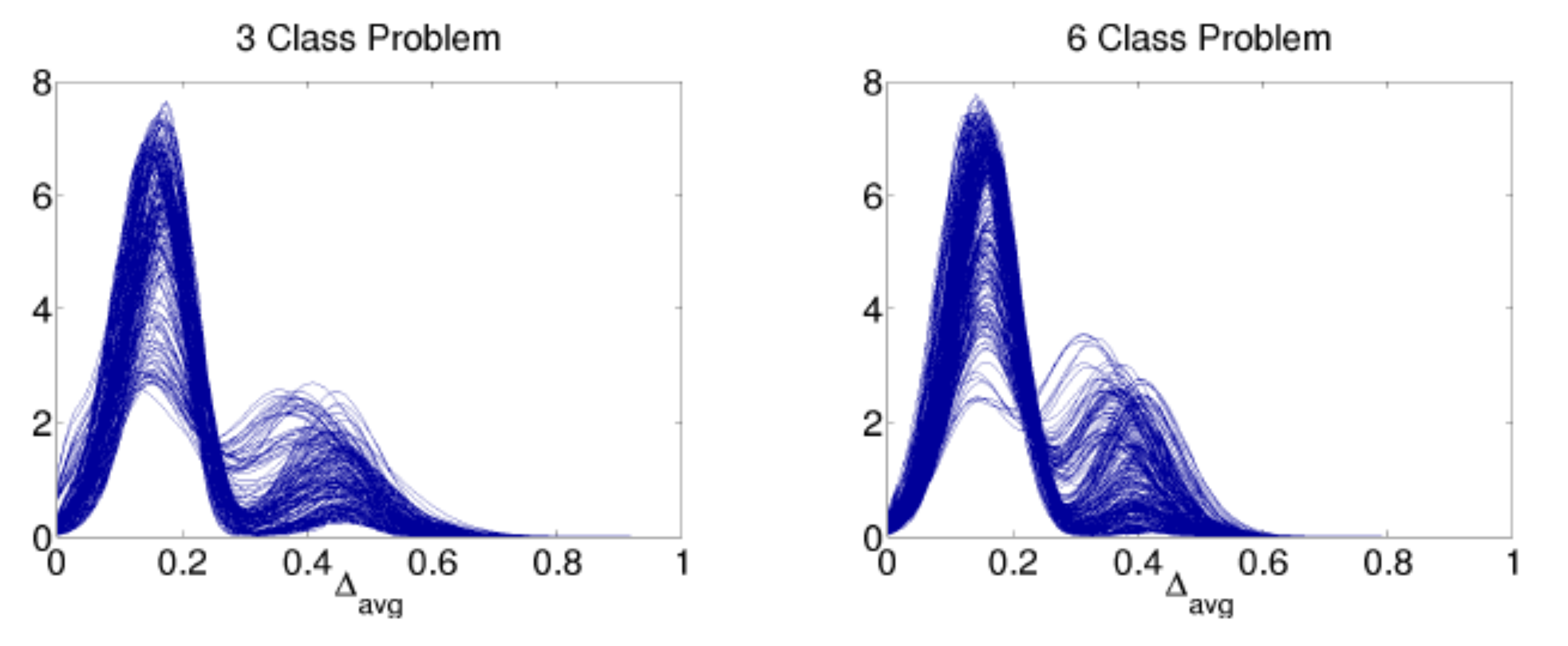
$$P \leq 0.5$$
$$p(\eta) = \begin{cases} 0 & \eta < -P \\ p(\eta) + p(-\eta - 2P) & \eta \geq -P \end{cases}$$
$$P > 0.5$$
$$p(\eta) = \begin{cases} 0 & \eta > 1 - P \\ p(\eta) + p(2 - 2P - \eta) & \eta \leq 1 - P \end{cases}$$



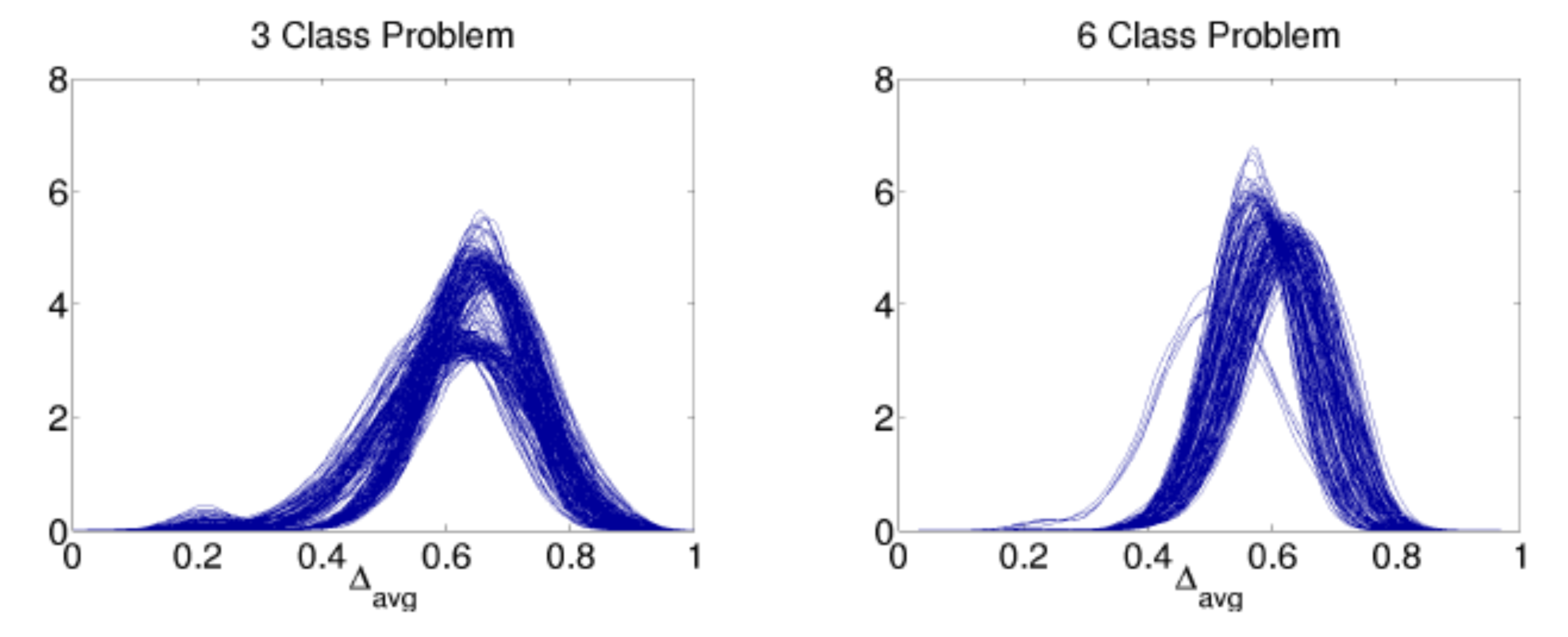
- Pdf curves of Δ_{avg} for classifier output similarity with estimation error noise N(0,0.1)



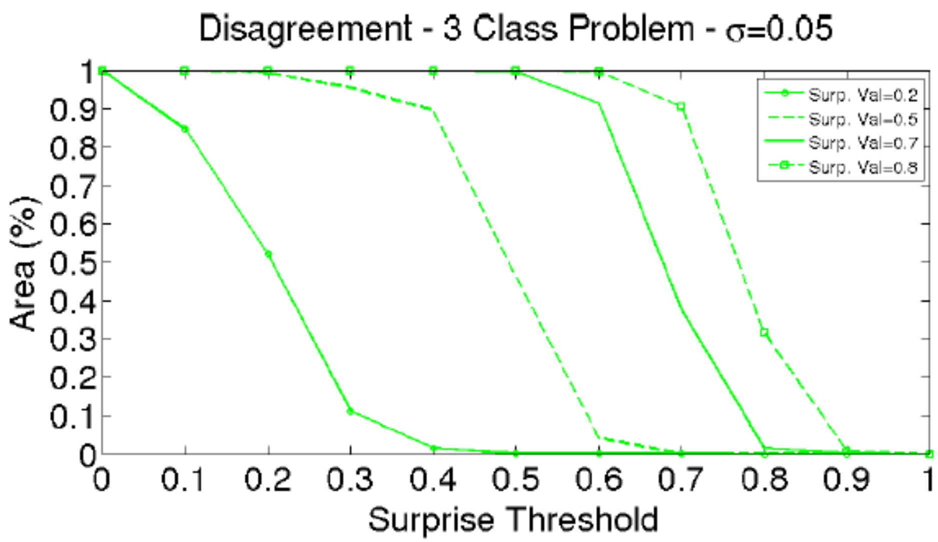
- Label agreement



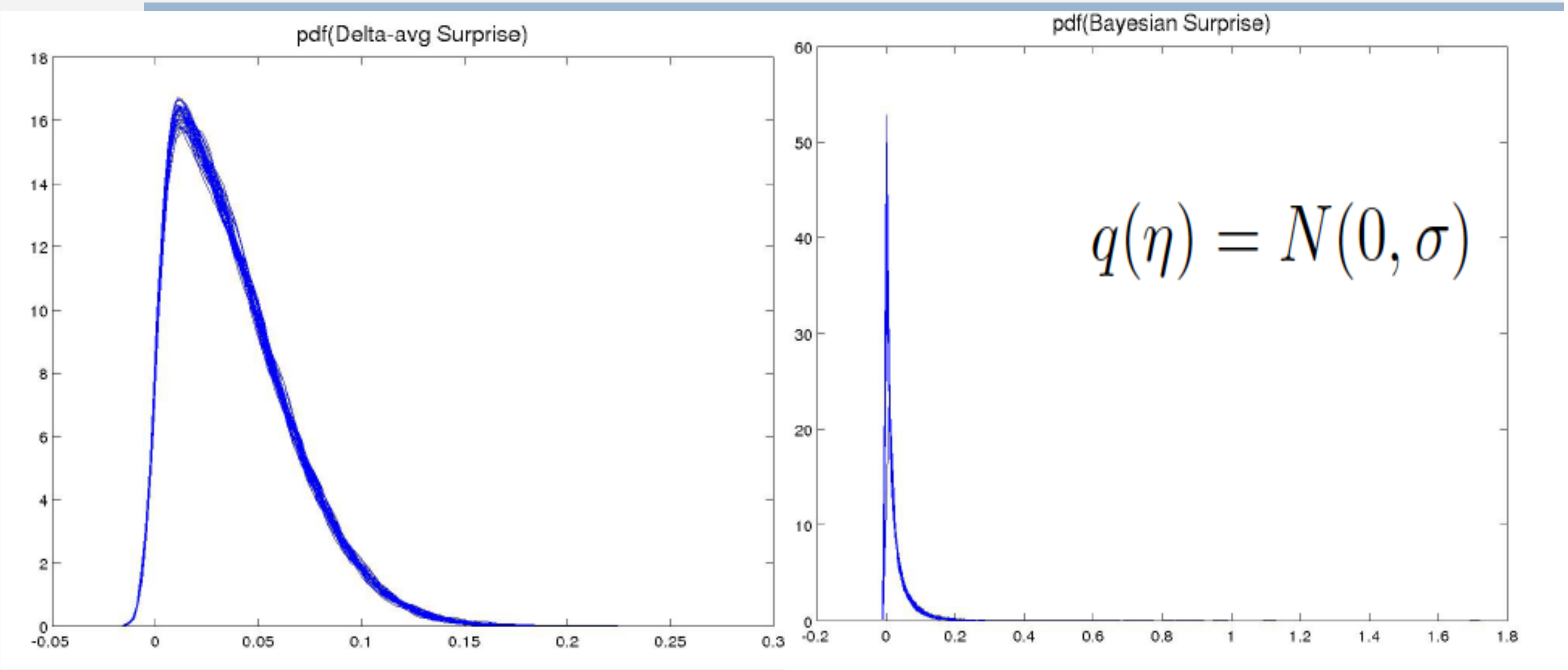
- Label disageement



- Results of simulation studies to determine decision threshold



Error sensitivity of incongruence measures



Scenario

- Identical class probabilities
- Estimation error st.dev 0.05

Thresholding

- One of the classifier incongruence measures can be used as a test statistics to detect incongruence
- An error sensitivity analysis would need to be carried for the chosen measure to estimate the test statistics distribution
- An appropriate decision threshold could then be determined to achieve a specified level of significance