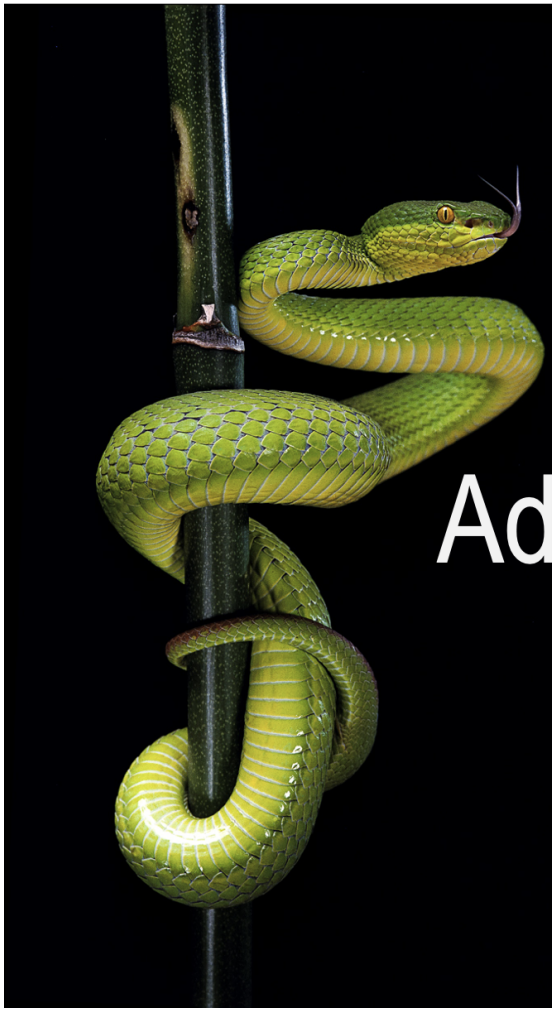


David J. Miller
Zhen Xiang
George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Chapter 03
Basics of Detection and Mixture Models

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Outline

- Mixture Densities
- Estimating the Parameters
  - Maximum Likelihood Estimation (MLE)
  - Expectation-Maximization algorithm (EM)
- K-means Clustering as a Special Case
- Model-Order Selection (BIC)
- Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)
- Some Detection Basics
- Performance Measures for Detection (ROC AUC)

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Preliminaries

- Given a dataset of "feature vectors,"

$$\mathcal{X} = \{\underline{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})' \ : \ i \in \{1, 2, \ldots, T\}\},$$

  each an independent realization of a random vector $\underline{X}$ with probability density function (pdf) $f_{\underline{X}}(\underline{x})$, $\underline{x} \in \mathbb{R}^N$.
- That is, for (Borel) subset $B \subset \mathbb{R}^N$,

$$\mathrm{P}(\underline{X} \in B) = \int_B f_{\underline{X}}(\underline{x}) \mathrm{d}\underline{x}.$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Example: The Multivariate Gaussian Density

- The multivariate (N-dim.) Gaussian density is

$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{C}|}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})' \mathbf{C}^{-1} (\underline{x} - \underline{\mu})\right), \text{ where}$$

$$\underline{\mu} = \mathrm{E}\underline{X} \in \mathbb{R}^N$$

is the mean vector and $|\mathbf{C}|$ is the determinant of the $N \times N$ positive-definite covariance matrix,

$$\mathbf{C} = \mathrm{E}[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'].$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

The Multivariate Gaussian Density (cont)

- Note that the Gaussian density is determined by just the first-order and second-order statistics (parameters),
- If $\mathbf{C}$ is non-singular, the *Mahalanobis* distance between $\underline{x} \in \mathbb{R}^N$ and this multivariate Gaussian distribution is defined to be

$$\|\underline{x} - \underline{\mu}\|_{\text{mn}} := \sqrt{(\underline{x} - \underline{\mu})' \mathbf{C}^{-1} (\underline{x} - \underline{\mu})}.$$

In one dimension ($N = 1$), $\|\underline{x} - \underline{\mu}\|_{\text{mn}}$ is just the number of standard deviations ($\sigma = \sqrt{\mathbf{C}}$) between $\underline{x}$ and $\underline{\mu}$, *i.e.*, $\|\underline{x} - \underline{\mu}\|_2 / \sigma$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Discrete Distributions

- If the random vector $\underline{X}$'s features are discrete valued, its joint probability mass function (pmf) is $p_{\underline{X}}(\underline{x}) = \mathrm{P}(\underline{X} = \underline{x})$.
- The pmf satisfies $0 \leq p_{\underline{X}}(\underline{x}) \leq 1$ and

$$\sum_{\underline{x} \in R_{\underline{X}}} p_{\underline{X}}(\underline{x}) = 1,$$

  where $R_{\underline{X}} \subset \mathbb{R}^N$ is the countable strict-range of $\underline{X}$, *i.e.*, $p_{\underline{X}}(\underline{x}) > 0 \Leftrightarrow \underline{x} \in R_{\underline{X}}$.
- Using Dirac impulses δ in $\mathbb{R}^N$, one can express a pmf as a pdf,

$$f_{\underline{X}}(\underline{x}) = \sum_{\underline{z} \in R_{\underline{X}}} \delta(\underline{x} - \underline{z}) p_{\underline{X}}(\underline{z}).$$

- E.g., the multinomial distribution,

$$p_{\underline{X}}(\underline{x}) = \frac{W!}{x_1! x_2! \cdots x_N!} p_1^{x_1} p_2^{x_2} \cdots p_d^{x_N}.$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Mixture Densities

- A mixture density has the form:

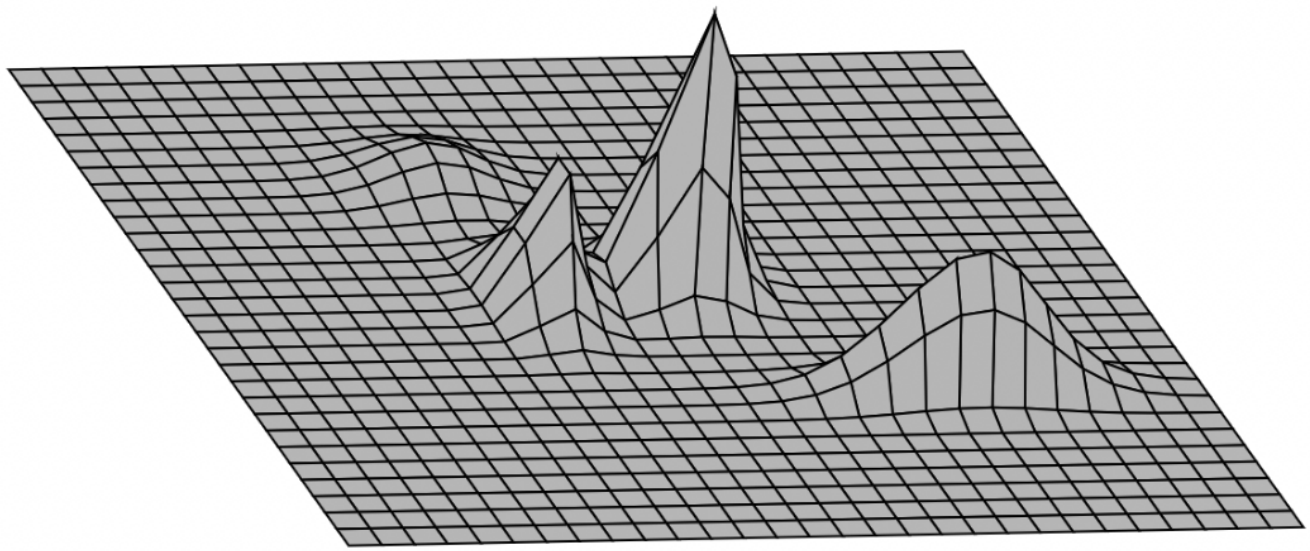
$$f_{\underline{X}}(\underline{x}) = \sum_{k=1}^{M} \alpha_k f_{\underline{X}|k}(\underline{x}; \Theta_k).$$

where:

- $f_{\underline{X}|k}(\cdot; \Theta_k)$ is a valid density function (often referred to as a *component* density) specified by parameters Θ_k,
- α_k is the *prior* probability that a feature vector is generated according to the k^{th} component, *i.e.*, $\{\alpha_k\}_{k=1}^{M}$ is a pmf, and
- the mixture model's parameters are $\Theta = \{\Theta_k, \alpha_k \mid k = 1, \ldots, M\}$.
- The components need not be all Gaussian densities.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Example GMM with $M = 4$ 2D Gaussian components



(Plotted using CalcPlot3D)

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Component inference

- The *a posteriori* probability that a data sample $\underline{x}$ was generated by each of the components:

$$p(k|\underline{x}) = \frac{\alpha_k f_{\underline{X}|k}(\underline{x}; \Theta_k)}{\sum_{j=1}^{M} \alpha_j f_{\underline{X}|j}(\underline{x}; \Theta_j)}, \quad k = 1, \ldots, M, \quad \text{where}$$

$$p(k|\underline{x}) := \mathrm{P}(Y(\underline{X}) = k | \underline{X} = \underline{x}), \quad \text{and}$$

- $Y(\underline{x})$ is the mixture component label for $\underline{x}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Component inference: GMM example

- E.g., for a GMM

$$p(k|\underline{x}) = \frac{\alpha_k |C_k|^{-1/2} e^{-(\underline{x}-\underline{\mu}_k)' \mathbf{C}_k^{-1} (\underline{x}-\underline{\mu}_k)/2}}{\sum_{j=1}^{M} \alpha_j |\mathbf{C}_j|^{-1/2} e^{-(\underline{x}-\underline{\mu}_j)' \mathbf{C}_j^{-1} (\underline{x}-\underline{\mu}_j)/2}}, \quad k = 1, \ldots, M.$$

- If a component's covariance matrix is diagonal, then the joint component density factors as a product of marginal Gaussian densities over the individual features (*i.e.*, the features are independent under the given component).
- This simplifies even further if the covariance matrix is a scaled identity matrix (in this case, the features all have the same variance).

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Estimating the Parameters

- Assuming the data $\mathcal{X}$ is T independent realizations $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_T$ of random vector $\underline{X}$, the maximum-likelihood (ML) parameters are:

$$\hat{\Theta}_{\mathrm{ML}} = \arg\max_{\Theta} \prod_{i=1}^{T} f_{\underline{X}}(\underline{x}_i; \Theta) = \arg\max_{\Theta} \sum_{i=1}^{T} \log f_{\underline{X}}(\underline{x}_i; \Theta).$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Estimating the Parameters: GMM example

- Though a non-convex optimization problem in general, a single globally optimal ML solution in closed form is available for a single-component GMM (a multivariate Gaussian):

$$\hat{\underline{\mu}} = \frac{1}{T}\sum_{i=1}^{T}\underline{x}_i, \hat{\mathbf{C}} = \frac{1}{T}\sum_{i=1}^{T}(\underline{x}_i - \hat{\underline{\mu}})(\underline{x}_i - \hat{\underline{\mu}})'.$$

- That is, $\hat{\mathbf{C}}$ is the average outer product of the centered data samples $\underline{x} \in \mathcal{X}$.
- Note that $\hat{\underline{\mu}}$ is an unbiased estimator of $\underline{\mu}$ ($\mathrm{E}\hat{\underline{\mu}} = \underline{\mu}$),
- but though $\hat{\mathbf{C}}$ is a biased estimator of $\mathbf{C}$ ($\mathrm{E}\hat{\mathbf{C}} = \mathbf{C}T/(T-1)$), it is consistent (asymptotically unbiased).

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

The EM algorithm for MM parameter estimation

- Expectation-Maximization (EM)
  - is guaranteed to converge to an extremum of the likelihood function;
  - performs a number of iterations, with each iteration guaranteed to increase the likelihood function;
  - unlike gradient ascent, EM does *not* require the (complicating) choice of a step size hyperparameter; and
  - for some complicated density functions, it converts an apparently intractable problem into a tractable one.
- First write the *incomplete* log-likelihood of the data $\mathcal{X}$ as:

$$\mathcal{L} = \sum_{i=1}^{T} \log\Big(\sum_{j=1}^{M} \alpha_j f_{\underline{X}|j}(\underline{x}_i; \Theta_j)\Big).$$

- Use binary (indicator) data $\{v_{ij} \mid i = 1, \ldots, T,\ j = 1, \ldots, M\}$:
  - $v_{ij} = 1$ if sample $\underline{x}_i$ was generated by component j, and
  - $v_{ij} = 0$ otherwise.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

EM: complete data log-likelihood

- Given the v data, the the *complete data log-likelihood function* is

$$\begin{aligned}\mathcal{L}_c &= \sum_{i=1}^{T} \log\left(\sum_{j=1}^{M} v_{ij}\alpha_j f_{\underline{X}|j}(\underline{x}_i;\Theta_j)\right) \\ &= \sum_{i=1}^{T}\sum_{j=1}^{M} v_{ij}\log(\alpha_j f_{\underline{X}|j}(\underline{x}_i;\Theta_j)).\end{aligned}$$

- Note that the v data simplifies things with the sum now outside of the log.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

EM: estimating the binary data v

- Treating the v data as random variables V, the *expected complete data log-likelihood* is

$$\mathrm{E}(\mathcal{L}_c|\mathcal{X};\Theta) = \sum_{i=1}^{T}\sum_{j=1}^{M} \mathrm{E}(V_{ij}|\mathcal{X};\Theta)\log(\alpha_j f_{\underline{X}|j}(\underline{x}_i;\Theta_j)),$$

where

$$\mathrm{E}(V_{ij}|\mathcal{X};\Theta) \;=\; p[j|i] \;=\; \frac{\alpha_j f_{\underline{X}|j}(\underline{x}_i;\Theta_j)}{\sum_{l=1}^{M}\alpha_l f_{\underline{X}|l}(\underline{x}_i;\Theta_l)}$$

and

$$p[j|i] \;:=\; p(j|\underline{x}_i) \;:=\; \mathrm{P}(Y(\underline{X}) = j|\underline{X} = \underline{x}_i).$$

- Thus, the expected complete data log-likelihood is:

$$\mathrm{E}(\mathcal{L}_c|\mathcal{X};\Theta) = \sum_{i=1}^{T}\sum_{j=1}^{M} p[j|i]\log(\alpha_j f_{\underline{X}|j}(\underline{x}_i;\Theta_j)).$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

EM: auxiliary function

- EM maximizes the *auxiliary function*

$$\mathcal{F} = \sum_{i=1}^{T}\sum_{j=1}^{M} p[j|i] \log(\alpha_j f_{\underline{X}|j}(\underline{x}_i; \Theta_j)) - \sum_{i=1}^{T}\sum_{j=1}^{M} p[j|i] \log p[j|i]$$

  over *both* the model parameters Θ and the *a posteriori* probabilities $\{p[j|i] \mid j = 1, ..., M, \ i = 1, ..., T\}$.

- Note that the second term of $\mathcal{F}$ is Shannon's entropy for Y given $\underline{X} = \underline{x}_i$:

$$H = -\sum_{j=1}^{M} p[j|i] \log p[j|i].$$

- $\mathcal{F}$ is optimized by alternating two optimization steps, the E-step and the M-step, until convergence.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

EM: E-step and M-step

- E-step maximizes the auxiliary function over the posteriors $\{p[j|i]\}$ given the parameters Θ held fixed. This yields the closed-form expression at iteration $t+1$:

$$p[j|i]^{(t+1)} = \frac{\alpha_j^{(t)} f_{\underline{X}|j}(\underline{x}_i; \Theta_j^{(t)})}{\sum_{l=1}^{M} \alpha_l^{(t)} f_{\underline{X}|l}(\underline{x}_i; \Theta_l^{(t)})},$$

  where parameters at t, $\Theta^{(t)} = \{\underline{\alpha}^{(t)}, \Theta_l^{(t)} \mid l = 1, 2, ..., M\}$, are plugged in to compute the *a posteriori* probabilities at $t+1$.

- Then the M-step maximizes F over Θ yielding $\Theta^{(t+1)}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

EM: GMM example

- For a GMM, the *closed form* M-step update is:

$$\underline{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{T} p[j|i]^{(t+1)} \underline{x}_i}{\sum_{i=1}^{T} p[j|i]^{(t+1)}}, \quad j = 1, \ldots, M,$$

$$\alpha_j^{(t+1)} = \frac{1}{T} \sum_{i=1}^{T} p[j|i]^{(t+1)}, \quad j = 1, \ldots, M,$$

followed by

$$\mathbf{C}_j^{(t+1)} = \frac{\sum_{i=1}^{T} p[j|i]^{(t+1)} (\underline{x}_i - \underline{\mu}_j^{(t+1)})(\underline{x}_i - \underline{\mu}_j^{(t+1)})'}{\sum_{i=1}^{T} p[j|i]^{(t+1)}}, \quad j = 1, \ldots, M$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

EM: general properties

- Each step (E or M above) is non-decreasing in the auxiliary function, $\mathcal{F}$ (which is the theoretical basis for EM convergence).
- Each M-step increases the *incomplete* data log-likelihood function (the original objective).
- Closed-form steps imply no step-size hyperparameter as needed for gradient based optimization.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Gaussian Kernel Density

- In a Gaussian *kernel* density, each data point $\underline{x} \in \mathcal{X} \subset \mathbb{R}^N$ is the center of an isotropic Gaussian density with common variance σ^2 and uncorrelated features: $\forall \underline{y} \in \mathbb{R}^N$,

$$f(\underline{y}) = \sum_{\underline{x} \in \mathcal{X}} \frac{1}{\sigma\sqrt{(2\pi)^N}} \exp(-\frac{1}{2\sigma^2}\|\underline{y} - \underline{x}\|^2)$$

- **Exercise:** Show how the common parameter σ^2 can be estimated to maximize the likelihood of the data $\mathcal{X}$ under f.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

K-Means Clustering as a Special Case of EM

- Suppose for a GMM that $\mathbf{C}_j = \sigma^2 I$ and $\alpha_j = \frac{1}{K}\ \forall j$.
- Also suppose that in the E-step one forces *hard* $\in \{0, 1\}$ assignments of data points to components (clusters),
  - $v_{ij}^{(t+1)} = 1$ for $j = \arg\max_k p[k|i]^{(t+1)}$ and
  - $v_{il}^{(t+1)} = 0 \ \ \forall l \neq j$,

  which reduces the E-step to the *nearest-neighbor data assignment rule* ($\forall \sigma^2$):

$$v_{ij}^{(t+1)} = 1 \ \ \text{if} \ \ \|\underline{x}_i - \underline{\mu}_j^{(t+1)}\| \leq \|\underline{x}_i - \underline{\mu}_k^{(t+1)}\| \quad \forall k \neq j,$$

  and the M-step (optimizing over the mean parameters) reduces to the *centroid rule*,

$$\underline{\mu}_j^{(t+1)} = \sum_{i=1}^{T} v_{ij}^{(t+1)} \underline{x}_i \Bigg/ \sum_{i=1}^{T} v_{ij}^{(t+1)} \ .$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

K-Means Clustering (cont)

- So this variant of EM gives a local min of K-Means clustering distortion:

$$\sum_{i=1}^{T}\sum_{j=1}^{K} v_{ij}\|\underline{x}_i - \underline{\mu}_j\|^2.$$

- Note that here "K" does not stand for the number of classes (indeed, unsupervised K-Means does not rely on class labels); rather, it is the number of clusters.
- It is ill-posed to seek to also maximize the data log-likelihood (DLL) over the hyperparameter K,
- *e.g.*, if a Gaussian kernel component is centered on a data point and the variance $\rightarrow 0$ then the DLL $\rightarrow \infty$.
- But K can be automatically selected by, *e.g.*, optimizing over the clustering-distortion objective with a model-order penalty.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Model-Order Selection

- Suppose a model type $m \in \mathcal{M}$ with DLL

$$\mathcal{L}(\underline{\theta}; m) = \sum_{\underline{x} \in \mathcal{X}} \log f(\underline{x}; \underline{\theta}, m)$$

has $d(m)$ associated ML parameters $\in \Theta_m \subset \mathbb{R}^{d(m)}$:

$$\underline{\theta}_m^* = \arg\max_{\underline{\theta} \in \Theta_m} \mathcal{L}(\theta; m).$$

- *Bayesian Information Criterion* (BIC) cost

$$-\mathcal{L}(\underline{\theta}_m^*; m) + \frac{d(m)}{2} \log(2\pi T),$$

can be minimized to select among different models $m \in \mathcal{M}$ for the dataset $\mathcal{X}$.

- The second term be interpreted as the number of bits needed to describe the model, and
- The first (negative data log-likelihood) term as the number of bits to describe the data given knowledge of the model.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Model-Order Selection (cont)

- Thus, the data log-likelihood trades off with the model order (model complexity) in the BIC objective.
- For high-dimensional data:
  - Minimizing BIC may result in "degenerative," e.g., single component, solutions.
  - To address this problem, [Graham & Miller, TSP'06] considered how model parameters may be shared across components.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Principal Component Analysis (PCA) and Singular-Value Decomposition (SVD)

- PCA is a linear transform technique for reducing the effective dimensionality of a feature vector $\underline{x} \in \mathbb{R}^N$, while introducing the least amount of distortion/error in the resulting approximation $\hat{\underline{x}}$ of $\underline{x}$.
- For classification problems, using PCA may
  - reduce training data requirements to achieve an accurate classifier, *i.e.*, to combat the so-called "curse of dimensionality," and
  - avoid "gross model order under-estimation" for mixture models.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

PCA

- In PCA, a feature vector $\underline{x} \in \mathcal{X} = \{\underline{x}_i, i = 1, \ldots, T\}$ is approximated as:

$$\hat{\underline{x}} = \sum_{j=1}^{J} \beta_j \underline{q}_j + \underline{m},$$

where:

- $\{\underline{q}_j\}_{j=1}^{J}$ are a set of *orthonormal vectors* for representing any $\underline{x} \in \mathcal{X}$ with $J < N$;
- the "mean" vector $\underline{m}$ is also common to all $\underline{x} \in \mathcal{X}$; and
- $\underline{\beta} = (\beta_1, ..., \beta_J)' \in \mathbb{R}^J$ are the corresponding "optimal" coefficients for representing $\underline{x}$, *i.e.*, these coefficients are chosen to minimize the MSE "distortion":

$$\frac{1}{T} \sum_{i=1}^{T} \|\underline{x}_i - \hat{\underline{x}}_i\|^2.$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

PCA (cont)

- The optimal choice of the mean vector, again in the sense of this MSE distortion, is the empirical mean of $\mathcal{X}$,

$$\underline{m} = \frac{1}{T} \sum_{i=1}^{T} \underline{x}_i.$$

- Basis vectors $\underline{q}_j$ are called "components" (different from components of mixture densities).

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

PCA pre-processing before classification or prediction

- Rather than classifying (or detecting) based on $\underline{\hat{x}}$ explicitly, when PCA is used the input to the classifier (or detector) is the vector of coefficients $\underline{\beta}$,
- where in general one chooses $J \ll N$ to achieve substantial dimension reduction.
- Likewise, in a density modelling framework (that could be part of a statistical anomaly detector), one would learn the joint density for $\underline{\beta}$, rather than for $\underline{\hat{x}}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

PCA: A simple example

- Suppose $J = 1$, *i.e.*, where $\hat{\underline{x}}_i = \beta_{1,i}\underline{q}_1 + \underline{m}$.
- To find the basis vector $\underline{q}_1$ and coefficients (with one coefficient β per data point $\underline{x} \in \mathcal{X}$), one poses a squared error estimation problem on the given dataset $\mathcal{X}$,

$$\min_{\beta_{1,1},\beta_{1,2},\ldots,\beta_{1,T},\underline{q}_1} \sum_{i=1}^{T} \|\underline{x}_i - \beta_{1,i}\underline{q}_1 - \underline{m}\|^2.$$

- First, suppose that the optimal (unit-norm) $\underline{q}_1$ has already been determined.
- Then, it is easily found (*e.g.*, by taking derivatives) that the optimal $\beta_{1,i}$ (in the MSE sense) are:
  $\beta_{1,i} = \underline{q}_1'(\underline{x}_i - \underline{m}),\ \ i = 1, \ldots, T.$
- That is, they are obtained simply by *projecting* each (centered) data point onto the basis vector $\underline{q}_1$.
- Note that this is true *irrespective of* the choice of $\underline{q}_1$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

PCA: A simple example (cont)

- Substituting the expression for the optimizing $\beta_{1,i}$ gives

$$-\underline{q}_1' \left(\sum_{i=1}^{T} (\underline{x}_i - \underline{m})(\underline{x}_i - \underline{m})' \right) \underline{q}_1 + \sum_{i=1}^{T} \|\underline{x}_i - \underline{m}\|^2.$$

- Note that the term in the large parentheses is a scaled sample covariance matrix, often referred to as the *scatter matrix*.
- Choosing $\underline{q}_1$ to minimize this expression, s.t. $\underline{q}_1$ is a unit vector, yields the solution that $\underline{q}_1$ is the *principal eigenvector of the scatter matrix*,
- *i.e.*, the eigenvector with largest eigenvalue. Moreover, it is also clear from this expression that the minimum MSE choice for $\underline{m}$ is the empirical mean.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

PCA: A simple example (cont)

- Similarly, if one considers $\hat{\underline{x}} = \sum_{j=1}^{J} \beta_j \underline{q}_j + \underline{m}$ for $J > 1$, the minimum MSE solution is to choose $\{\underline{q}_j\}_{j=1}^{J}$ as the J principal (orthonormal) eigenvectors of the scatter matrix (*i.e.*, those with the largest eigenvalues),
- with the optimal coefficients $\underline{\beta}_i = \{\beta_{j,i}\}_{j=1}^{J}$ obtained by projecting $\underline{x}_i$ onto each of the $\underline{q}_j$, as above for $\underline{q}_1$.
- Thus, PCA can be performed in practice via eigendecomposition on the sample covariance matrix (using any established technique), retaining the J components corresponding to the largest eigenvalues.
- Note that feature compaction can instead be achieved by an auto-encoder DNN.
- Also, a technique of feature *selection* is described in the Appendix on SVMs.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Some Detection Basics

- Given an observed feature vector $\underline{x}$, suppose one wishes to distinguish between two generative hypotheses:
  - H_0 is by convention referred to as the "null" hypothesis, while
  - H_1 is the "alternative" hypothesis.
- Also if, optimistically, one has a training set for *each* of these two hypotheses, then one possibility is to train a statistical classifier via supervised learning which can then be used to decide between the hypotheses for any given (unlabeled) $\underline{x}$.
- Another approach is to estimate a generative model (density) f_i for each H_i using its training set, and decide H_0 if the likelihood ratio $f_0(\underline{x})/f_1(\underline{x}) \geq \eta > 0$ (some threshold η), else decide H_1.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

True and False Positive Rates

- The true positive rate (TPR), also called the *power* of the test, is the probability of correctly deciding H_1, *i.e.*,

$$\mathrm{P}(f_0(\underline{X})/f_1(\underline{X}) < \eta \mid H_1).$$

- The false positive rate (FPR, or false alarm probability) is the probability of deciding H_1 when H_0 is true, *i.e.*,

$$\mathrm{P}(f_0(\underline{X})/f_1(\underline{X}) < \eta \mid H_0).$$

- Each threshold value η represents a distinct tradeoff between power and FPR.
- The likelihood ratio test, assuming accurate null and alternative models, has the highest power among all possible tests, given the FPR fixed.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Statistical Anomaly Detection

- What if there is a training set of examples from H_0, but no examples from H_1?
- Here, H_0 represents something that is known or "normal", and one wishes to identify whether a given $\underline{x}$ is unlikely to have been generated under H_0.
- It is also useful to assess *how unlikely* H_0 is.
- Statistical anomaly detection applies a threshold to a *detection statistic* that is a function of the observation $\underline{x}$,
- e.g., $f_0(\underline{x})$.
- But if f_0 is multimodal, small $f_0(\underline{x})$ may not reliably detect anomalies.
- Instead use a *p-value*, i.e., the probability of observing a value "more extreme" than $\underline{x}$ under H_0,

$$\mathrm{P}(f_{\underline{X}|H_0}(\underline{X}) < f_{\underline{X}|H_0}(\underline{x}) \mid H_0).$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

p-values for Scalar Random Objects (Random Variables)

- For a (scalar) random variable X, one can define "one-sided" p-values of x as $\mathrm{P}(X > x|H_0)$ (right tail) or $\mathrm{P}(X < x|H_0)$ (left tail), particularly when $x > \mathrm{E}(X|H_0)$ or $x < \mathrm{E}(X|H_0)$ respectively.
- Such one-sided p-values represent the FPR when the detection threshold is x itself.
- For a Gaussian null density the right-sided p-value at $x > \mu$ is

$$\int_x^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}} \mathrm{d}z,$$

where σ^2 is the conditional variance of X given H_0.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

p-values for Random Variables (cont)

- Note that if $F(x) = \mathrm{P}(X \leq x)$, $x \in \mathbb{R}$, *i.e.* if it is the cumulative distribution function (cdf) of random variable X and $1 - F(X)$ is the right-sided p-value of a random sample X, then for arbitrary $x \in [0, 1]$,

$$\mathrm{P}(1-F(X) \geq x) = \mathrm{P}(X \leq F^{-1}(1-x)) = F(F^{-1}(1-x)) = 1-x.$$

- Thus, the right-sided p-value

$$1 - F(X) \sim \text{uniform}[0, 1].$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023

p-values for Anomaly Detection

- If the p-value is ϕ then thresholding at ϕ sets the FPR to ϕ when deciding for $\underline{x}$, assuming the null model is accurate.
- If the null model is inaccurate,
  - then the detection rule could either be too liberal (making too many false detections)
  - or conservative (making too few false detections,
  - and possibly too few true detections).

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Performance Measures for Detection: TPR, FPR, ROC

- TPR (power) is the probability of correctly rejecting the null hypothesis, H_0, and
- FPR is the probability of *falsely* rejecting the null hypothesis.
- Clearly, there is a tradeoff between power and FPR, which can be controlled by the choice of the detection threshold.
- The Receiver Operating Characteristic (ROC) curve is a plot of power versus FPR, which can be generated by sweeping over a sequence of detection thresholds with increasing FPR,
  - starting with a threshold achieving zero FPR (and possibly zero power as well), and
  - ending with a threshold achieving both FPR and power equal to 1 (*i.e.*, where everything is decided to H_1).

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Performance Measures for Detection: ROC AUC, ACC

- The ROC Area Under the Curve (ROC AUC) is a comprehensive measure of detection performance, with a maximum value of 1.0.
- One can also evaluate the curve only up to a maximum tolerable FPR, *e.g.*, $\delta < 1$, and then measure the area under the partial ROC curve (with maximal attainable value of δ).
- Accuracy (ACC) is just the probability of a correct decision,
- e.g., when there are just two choices, ACC is just the (unconditional) probability of a true positive or true negative:

$$\mathrm{P}(f_0(\underline{X})/f_1(\underline{X}) < \eta \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(f_0(\underline{X})/f_1(\underline{X}) \geq \eta \mid H_0)\mathrm{P}(H_0).$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Performance Measures: Discussion

- In some applications the FPR must be kept to a very small value, even though this will also limit the attainable TPR.
- E.g., in a cyber security context, each flagged detection may require human analyst confirmation or action (to mitigate the attack).
- Thus, a "too-high" FPR will overwhelm the analyst.
- In practice, one can use a held-out set of samples generated according to H_0 to estimate the FPR, with the detection threshold varied until the desired (estimated) FPR is achieved.
- The power can also be estimated *operationally*, *e.g.*, using the human analyst to label detections as true or false, and estimating the TPR based on the normalized count of true positives (similarly, one can operationally estimate the FPR).
- In a supervised setting, where one has access to known examples from both H_0 and H_1, ROC AUC, for example, can be used to benchmark-compare different detectors.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Confidence Intervals and Cross Validation

- Experiments involving deep neural networks have substantial sources of randomness together with a potentially enormous number of hyperparameters.
- This is why it's often very difficult for other parties to precisely replicate deep learning experiments.
- In the presence of experimental randomness, independent trials can be repeated to obtain and report statistical confidence intervals for detection performance.
- Statisical confidence is based both on the law of large numbers and the central limit theorem.
- E.g., for an estimated probability p over n trials of a Boolean detection decision, the sample standard deviation is $\sqrt{p(1-p)/n}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Confidence Intervals and Cross Validation (cont)

- Under cross validation, one can, e.g., randomly partition the labelled data into ten equally sized sets and conduct experiments wherein, e.g.,
  - eight of the sets are used for deep learning,
  - one of the sets is used to set some of the hyperparameters, and
  - the remaining set is used as a test-set to evaluate performance.
- Such experiments are repeated so that all sets have a "turn" as the test and hyperparameter folds, and
- performance results are averaged over the experiments.
- A challenge with deep learning experiments is that each such experiment/trial often requires a lot of computation.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT