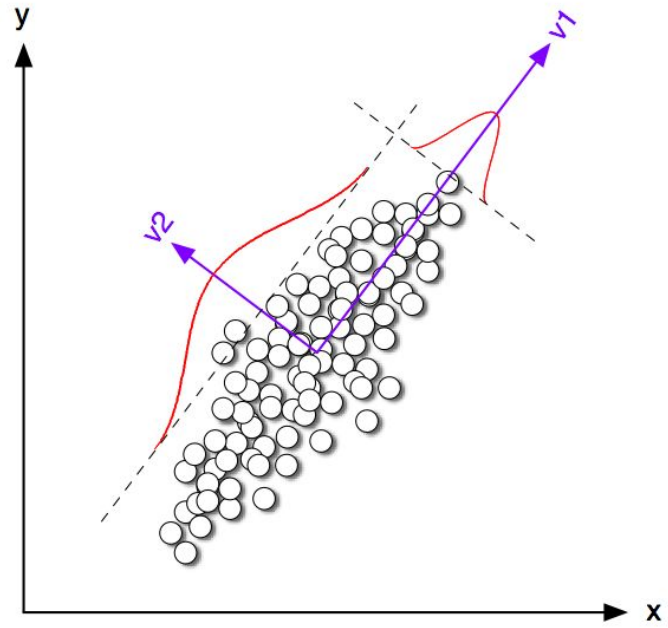
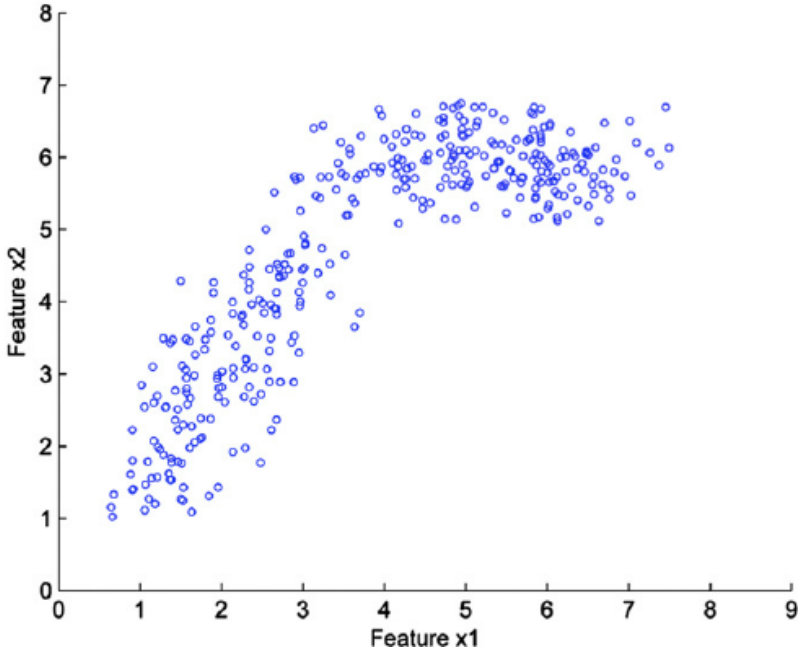
Principal Component Analysis: Linear and Nonlinear Methods

Suyash P. Awate

Principal Component Analysis

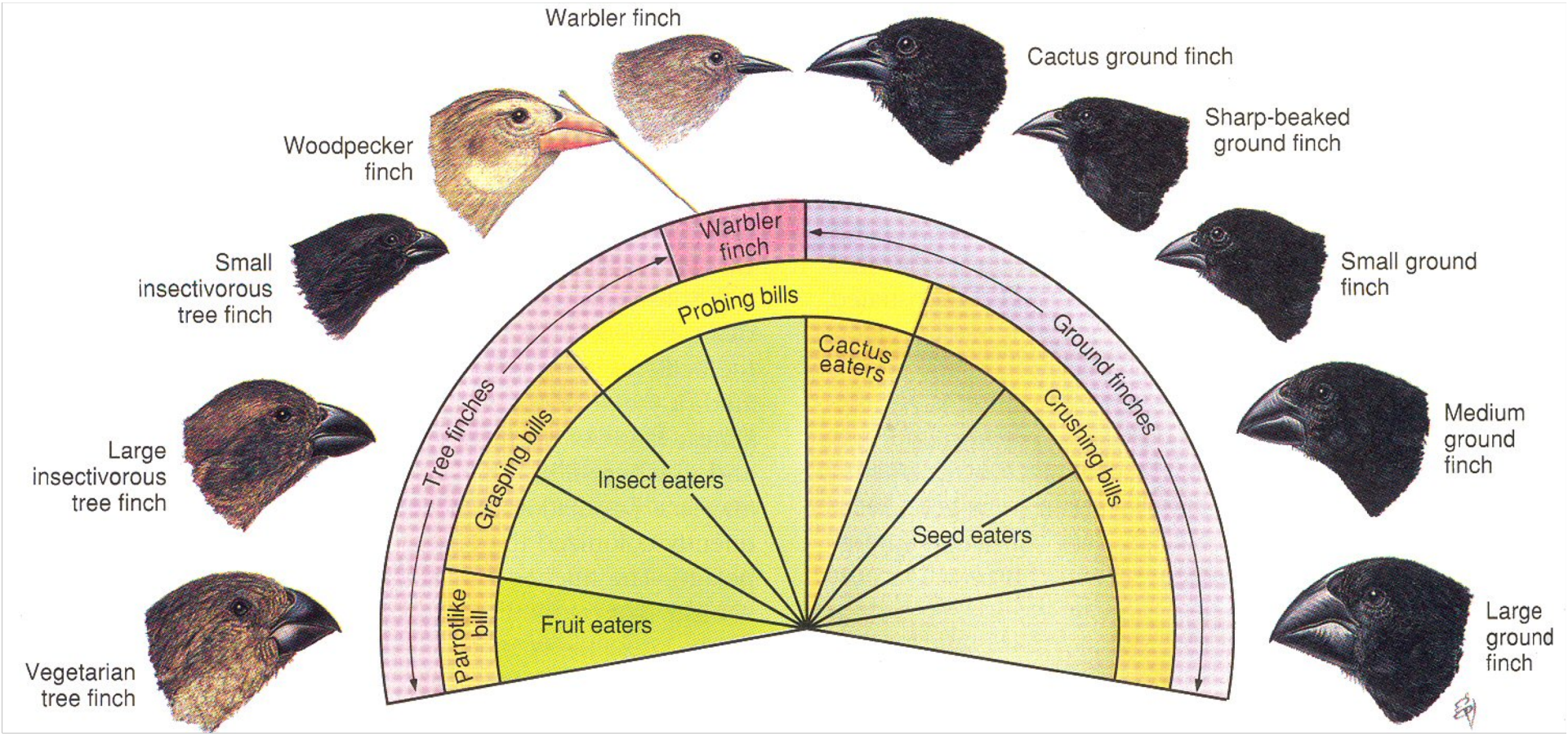
- Statistical model of the data
  - Finding the mean of the data sample
  - Finding modes of variation in data sample
- Very closely related to fitting a multivariate Gaussian model to the data





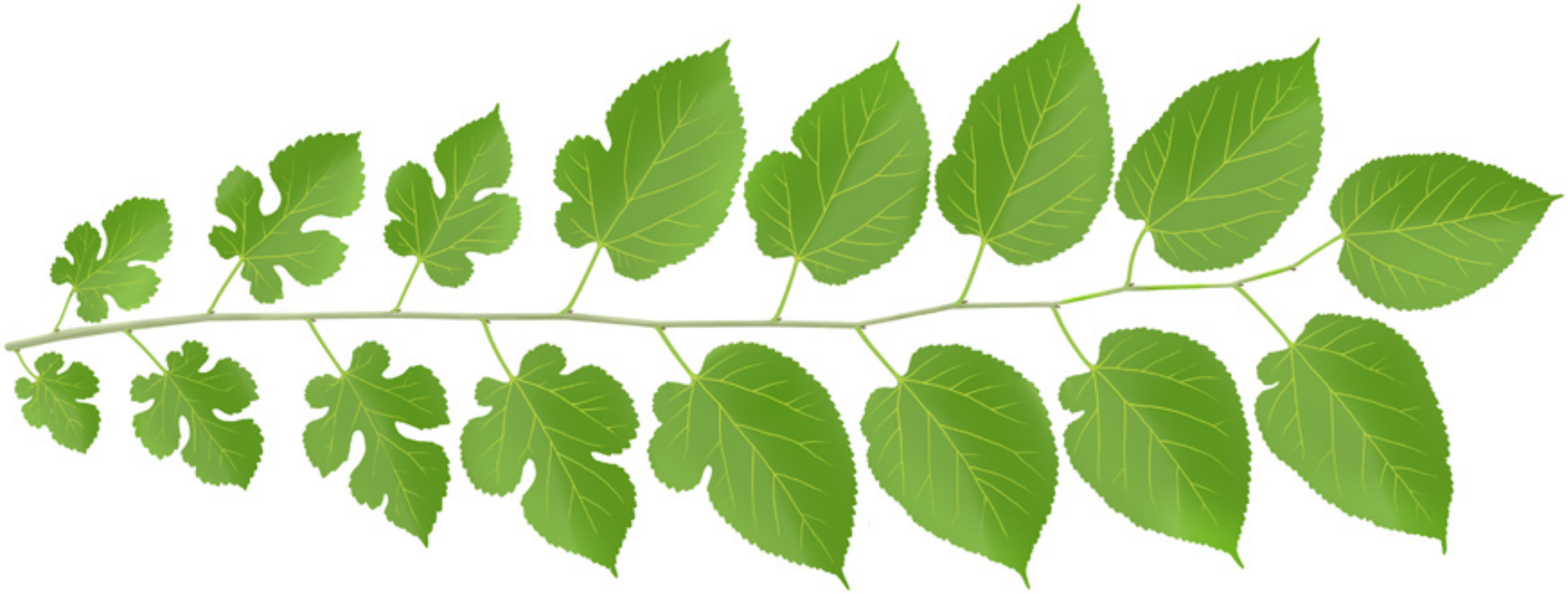
Principal Component Analysis

- Darwin's finches
  - Variation of beak's shape and size



Principal Component Analysis

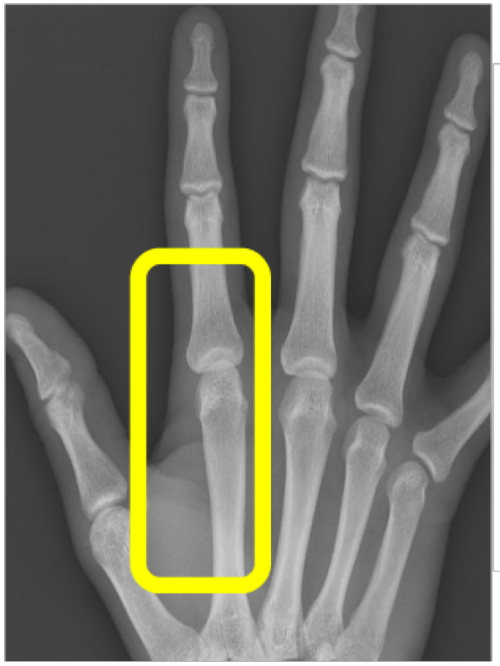
- White mulberry leaves
  - Simpler leaves reflect maturity



http://www.washingtonpost.com/wp-srv/special/metro/urban-jungle/pages/120724.html

Principal Component Analysis

- Human wrist bone



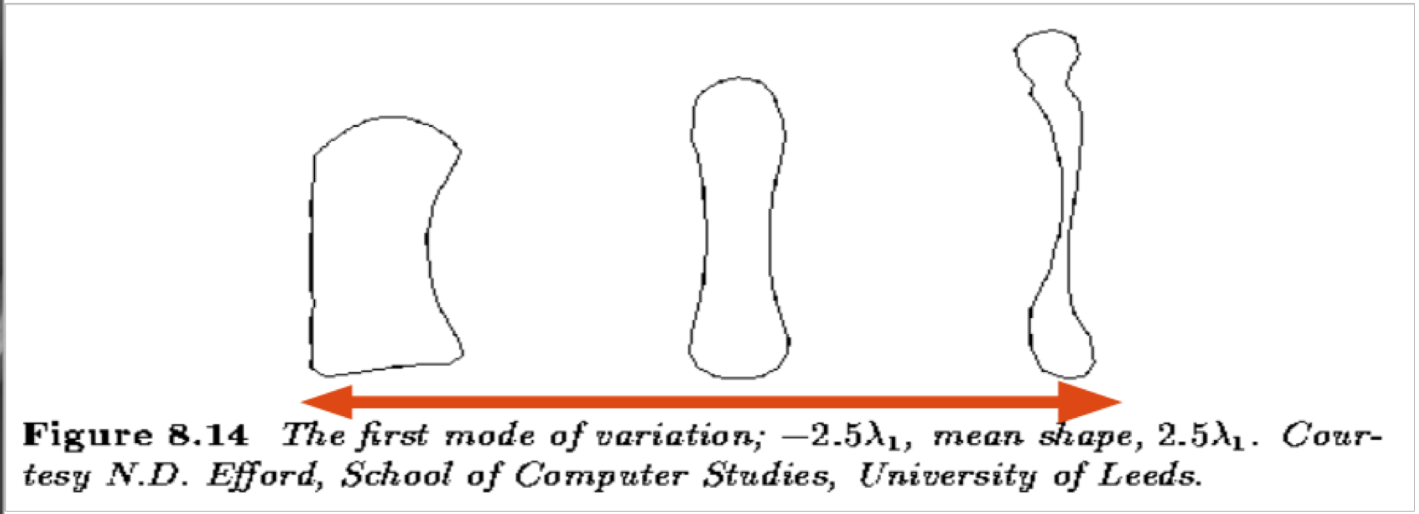
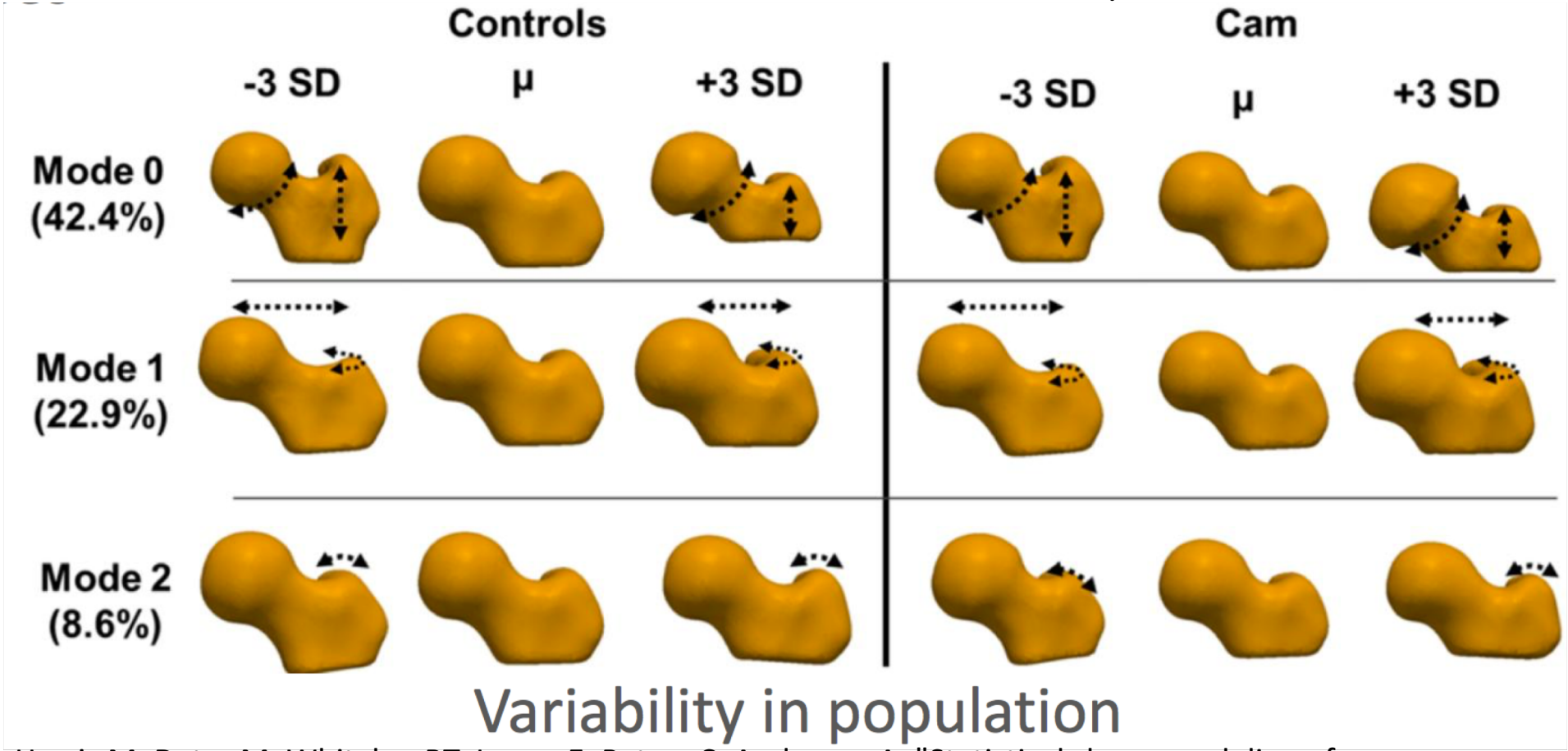


Figure 8.14 *The first mode of variation; $-2.5\lambda_1$, mean shape, $2.5\lambda_1$. Courtesy N.D. Efford, School of Computer Studies, University of Leeds.*

Principal Component Analysis

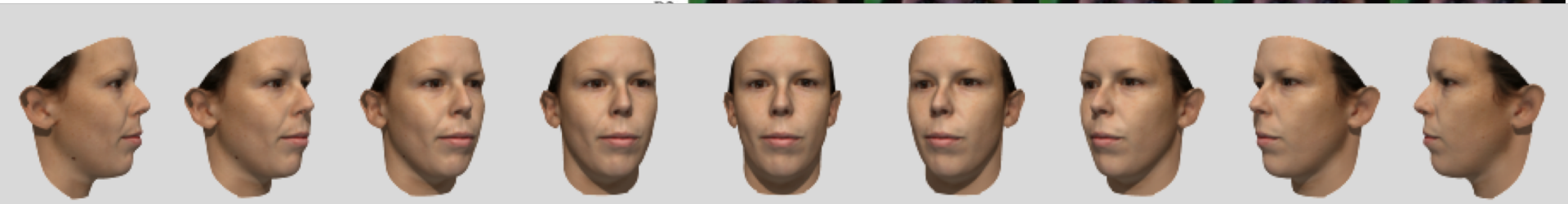
- Human hip bone
  - Modes of variation in normal cases, diseased cases



Harris M, Datar M, Whitaker RT, Jurrus E, Peters C, Anderson A. "Statistical shape modeling of cam femoroacetabular impingement." *Journal of Orthopaedic Research* 31, no. 10 (2013): 1620-1626.

Principal Component Analysis

- Face recognition
  - Variability in appearance



Pose Variations (-70, -50, -30, -15, 0, 15, 30, 50, 70 degree rotation around Y-axis)

Illumination Variations (0, 30, 50 degree rotation around Y-axis)

Multivariate Gaussian

- Generalizes a univariate Gaussian probability density function (PDF)
  - Standard normal density function = 1D Gaussian with mean 0, standard deviation 1
- Consider a vector random variable (RV) $X := [X_1, X_2, \ldots, X_D]^T$

Multivariate Gaussian

- **Definition**
  - The random variable (RV) $X := [X_1, X_2, \ldots, X_D]^T$ has a multivariate (jointly) Gaussian PDF if there exists a finite set of independent and identically distributed (i.i.d.) univariate standard-normal RVs $W_1, \ldots, W_N$ (with $D <= N$) such that each X_d can be expressed as

$$X_d = \mu_d + \sum_n A_{dn} W_n \text{ (i.e., } X = AW + \mu)$$

Multivariate Gaussian

- Example 1: **Zero-Mean Isotropic Gaussian**
  - Consider independent standard-normal RVs $W_1,\ldots,W_D$ with $A = I_{D\times D}$ and $\mu = 0$
  - Then, $X = AW + \mu = W$
  - Then, the Gaussian PDF is:

$$p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{d/2}} \exp(-0.5 w^T w)$$

Multivariate Gaussian

- Transformation of random variables
  - Univariate case: W is 1-dimensional with PDF P(W)
    - Consider a function g(.) such that X := g(W)
    - Then, PDF of X is:

$$p(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right|$$

  - Multivariate case: W is D-dimensional with PDF P(W)
    - Consider the transformation X := A W
    - Then, the PDF of X is:

$$q(X) = p(A^{-1}X) \frac{1}{\det(A)} = \frac{1}{(2\pi)^{d/2} \det(A)} \exp(-0.5 X^T (A^{-1})^T A^{-1} X)$$

    - If C := AA^T (where C has a special name), then $q(X) = \frac{1}{(2\pi)^{d/2} |C|^{0.5}} \exp(-0.5 X^T C^{-1} X)$

Multivariate Gaussian

- Example 2: **Zero-Mean Anisotropic Gaussian**
  - Consider the case where
    $\mu = 0$ and A is a square non-singular matrix
  - Then, the Gaussian PDF is:

$$q(X) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5X^T C^{-1} X)$$

  - Property
    - When X = AW,
      E[X] = E[AW] = A E[W] = 0

Multivariate Gaussian

- Example 3: **Anisotropic Gaussian**
  - When X = AW and Y = X + μ = **AW + μ**
    - E[Y] = E[AW + μ] = A E[W] + μ = μ
    - And the PDF is:

$$p(y) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5(y-\mu)^T C^{-1}(y-\mu))$$

Multivariate Gaussian

- For any multivariate RV X
  - Covariance C := E [$(X-\mu)(X-\mu)^T$]
  - Thus, C_{ij} = E [$(X_i-\mu_i)(X_j-\mu_j)^T$] = Cov (X_i, X_j)
- Property: Cov (W) = E [WW^T] = I
  - Cov (W_i, W_i) = 1 (standard normal PDF)
  - Cov (W_i, $W_{j\sim i}$) = 0 (independence)
- For a multivariate Gaussian RV X = AW + μ
  - Property: Cov (X) = Cov (AW + μ)
    = E [$(AW)(AW)^T$] = A^T E[WW^T] A = A^T A

$$p(y) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5(y-\mu)^T C^{-1}(y-\mu))$$

Multivariate Gaussian

- **Covariance matrix**
  - Property: C is **symmetric**
    - $C_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = C_{ji}$
  - Property: C is **positive semi-definite** (PSD)
    - For any non-zero vector 'a',
      $E[a^T (X - E[X])(X - E[X])^T a] = E[(f(X))^T f(X)] >= 0$
      = variance of scalar RV $f(X) := (X - E[X])^T a$

Multivariate Gaussian

- **Eigen Decomposition**
  - Every square matrix M
    has an eigen-decomposition $M = Q \Lambda Q^{-1}$
    where
    Q is an invertible matrix
    Λ is a diagonal matrix
  - Every square real symmetric matrix M
    (like the covariance matrix C)
    has an eigen-decomposition $M = Q \Lambda Q^{T}$
    where
    Q is an orthogonal matrix
    Λ is a diagonal matrix

Multivariate Gaussian

- Eigen Decomposition
  - Every square real symmetric PSD matrix M
    (like the covariance matrix C)
    has an eigen-decomposition $M = Q \Lambda Q^T$
    where
    Q is an orthogonal matrix
    Λ is a diagonal matrix with nonnegative entries

Multivariate Gaussian

- **Contours of the multivariate Gaussian PDF**
  - Property: If X is multivariate Gaussian in 2D with a diagonal (invertible) covariance C, then iso-probability contours of P(X) are ellipses whose axes are aligned with the cardinal axes
  - Proof:
    - Let C be symmetric positive definite (SPD)
    - C^{-1} is also SPD
      - C has a Cholesky decomposition $C = MM^T$, where M is upper triangular with positive diagonal entries
      - $C^{-1} = (M^T)^{-1} (M)^{-1} = NN^T \rightarrow$ SPD
    - A contour $\{x: P(x) = \alpha\}$
      $= \{x: (x-\mu)^T C^{-1} (x-\mu) = \beta\}$ where $\beta>0$ because C^{-1} is SPD
      $= \{x: \Sigma_d (x_d - \mu_d)^2 / (\beta\sigma_d^2) = 1\}$ that is an ellipse

Multivariate Gaussian

- **Mahalanobis Distance**
  - $(x-\mu)^T C^{-1} (x-\mu)$
  - Multidimensional generalization of Euclidean distance from mean μ
    - Rescales units along each dimension, based on the variance of the data along that dimension
  - Iso-probability contour of a Gaussian is the locus of points with the same Mahalanobis distance
  - Theorem:
    Mahalanobis distance is a true distance metric
    - Satisfies positivity, symmetry, triangular inequality

Multivariate Gaussian

- **Scaling and Rotating the coordinate frame**
  - Let X := SW
    - S = diagonal matrix that rescales units along each coordinate axis
    - Covariance matrix $C = SS^T = S^2$
    - Mahalanobis distance between x and mean (origin) is $x^T C^{-1} x = x^T S^{-2} x$

Multivariate Gaussian

- Scaling and Rotating the coordinate frame
  - Let **Y := UX = USW**, where U is a rotation matrix that rotates the coordinate frame
    - Covariance matrix $C' = AA^T = (US)(US)^T = US^2U^T$
    - Mahalanobis distance, w.r.t. C', from **0** to $y:=Ux$ is $y^TC^{-1}y$ $= (Ux)^T (US^{-2}U^T) (Ux) = x^T S^{-2} x$
      = same as Mahalanobis distance, w.r.t. C, from **0** to x
    - Rotating data x simply rotates the iso-probability contours of P(X)

Multivariate Gaussian

- **Connections to PCA**
  - Let us apply **$A := USV^T$** to a **zero-mean isotropic Gaussian PDF**
  - (1) Rotation V^T doesn't change PDF
    - Covariance C = Identity
  - (2) Scaling S scales each dimension d by S_{dd} making the PDF anisotropic
    - Consider distinct $S_{11} > S_{22} > \ldots$
    - Covariance $C = S^2$
    - Variance is S_{dd}^2 along axis d (modes of variation = axes)
  - (3) Rotation U rotates the anisotropic PDF
    - Variance is S_{dd}^2 along U_d (modes of variation = U_d)
    - Covariance $C = US^2U^T$

Multivariate Gaussian

- Connections to PCA
  - Given: Data $\{x_i : i=1,\ldots,n\}$
  - Goal: Estimate modes of variation
  - Algorithm:
    1) Find maximum likelihood estimate (MLE) of covariance C
       - MLE for Gaussian mean is sample mean
       - MLE for Gaussian covariance is sample covariance
       - Consistent estimator
    2) Find eigen decomposition of C to give U and S:
       - Modes of variation U_i
       - Variance S_{ii}^2 along each mode of variation

Multivariate Gaussian

- **Directions of maximal variance**
  - Consider data $\{x_i : i=1,...n\}$ drawn from a Gaussian PDF with mean **$\mu = 0$ and <u>diagonal covariance</u> C**
  - For a Gaussian PDF, a diagonal covariance implies:
    - $X^d = (C_{dd})^{0.5} W^d$ where
      W^d is a standard Normal RV
  - Question:
    Find the direction / unit-vector v such that the data projected on the subspace denoted by vector v passing through the mean (origin) has maximal variance
  - Answer (derivation) follows

Multivariate Gaussian

- Directions of maximal variance
  - Projected data = $\langle x_i, v \rangle$ v
  - Mean of the projected data = $\Sigma_i \langle x_i, v \rangle$ v
    = $\langle \Sigma_i x_i, v \rangle$ v = $\langle 0, v \rangle$ v = 0
  - Projected data is 1D
  - Distance of projected data from the mean
    = $\|\langle x_i, v \rangle v\|_2 = |\langle x_i, v \rangle|$
  - Variance of projected data = $\Sigma_i \langle x_i, v \rangle^2$
  - Optimal direction $= \arg\max_{v:\|v\|_2=1} \sum_i \langle x_i, v \rangle^2$

Multivariate Gaussian

- Directions of maximal variance
  - Optimal direction

$$
\begin{aligned}
&= \arg\max_{v:\|v\|_2=1} \sum_i (x_i^T v)^2 \\
&= \arg\max_{v:\|v\|_2=1} \sum_i (x_i^T v)^T (x_i^T v) \\
&= \arg\max_{v:\|v\|_2=1} \sum_i v^T x_i x_i^T v \\
&= \arg\max_{v:\|v\|_2=1} v^T \left(\sum_i x_i x_i^T\right) v \\
&= \arg\max_{v:\|v\|_2=1} v^T C v
\end{aligned}
$$

    - where C = sample covariance
    - This is the connection between sample covariance C and direction v maximizing variance of projected data

Multivariate Gaussian

- Directions of maximal variance
  - Optimal direction

    $= \arg\max_{v:\|v\|_2=1} \sum_d C_{dd}(v^d)^2$

    - because C is diagonal
  - This is maximized when

    $v^d = 1$ for $d = \arg\max_e C_{ee}$ and $v^d = 0$ otherwise

    - Put all "weight" on that component of v that is associated with the maximum of diagonal elements in C
    - Constraint set = hypersphere.
      Contours of objective function are ellipsoids; minor axis = dimension corresponding to $\min_d (1/C_{dd})$ or $\max_d (C_{dd})$.
      Point, on hypersphere, maximizing objective function lies at intersection of minor axis with hypersphere

Multivariate Gaussian

- Directions of maximal variance
  - Question: Find 2nd direction u that is:
    - (i) orthogonal to v and
    - (ii) maximizes variance of data projected onto it
  - Optimal direction

$$= \arg\max_{u:\|u\|_2=1, u\perp v} \sum_i \langle x_i, u\rangle^2$$

$$= \arg\max_{u:\|u\|_2=1, u\perp v} \sum_d C_{dd}(u^d)^2, \text{ where we know } C_{dd} \geq 0$$

  - This is maximized when

$$u^c = 1 \text{ for } c = \arg\max_{d\neq e} C_{dd} \text{ and } u^c = 0 \text{ otherwise}$$

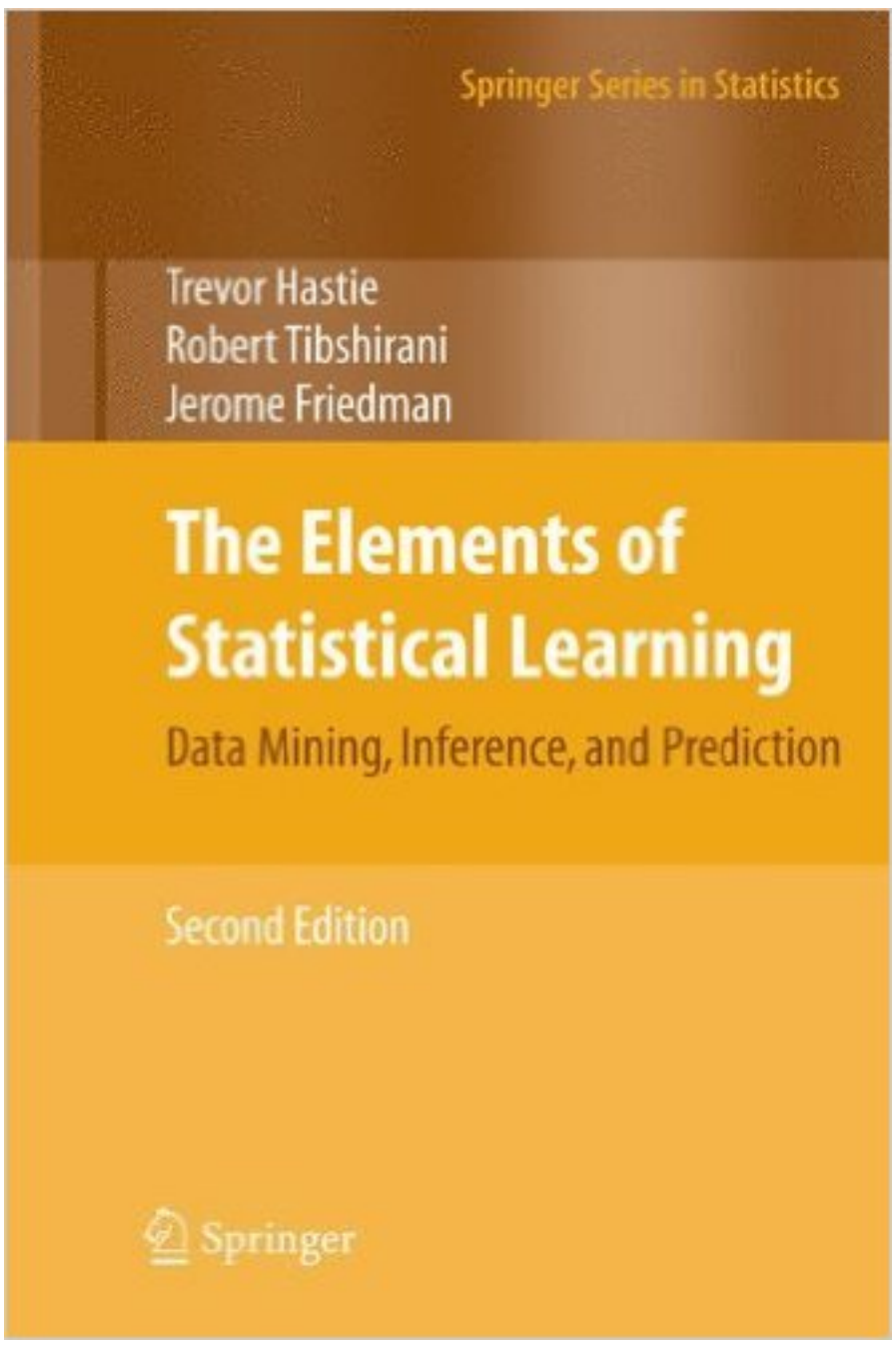
    - Put all "weight" on that component of u that is: (i) not the d chosen before and (ii) corresponds to the maximum of the remaining diagonal elements in C

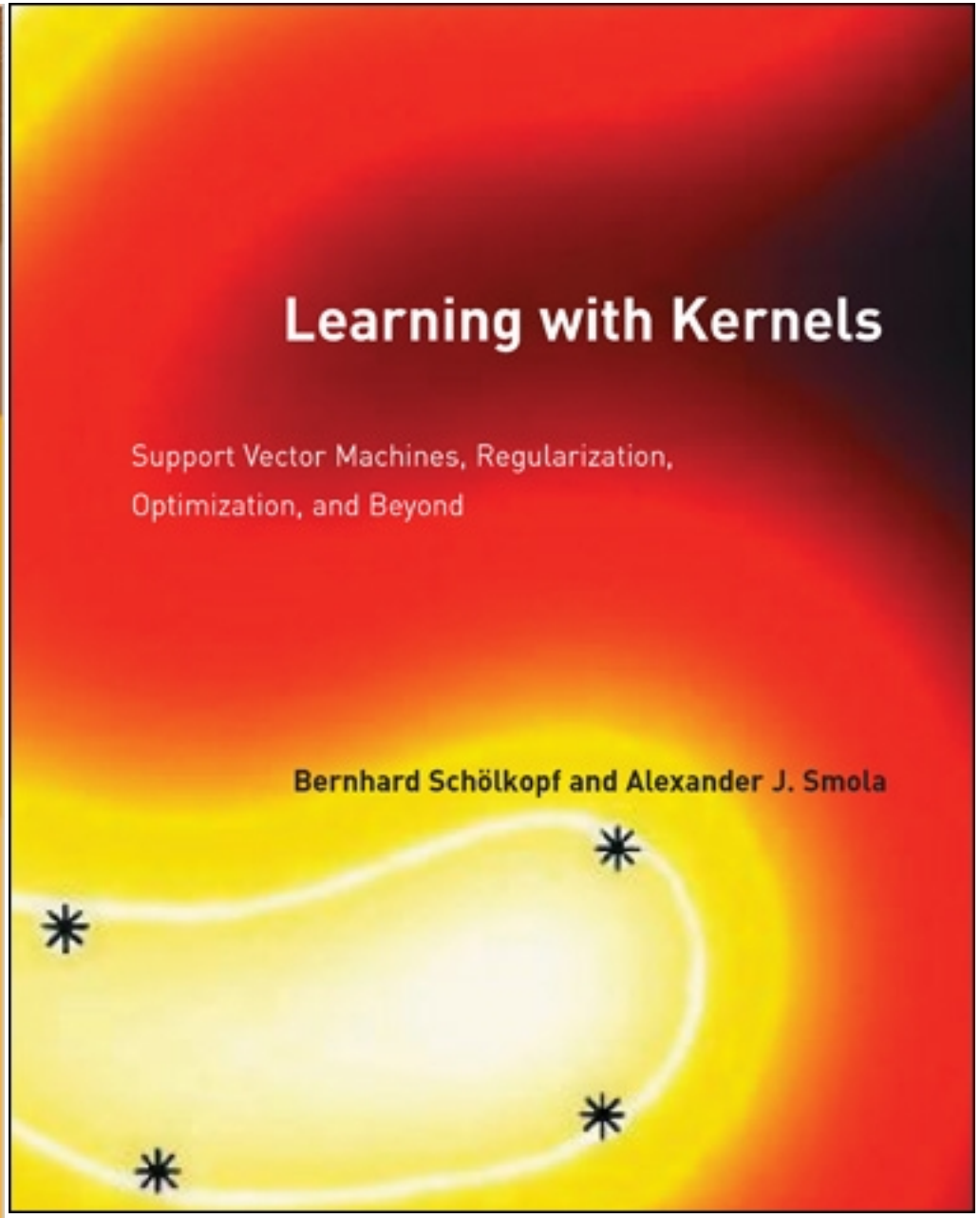
Multivariate Gaussian

- Directions of maximal variance
  - Similar arguments hold for 3rd, 4th, … directions
  - Theorem:
    For data drawn from a Gaussian
    with mean 0 and a diagonal covariance matrix,
    the directions maximizing variance are
    the cardinal directions
  - These directions = **principal components of variation**
  - Rotating the coordinate frame by pre-multiplication with a orthogonal matrix U simply rotates the principal components

Multivariate Gaussian

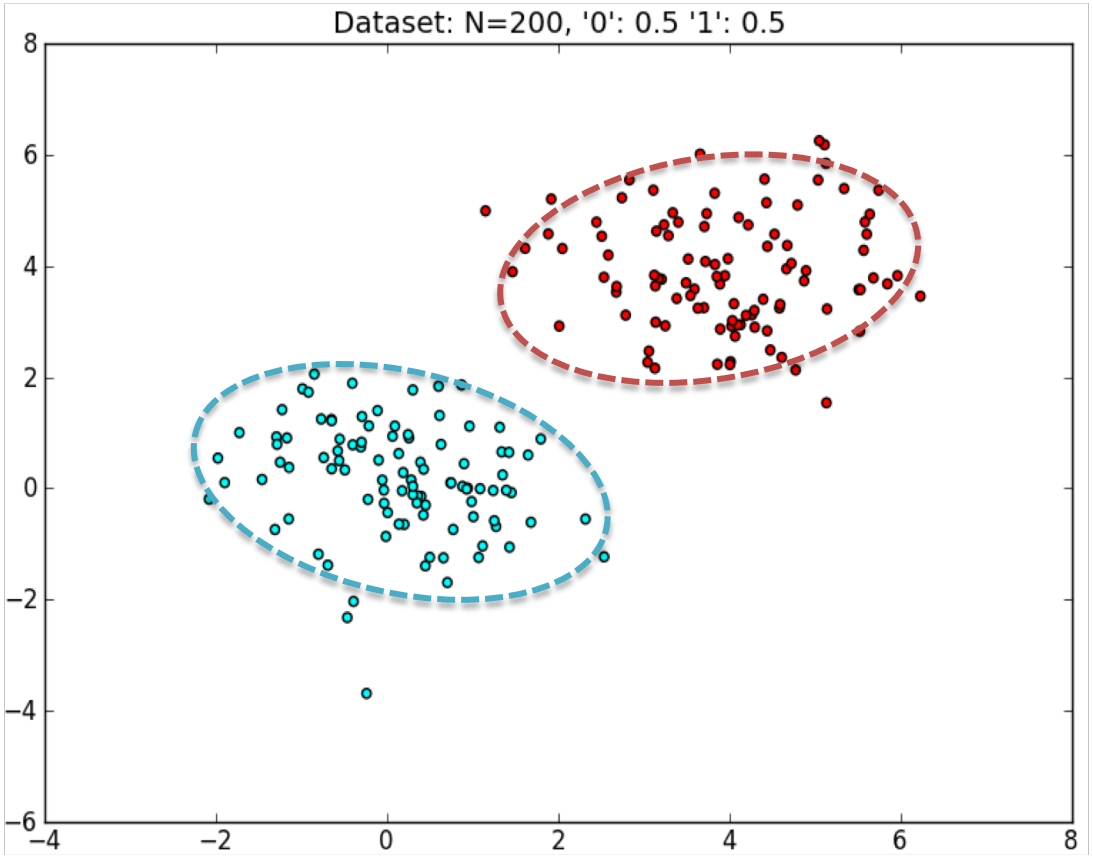
- PCA and eigen decomposition
  - Theorem:
    The **principal directions** U for
    data that are observations of $X := AW + \mu$
    are the **eigenvectors** of the sample (empirical)
    covariance $C = AA^T$
  - **Variances** along principal directions are the **eigenvalues** of the sample (empirical) covariance matrix C


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer


Learning with Kernels
Support Vector Machines, Regularization, Optimization, and Beyond
Bernhard Schölkopf and Alexander J. Smola

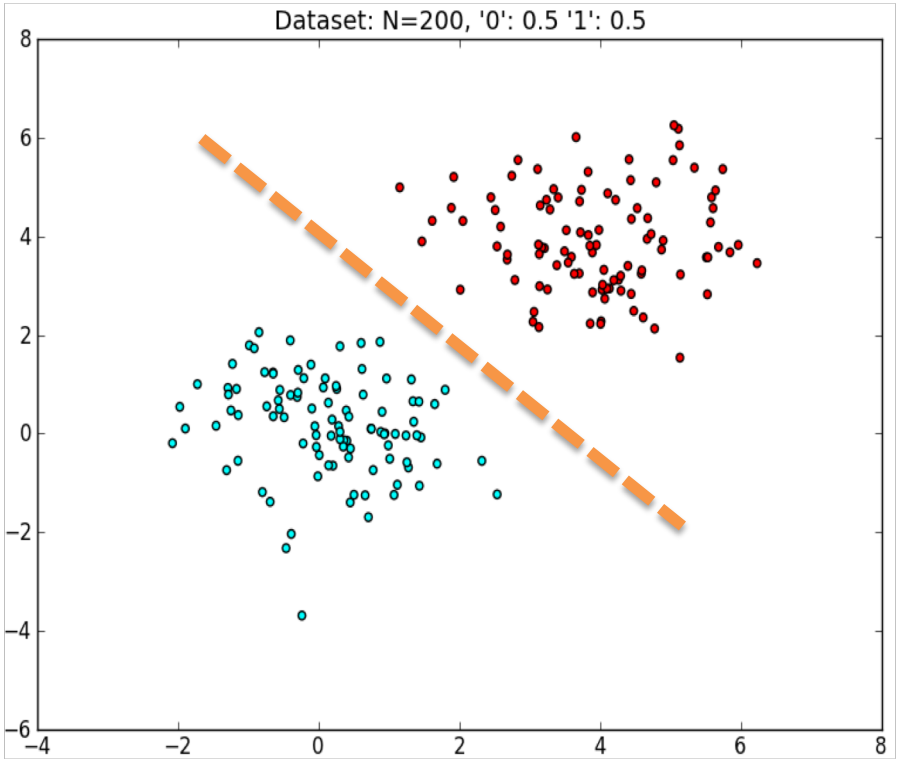
Modeling Data

- **Generative modeling**
  - Data in ellipsoidal clusters / classes
  - Use a Gaussian to model the data
  - How to classify ?
    - For a test point, assign point to cluster that gives larger probability



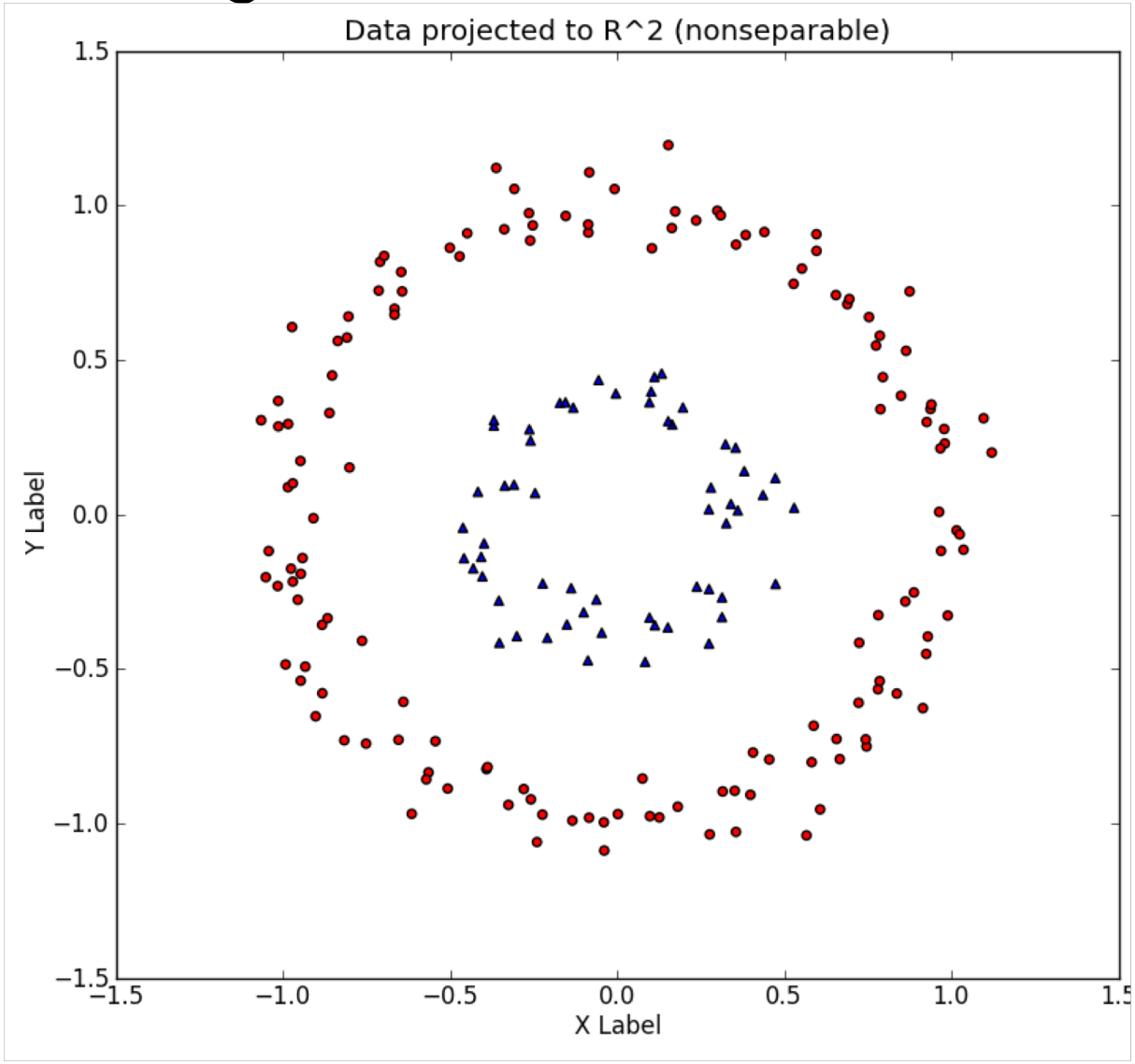
Modeling Data

- **Discriminative modeling** (for classification)
  - Data linearly separable
  - Use a line / plane to separate 2 classes
  - What if the data doesn't conform to this model ?



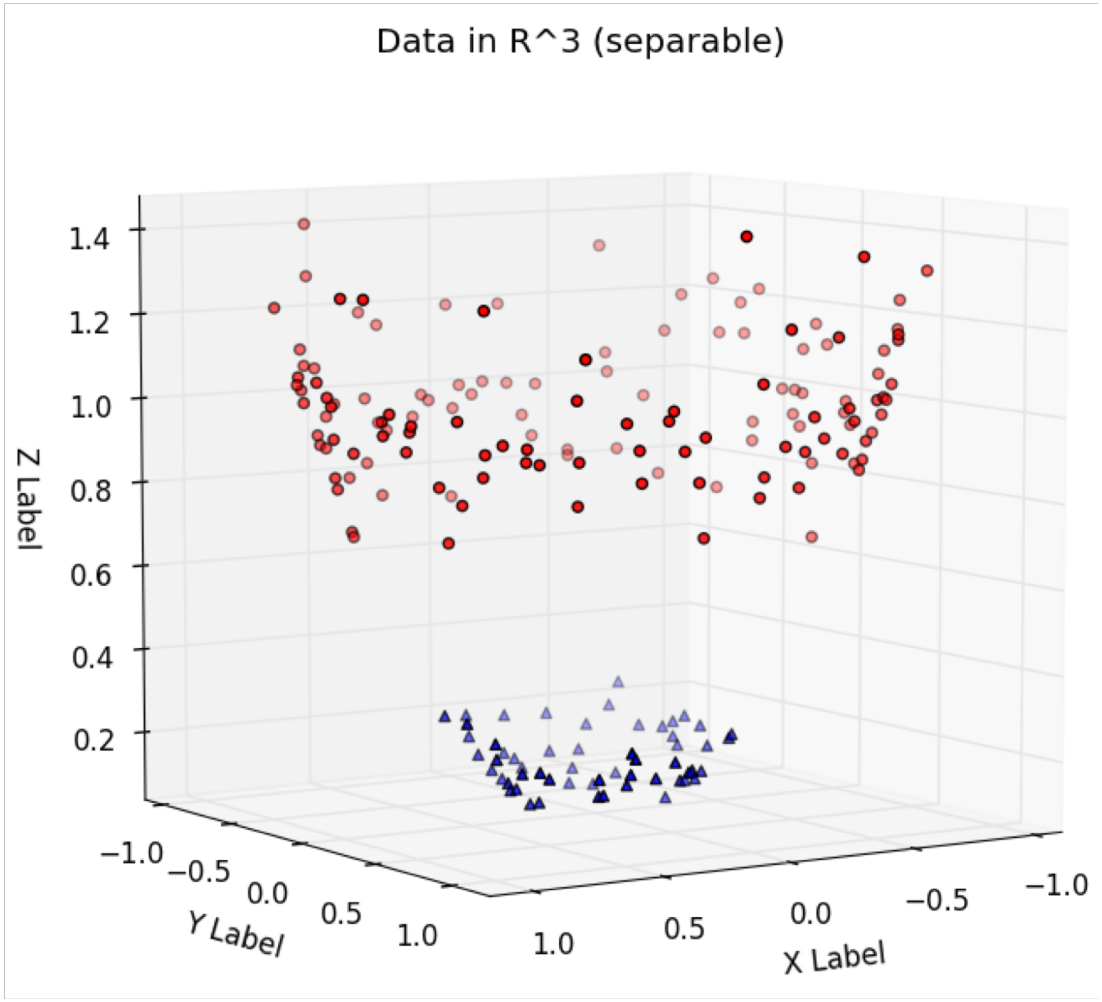
Modeling Complex Data

- Discriminative modeling (for classification)
- Example
  - Classes not linearly separable



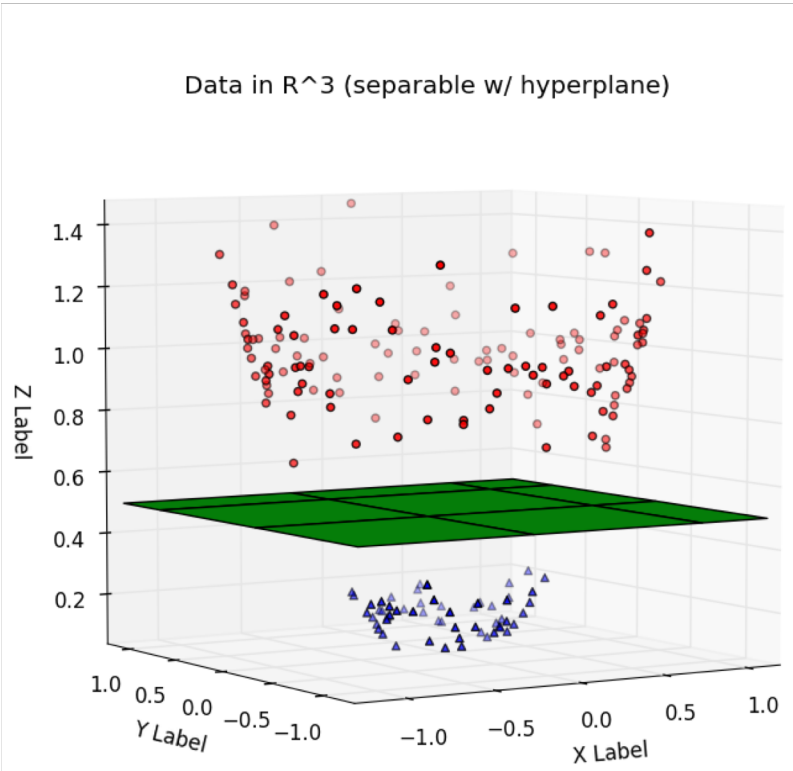
Modeling Complex Data

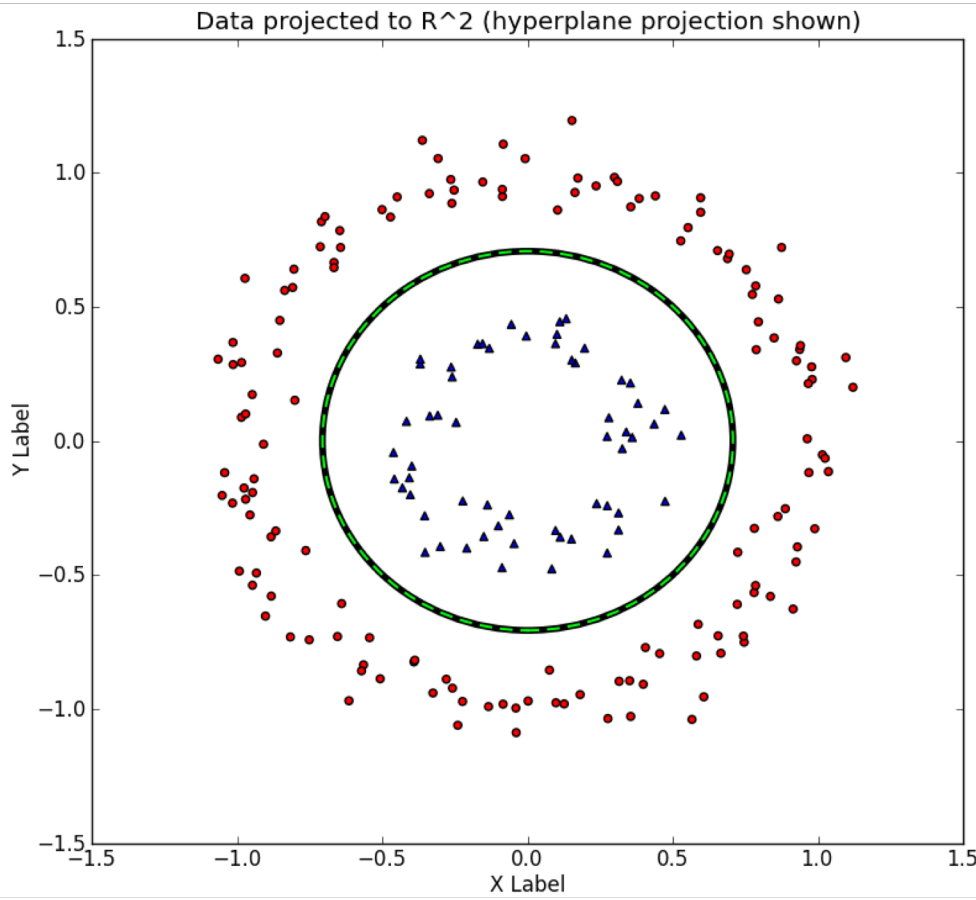
- Discriminative modeling (for classification)
  - Map [x, y] → [x, y, z := $x^2 + y^2$]



Modeling Complex Data

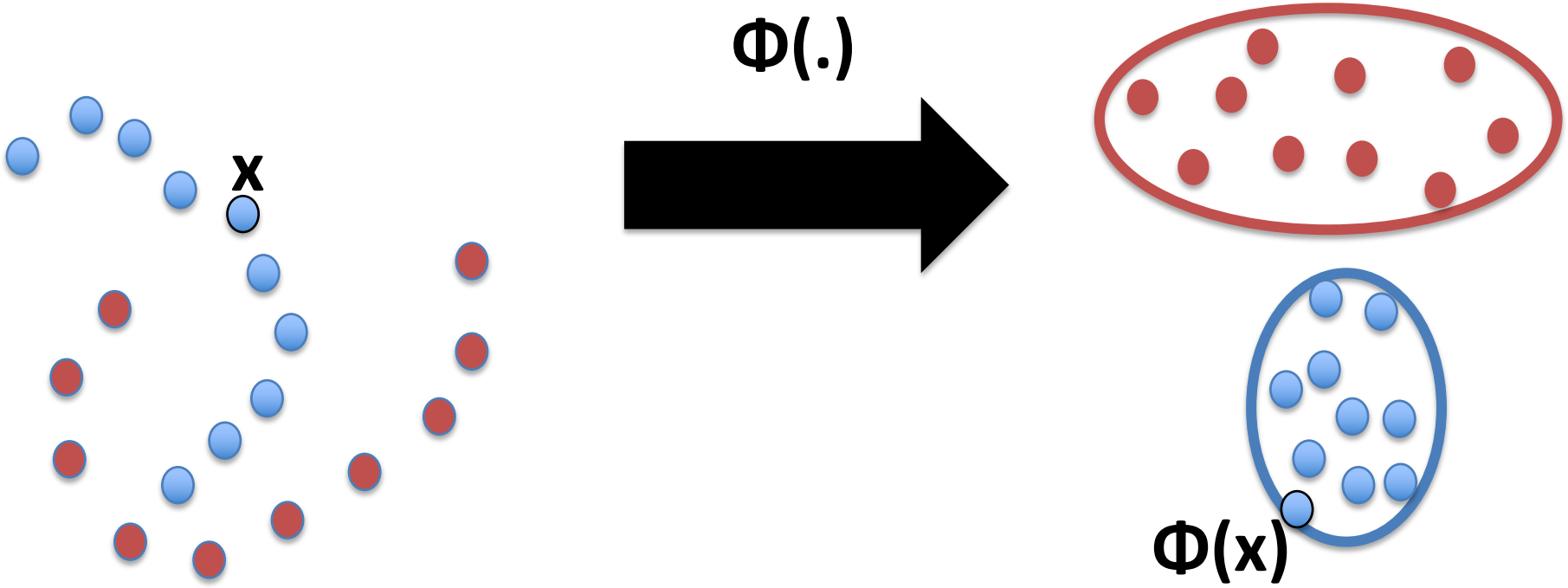
- Map [x, y] → [x, y, z := $x^2 + y^2$]
- In mapped space, data is linearly separable





Modeling Complex Data

- **Generative modeling**
  - Design a (nonlinear) mapping so that data becomes more Gaussian in the mapped space



Kernel Trick

- Mapped space can be very high dimensional
  - Datum = x, Mapped datum = $\Phi(x)$
- Mapping function $\Phi(.)$ can be very complex
- Don't work with mapped data explicitly
- Use a 'kernel' function
  - $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$
  - Kernel gives the inner product in mapped space
  - Don't store $\Phi(x)$ at all
  - Use k(x,y) for all computations
- Identity map $\Phi(.)$ => $k(x,y) = \langle x,y \rangle$

Kernel Trick

- Each mapping function equivalent to a kernel
- In practice, we choose a kernel first !
- Some theorems describe conditions on the kernel for the existence of an equivalent map
  - e.g., k(x,y) = exp (a $||x-y||^2$); a = free parameter
    - Gaussian / Radial basis function (RBF) kernel
  - Mapped space has infinite dimension

$$\exp\left(-\frac{1}{2}||\mathbf{x}-\mathbf{x}'||^2\right) = \sum_{j=0}^{\infty} \frac{(\mathbf{x}^\top \mathbf{x}')^j}{j!} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right) \exp\left(-\frac{1}{2}||\mathbf{x}'||^2\right)$$

    - e.g., for degree j = 2, points x=[x1,x2], y=[y1,y2] in 2D:

$$(\mathbf{x}\cdot\mathbf{y})^2 = (x_1^2, x_1x_2, x_2x_1, x_2^2)(y_1^2, y_1y_2, y_2y_1, y_2^2)^\top$$

Kernel Trick

- Hilbert space
  - Vector space having the structure of an inner product
  - Inner product allows us to measure:
    - Distances / lengths, via the induced norm
    - Angle between two vectors
  - Inner-product function satisfies properties of:
    - Symmetry
    - Linearity in first argument
    - Positive definiteness

Kernel Trick

- **Reproducing kernel**

Definition 2 (Reproducing Kernel). *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \longrightarrow \mathbb{R}$ defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies*

(1) $\forall x \in \mathcal{X}, k_x = k(x, \cdot) \in \mathcal{H}$,

(2) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k_x \rangle_{\mathcal{H}} = f(x)$ *(the reproducing property).*

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}} = \langle k_y, k_x \rangle_{\mathcal{H}} = k(y, x)$.

  - Kernel is *representor of evaluation* (or, *evaluation functional*)

- **Reproducing kernel Hilbert space**

Theorem 3. *A Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space if and only if it has a reproducing kernel.*

Theorem 4. *If it exists, reproducing kernel is unique. Equivalently, a reproducing kernel Hilbert space uniquely determines its reproducing kernel.*

  - Evaluation functionals are bounded, continuous

Kernel Trick

- **Kernel**

Definition 3 (Kernel). *A function* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ *is called a kernel on* $\mathcal{X}$ *is there exists a Hilbert space (not necessarily a reproducing kernel Hilbert space)* $\mathcal{F}$ *and a map* $\mathbf{\Phi} : \mathcal{X} \longrightarrow \mathcal{F}$ *such that* $k(x, y) = \langle \mathbf{\Phi}(x), \mathbf{\Phi}(y) \rangle_{\mathcal{F}}$.

- Corollary:
  Every reproducing kernel is a kernel

Kernel Trick

- **Positive definite function**

Definition 4. *A symmetric function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is positive definite if* $\forall n \in \mathbb{N}$, $\forall (a_1, \ldots, a_n) \in \mathbb{R}^n$ *and* $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

The function $k(\cdot, \cdot)$ *is strictly positive definite if for mutually distinct* x_i*, the equality holds only when all the* a_i *are zero.*

- Corollary:
  Every kernel is a positive definite function

Kernel Trick

- **Moore-Aronszajn Theorem**

Theorem 5 (Moore-Aronszajn). *Let $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be positive definite. Then there is a unique reproducing kernel Hilbert space with reproducing kernel k. In particular, the span of the reproducing kernel k is dense in the resulting reproducing kernel Hilbert space.*

  - Key construction: H is completion of H_0, where

Let the space $\mathcal{H}_0 = \text{span}\{k(x, \cdot) : x \in \mathcal{X}\}$ be endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j),$$

where $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ and $g = \sum_{j=1}^{m} \beta_j k(y_j, \cdot)$

  - Inner product is:
    - Linear: <u+v,w>=<u,w>+<v,w> & <au,v>=a<u,v>, by construction
    - Symmetric: <u,v>=<v,u>, **because k(.,.) is symmetric**
    - Non-degenerate: **because k(.,.) is positive definite;** <u,u> >=0 and <u,u> =0 iff u=0

Kernel Trick, Density Estimation

- Consider mean of mapped data

$$\mathbf{\Phi}_0 = 1/n \sum_{i=1}^{n} \mathbf{\Phi}(\mathbf{x}_i)$$

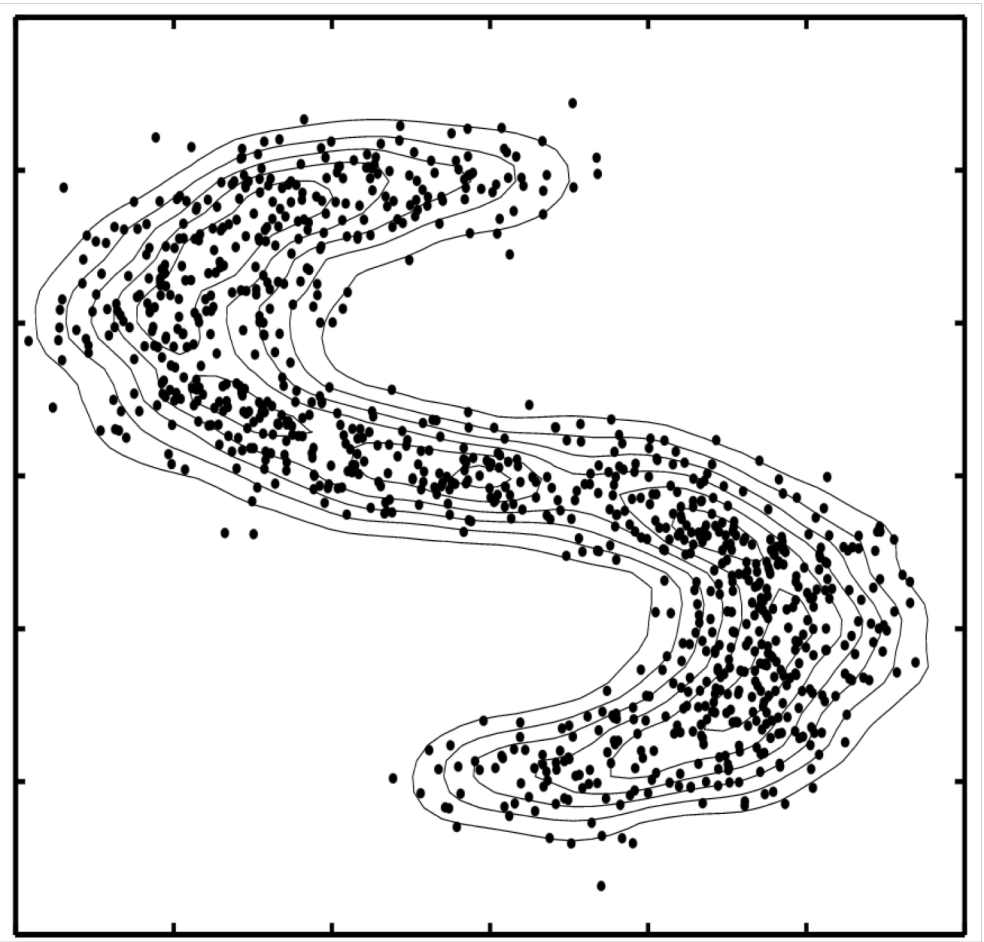
- Distance between mapped datum and mean

$$\|\mathbf{\Phi}(\mathbf{z}) - \mathbf{\Phi}_0\|^2 = k(\mathbf{z}, \mathbf{z}) - \frac{2}{n} \sum_{i=1}^{n} k(\mathbf{z}, \mathbf{x}_i) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j)$$

  - 3rd term = constant ⟶ ignore
  - 1st term = constant for RBF kernel ⟶ ignore
  - 2nd term equivalent to using density estimation using RBF (Gaussian) mixture model
    - Larger probability ⟶ closer to mean

Kernel Trick, Density Estimation

$$\sum_{i=1}^{n} k(\mathbf{z}, \mathbf{x}_i)$$



- How do its contours look like ?
  - Contours adapt to the nonlinear distribution of the data
  - Without kernels (with 1 isotropic Gaussian), contours will be circular, not adapted to data

Principal Component Analysis

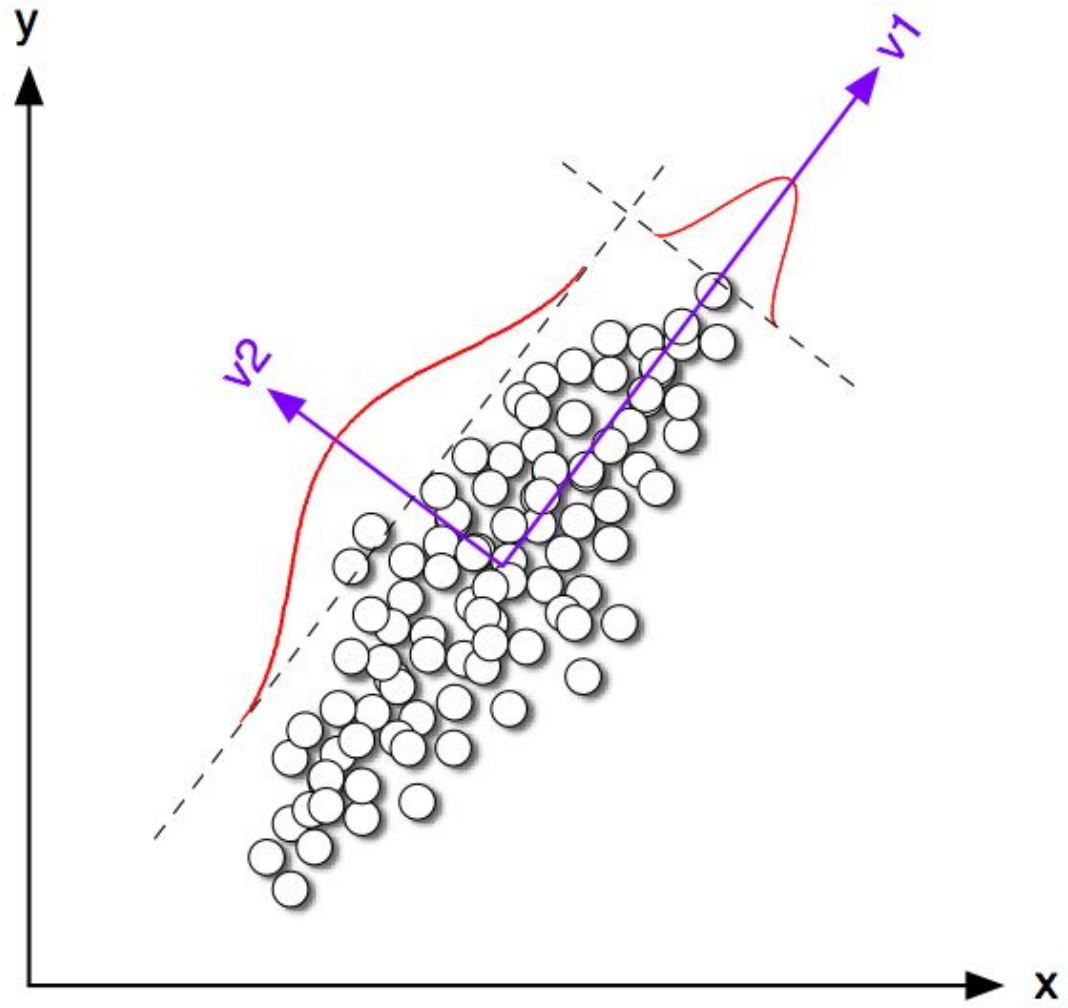
- Data sample = $\{x_j\}$
- Mean estimate = sample mean
- Now, considered centered data
  - Shift in coordinate frame
- Covariance estimate = sample covariance
  - Covariance matrix $C = \frac{1}{M}\sum_{j=1}^{M} \mathbf{x}_j \mathbf{x}_j^\top$
- Eigenvectors of C → directions of variation

$\lambda \mathbf{v} = C\mathbf{v}$ = $\frac{1}{M}\sum_{j=1}^{M}(\mathbf{x}_j \cdot \mathbf{v})\mathbf{x}_j$

  - All eigenvectors must lie in the span of data

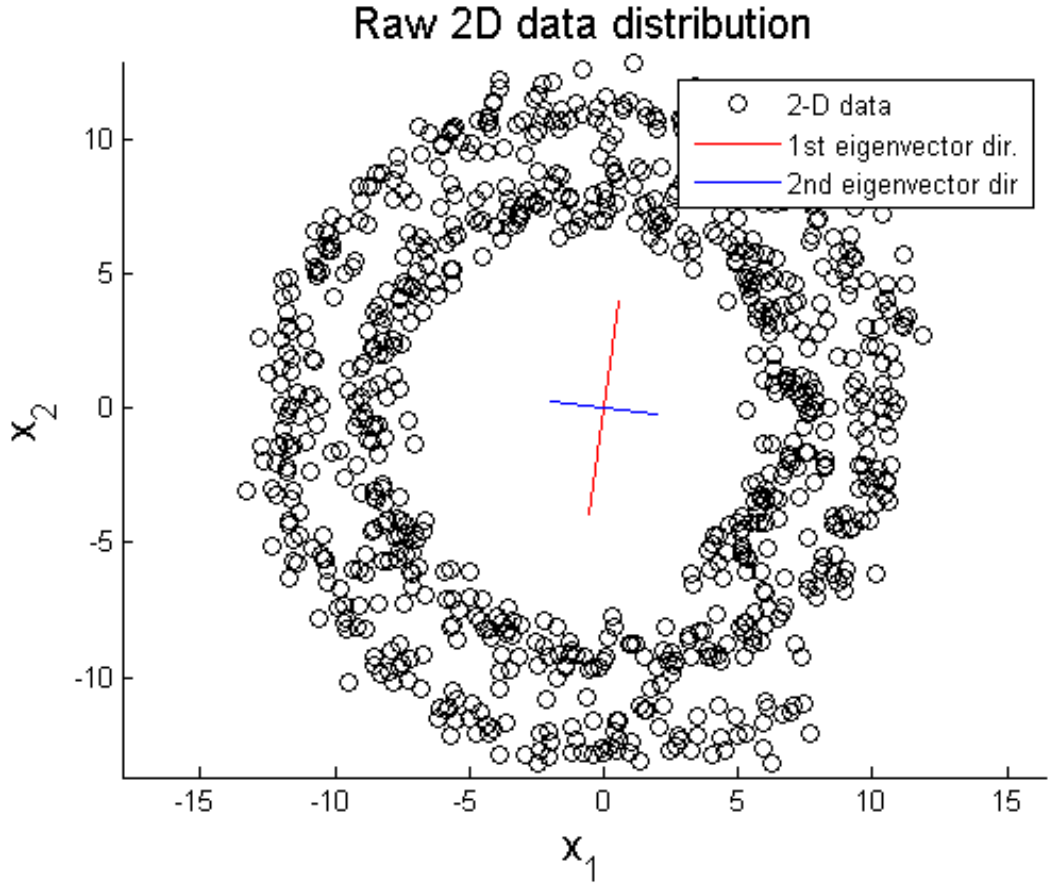
PCA

- PCA for Gaussian distributed data



PCA

- But PCA doesn't work well for a data cluster that isn't ellipsoidal
  - e.g., what if data has the following distribution ?



Kernel PCA

- For centered data, covariance matrix is:

$$\bar{C} = \frac{1}{M} \sum_{j=1}^{M} \Phi(\mathbf{x}_j) \Phi(\mathbf{x}_j)^\top$$

- We can model eigenvector V as $\mathbf{V} = \sum_{i=1}^{M} \alpha_i \Phi(\mathbf{x}_i)$

- For eigenvector V: $\lambda \mathbf{V} = \bar{C} \mathbf{V}$
  - Take inner product with one of the $\Phi(x_k)$

$$\lambda(\Phi(\mathbf{x}_k) \cdot \mathbf{V}) = (\Phi(\mathbf{x}_k) \cdot \bar{C}\mathbf{V})$$

- Substituting forms of V and C in this equation, we get, for each k, …

Kernel PCA

- This can be simplified to $M\lambda K\boldsymbol{\alpha} = K^2\boldsymbol{\alpha}$ where K = Gram matrix for centered data s.t. $K_{ij} := (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)),$
- Solve: $M\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha}$
  - All solutions α satisfy previous equation
  - Additional solutions α won't lead to difference in 'v'
- Eigenvectors of K give vectors α that in turn represent V (need to normalize to unit norm)
- λ = eigenvalues of covariance matrix
  = (1/M) eigenvalues of K

Kernel PCA

- Linear modes in mapped space -> nonlinear modes of variation in input space
  - Call 'feature' = < Φ(x), v >
  - Visualize contours of feature values over domain

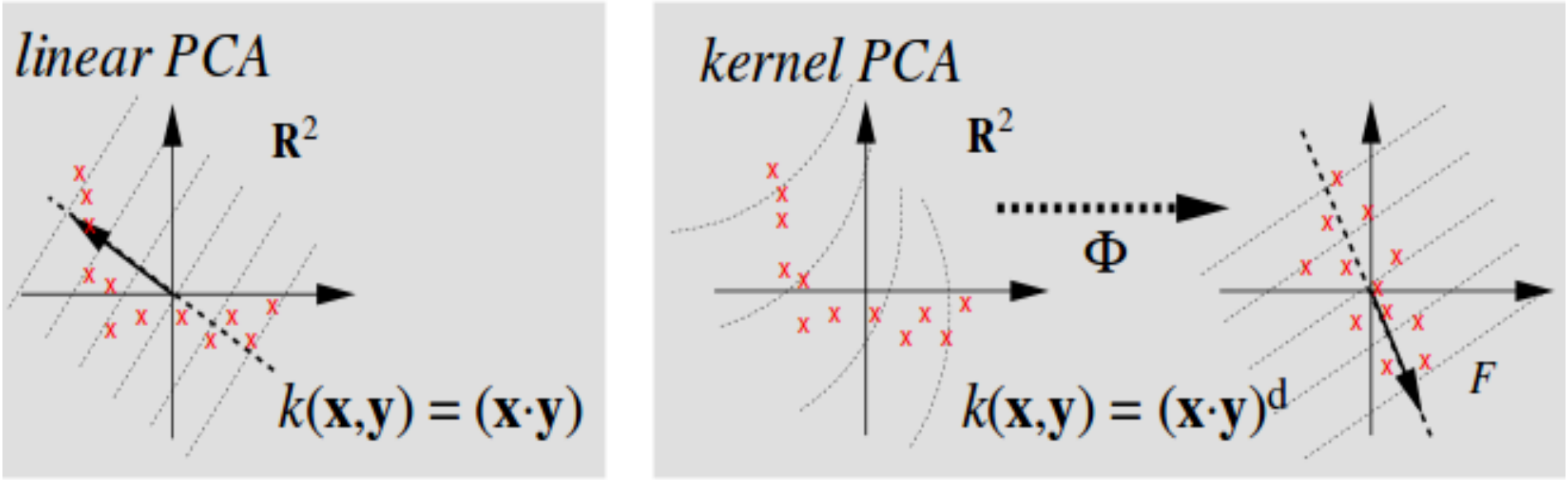


Fig. 1. Basic idea of kernel PCA: by using a nonlinear kernel function k instead of the standard dot product, we implicitly perform PCA in a possibly high–dimensional space F which is nonlinearly related to input space. The dotted lines are contour lines of constant feature value.

KPCA

- Plot contour for constant distance from subspace selected by few principal eigenvectors
  - Can be better than GMM
  - Kernel choice matters

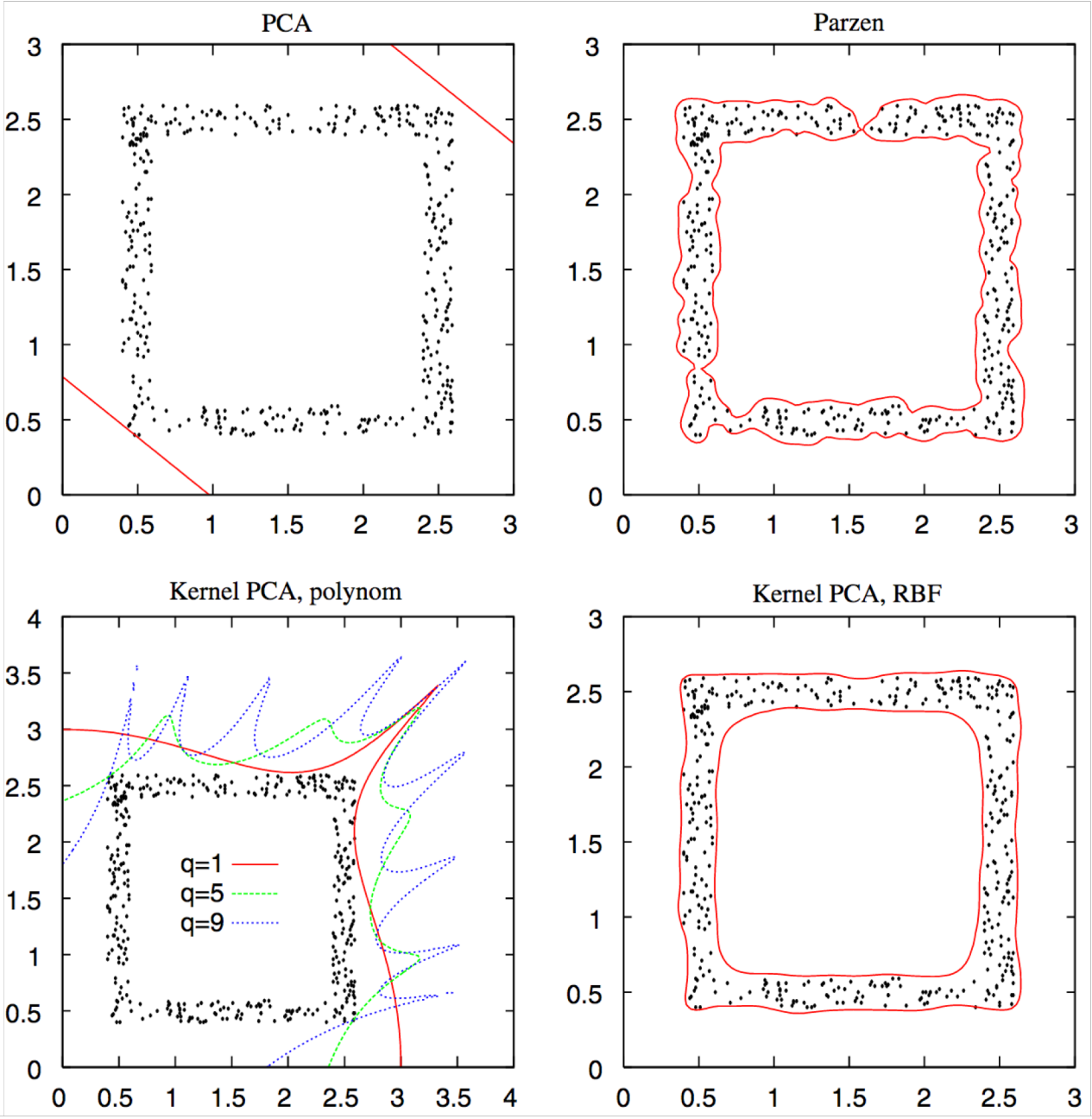
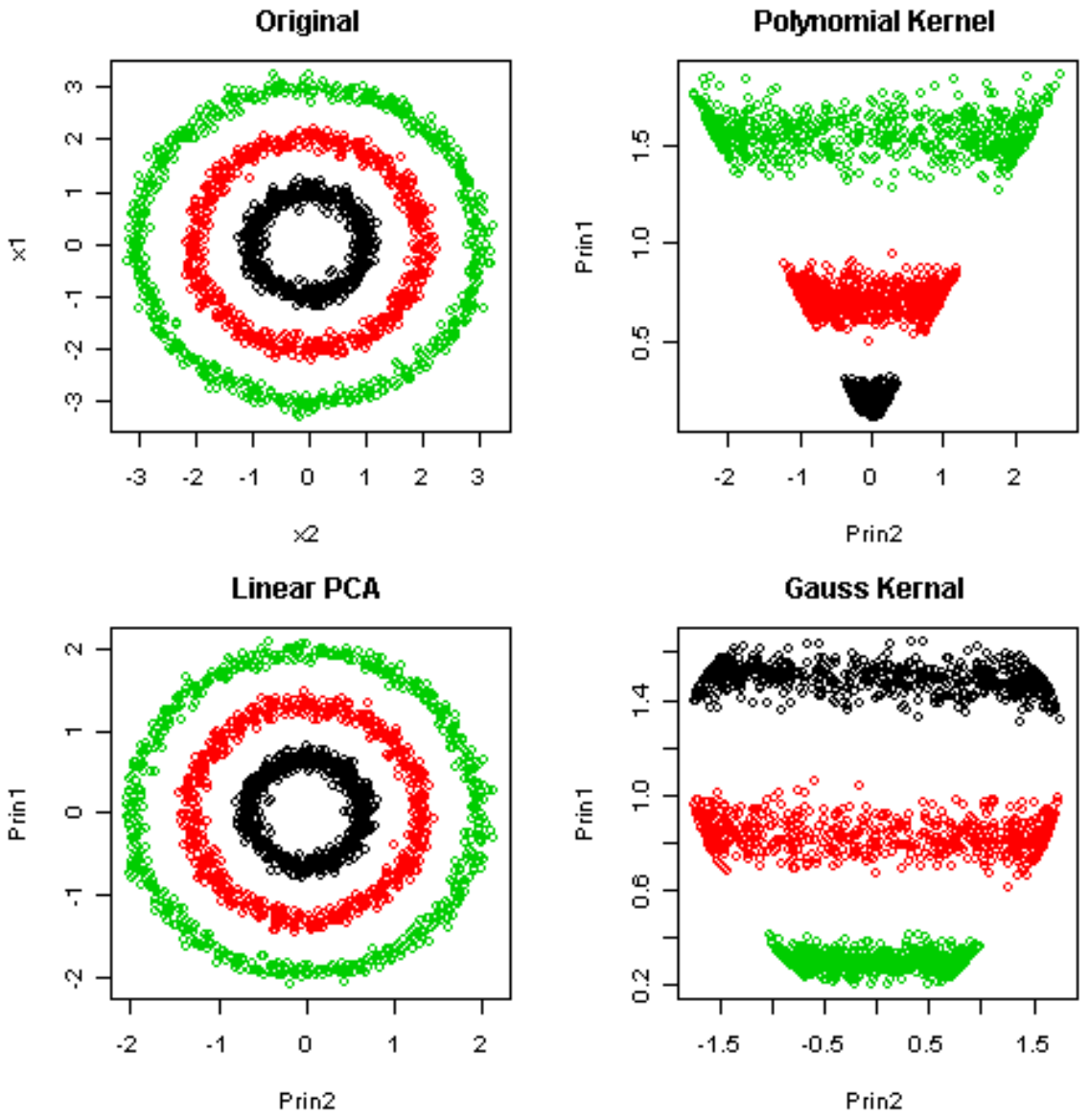


Fig. 2. Decision boundaries for various methods: (top left) PCA reconstruction error with $q = 1$ eigenvector, (top right) Parzen window density estimator with $\sigma = 0.05$, (bottom left) reconstruction error in $\mathcal{F}$ with polynomial kernel $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^{10}$ and various q values, (bottom right) reconstruction error in $\mathcal{F}$ with Gaussian kernel using $\sigma = 0.4$ and $q = 40$.

Kernel PCA

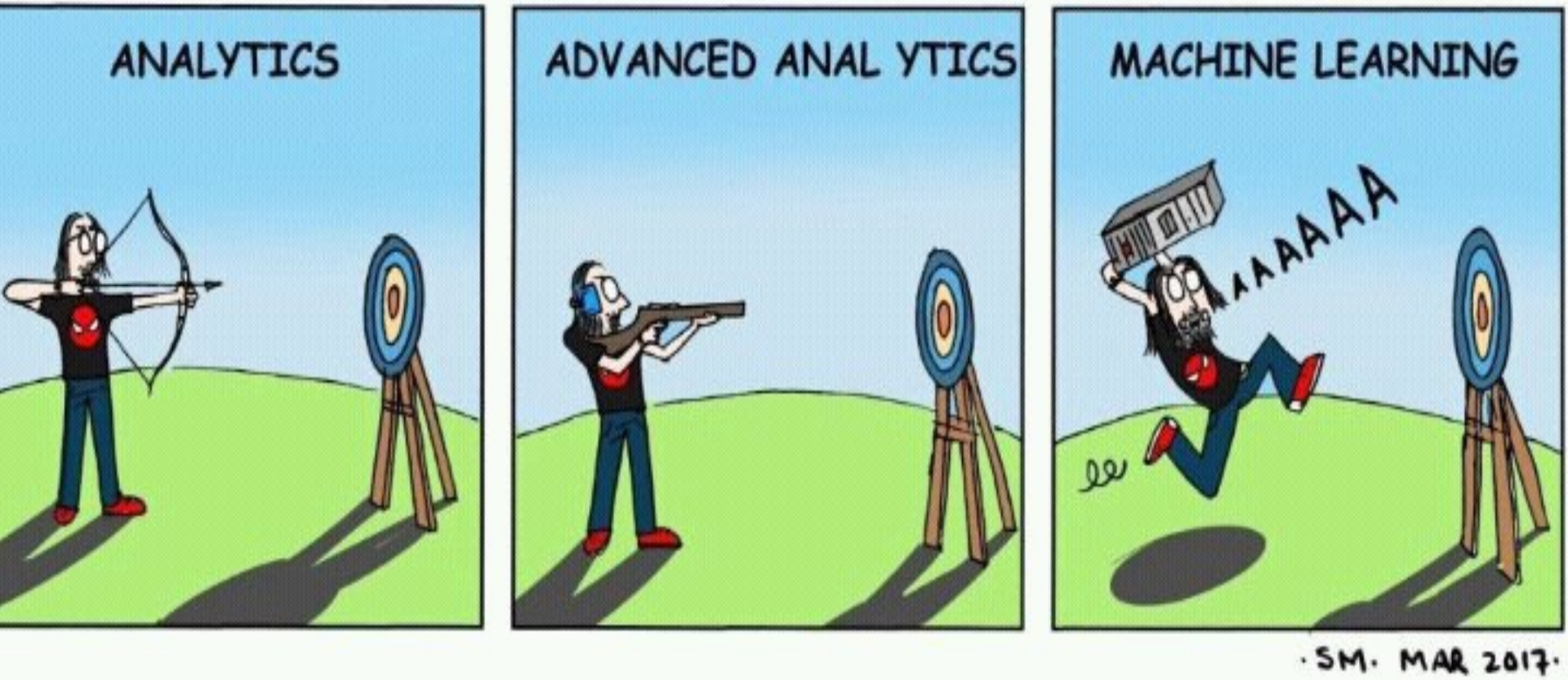
- More examples



- Preferred parameter-tuning method:
  "Grad student descent" =
  Grad student fiddles with parameters until it works
  [Ryan Adams, Harvard]

Data scientist = "Person who is better at statistics than any software engineer and better at software engineering than any statistician." [Josh Wills, Cloudera]

Data scientist = "Person who is worse at statistics than any statistician and worse at software engineering than any software engineer." [Will Cukierski, Kaggle]



"You've been traded for some big data,
two spreadsheets, and an algorithm."