CHAPTER 5:
MULTIVARIATE METHODS

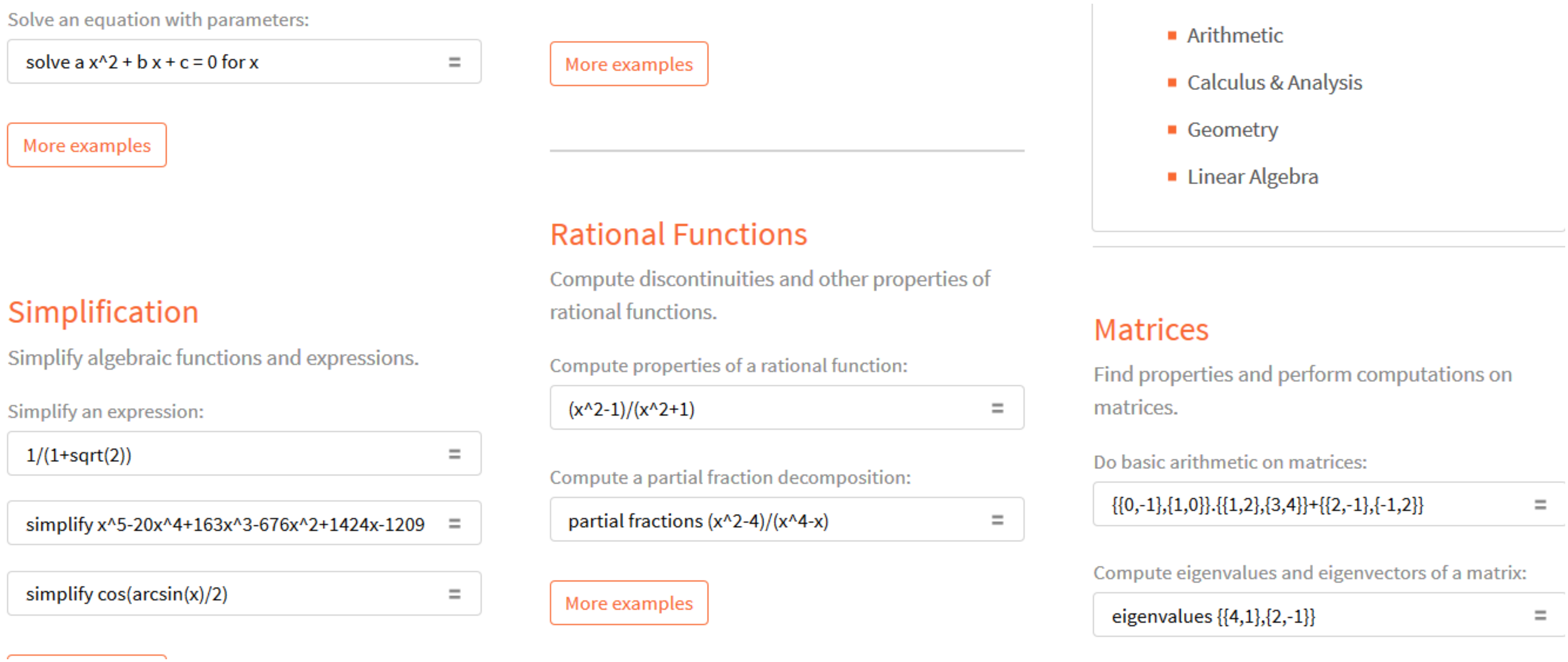
Stella Grasshof

Overview of today

1) reminder and hint
2) multivariate data
   - descriptive statistics
   - classification
   - model selection
3) regression
   - 1D regression revisted
   - multivariate regression
   - model selection

Reminder and Hint

- Pre-warning: many equations today (will be less Th.)
- List of errors in the book (incomplete) https://www.cmpe.boun.edu.tr/~ethem/i2ml3e/
- Helpful online tool: https://www.wolframalpha.com/



Multivariate Data

- Multiple measurements (sensors)
- *D* inputs/features/attributes: *D*-variate
- *N* instances/observations/examples

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_D^1 \\ X_1^2 & X_2^2 & \dots & X_D^2 \\ & & \vdots & \\ X_1^N & X_2^N & \dots & X_D^N \end{pmatrix} \in \mathbb{R}^{N \times D}$$

Multivariate Data

$$\mathbf{X} = \begin{pmatrix} \boxed{X_1^1 \quad X_2^1 \quad \dots \quad X_D^1} \\ X_1^2 \quad X_2^2 \quad \dots \quad X_D^2 \\ \vdots \\ X_1^N \quad X_2^N \quad \dots \quad X_D^N \end{pmatrix} \in \mathbb{R}^{N \times D}$$

data of person 1 of N

Medical Data per person

- Age 0 – 122 years
- Weight 0 – 560 kg
- Blood pressure
  systolic 120 mmHg
  diastolic 80 mmHg
- ...

Multivariate Parameters

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_D^1 \\ X_1^2 & X_2^2 & \dots & X_D^2 \\ & \vdots & & \\ X_1^N & X_2^N & \dots & X_D^N \end{pmatrix} \in \mathbb{R}^{N \times D}$$

- Mean $\quad \mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu} = [\mu_1, \dots, \mu_D]^{\mathrm{T}} \in \mathbb{R}^D$
- Covariance $\quad \sigma_{ij} = \mathrm{Cov}(X_i, X_j) \quad X_i \in \mathbb{R}^N$ column i of matrix $\boldsymbol{X}$

$$\boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2D} \\ & \vdots & & \\ \sigma_{D1} & \sigma_{D2} & \dots & \sigma_D^2 \end{pmatrix} \in \mathbb{R}^{D \times D}$$

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\mathbf{X}) = \mathrm{E}\boxed{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]}$$

Matrix minus vector

- Correlation $\quad \mathrm{Corr}(X_i, X_j) = \rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$

Multivariate Parameters

Adapted Notation

vector of random variables

$$\boldsymbol{x} = (x_1, \dots, x_D)^{\mathrm{T}} \in \mathbb{R}^D$$

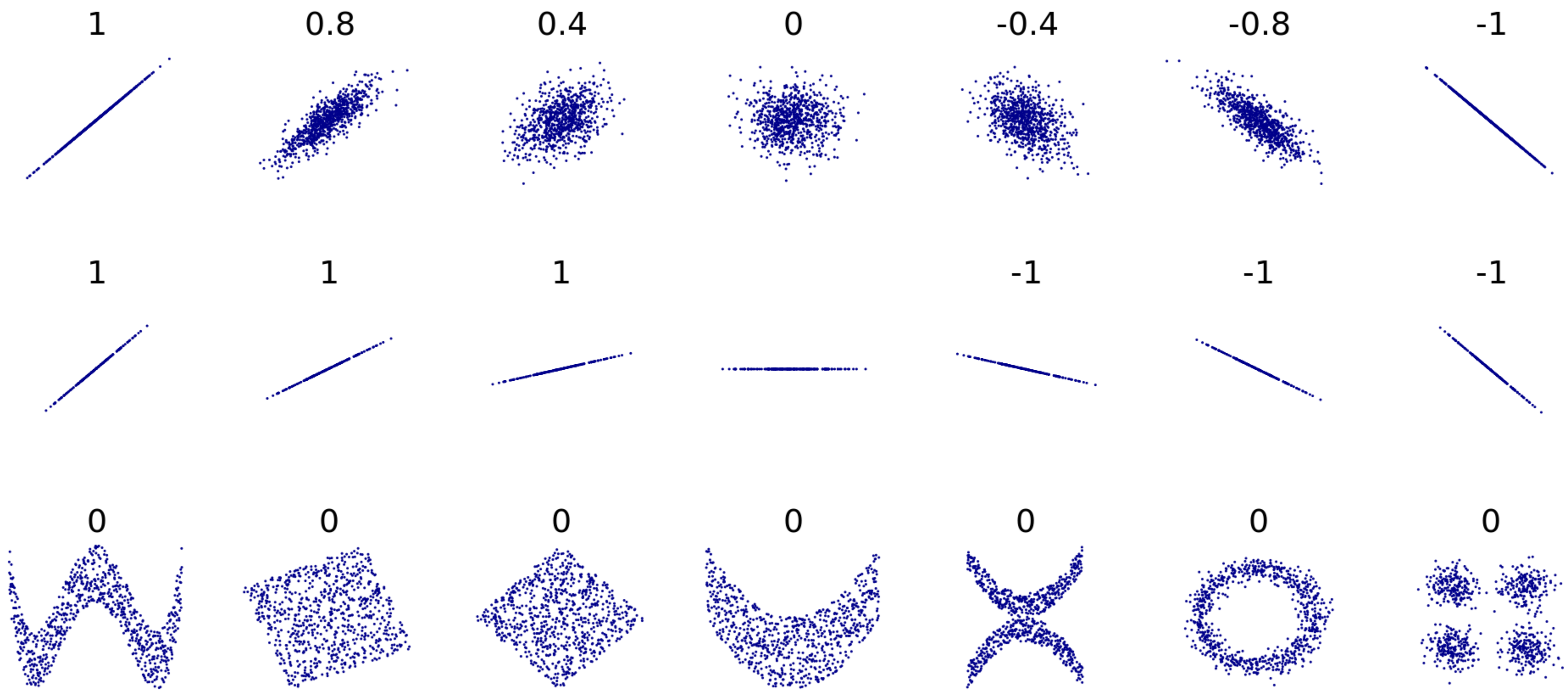
- Mean $$\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu} = [\mu_1, \dots, \mu_D]^{\mathrm{T}} \in \mathbb{R}^D$$
- Covariance $$\sigma_{ij} = \mathrm{Cov}(x_i, x_j)$$

$$\boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2D} \\ & \vdots & & \\ \sigma_{D1} & \sigma_{D2} & \dots & \sigma_D^2 \end{pmatrix} \in \mathbb{R}^{D \times D}$$

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]$$

- Correlation $$\mathrm{Corr}(x_i, x_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \in [-1, 1]$$

Correlation

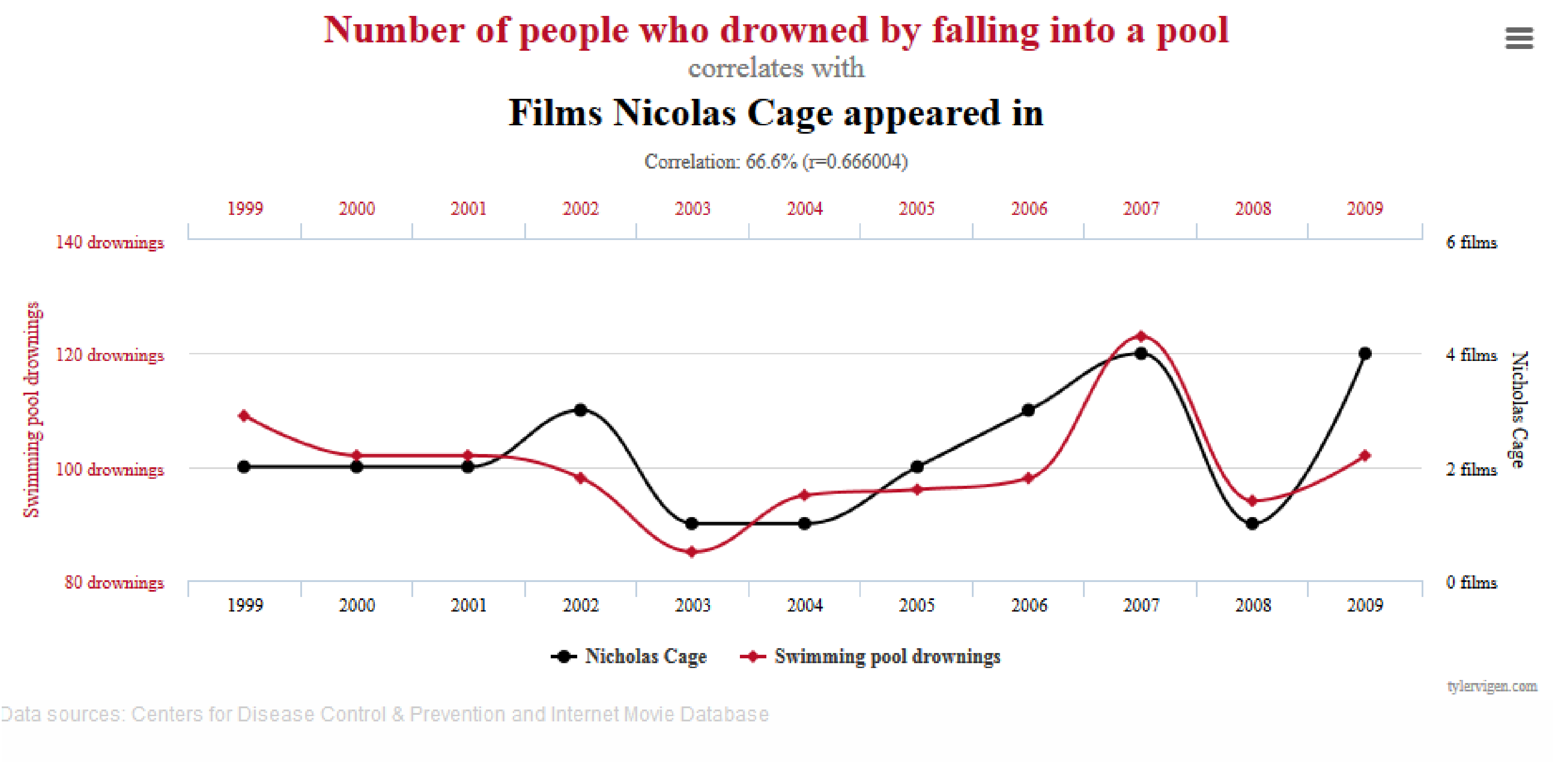


correlation ≠ causality

Image source: https://en.wikipedia.org/wiki/Correlation_and_dependence

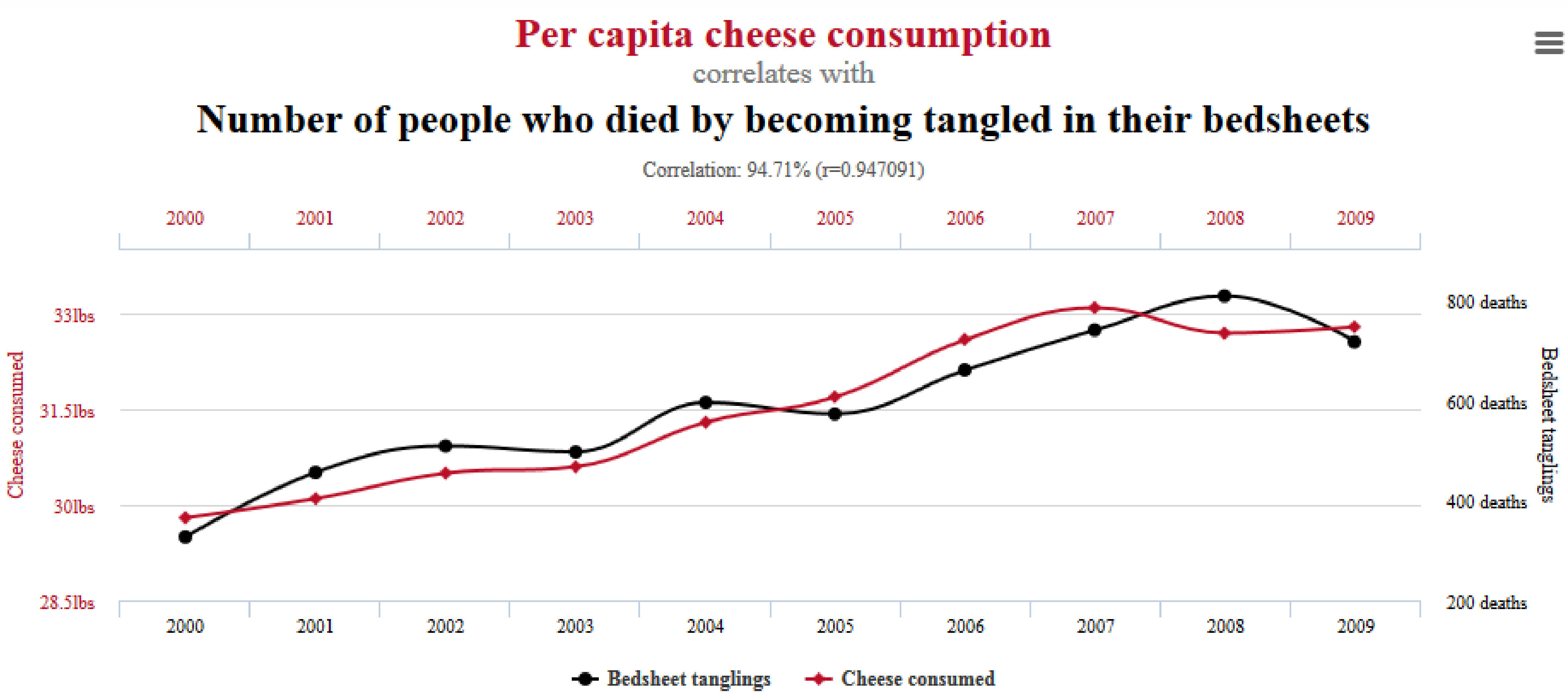
Correlation ≠ Causality

see: https://www.tylervigen.com/spurious-correlations



Correlation ≠ Causality

see: https://www.tylervigen.com/spurious-correlations



Uncorrelated ≠ Independence

- x, y uncorrelated, i. e. $\mathrm{Corr}(x,y) = \rho_{xy} = 0$ if they are not linearly related
- x, y are independent if joint p.d.f. can be factorized as $p(x,y) = p_x(x)p_y(y)$

x, y independent $\overset{\Rightarrow}{\not\Leftarrow}$ x, y uncorrelated

$\Leftarrow$

ONLY IF x, y are normally distributed

$x \sim \mathcal{N}(\mu_x, \sigma_x^2),\ y \sim \mathcal{N}(\mu_y, \sigma_y^2)$

Parameter Estimation from Data

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_D^1 \\ X_1^2 & X_2^2 & \dots & X_D^2 \\ & \vdots & & \\ X_1^N & X_2^N & \dots & X_D^N \end{pmatrix} \in \mathbb{R}^{N \times D}$$

- Sample Mean $$\boldsymbol{\mu} \approx \widehat{\boldsymbol{\mu}} = \boldsymbol{m} = \frac{1}{N} \left(\sum_{n=1}^{N} X_1^n, \dots, \sum_{n=1}^{N} X_D^n \right)^{\mathrm{T}} \in \mathbb{R}^D$$
- Covariance $$\sigma_{ij} \approx \widehat{\sigma}_{ij} = s_{ij} = \frac{1}{N} \sum_{n=1}^{N} (X_i^n - m_i)(X_j^n - m_j)$$

$$s_i^2 = \frac{1}{N} \sum_{n=1}^{N} (X_i^n - m_i)^2$$

- Correlation $$\rho_{ij} \approx \widehat{\rho}_{ij} = r_{ij} = \frac{s_{ij}}{s_i s_j}$$

Parameter Estimation from Data

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_D^1 \\ X_1^2 & X_2^2 & \dots & X_D^2 \\ & \vdots & & \\ X_1^N & X_2^N & \dots & X_D^N \end{pmatrix} \in \mathbb{R}^{N \times D}$$

- Sample Mean $\boldsymbol{\mu} \approx \widehat{\boldsymbol{\mu}} = \boldsymbol{m} = \frac{1}{N} \left(\sum_{n=1}^{N} X_1^n, \dots, \sum_{n=1}^{N} X_D^n \right)^{\mathrm{T}} \in \mathbb{R}^D$
- Covariance $\sigma_{ij} \approx \widehat{\sigma}_{ij} = s_{ij} = \frac{1}{N-1} \sum_{n=1}^{N} (X_i^n - m_i)(X_j^n - m_j)$

unbiased estimator $s_i^2 = \frac{1}{N-1} \sum_{n=1}^{N} (X_i^n - m_i)^2$

- Correlation $\rho_{ij} \approx \widehat{\rho}_{ij} = r_{ij} = \frac{s_{ij}}{s_i s_j}$

Missing Values ?

What to do if certain instances have missing attributes?

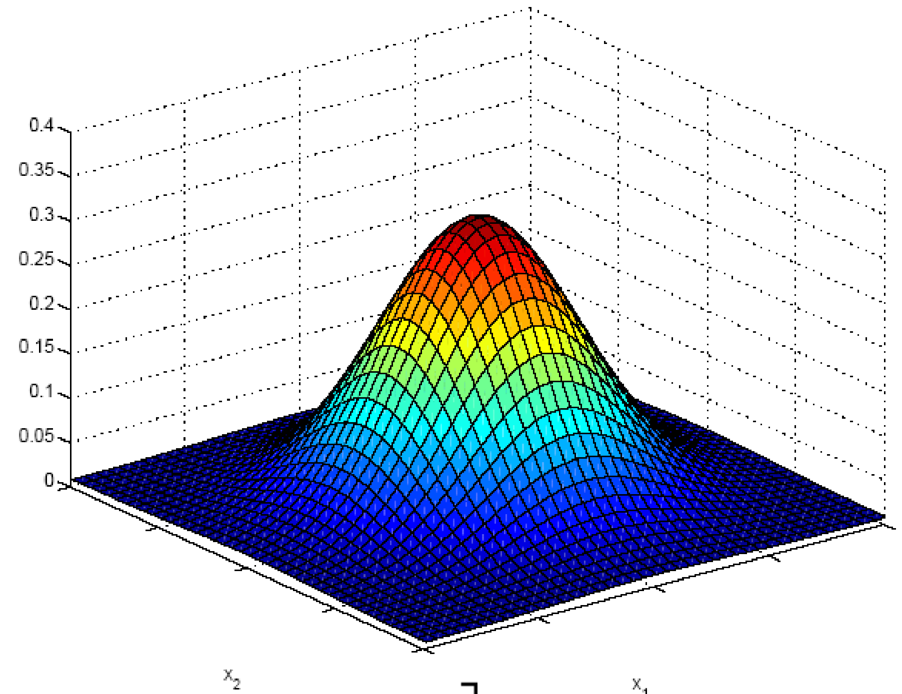
- **Ignore** those instances:
  not a good idea if the sample is small
- **Use 'missing' as an attribute:**
  may give information
- **Imputation**: Fill in the missing value
  - Mean imputation: Use the most likely value (e.g., mean)
  - Imputation by regression:
    Predict based on other attributes

Multivariate Normal Distribution

1-dimensional: $x \in \mathbb{R}$

$x \sim \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$



D-dimensional: $\boldsymbol{x} \in \mathbb{R}^D$

$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Multivariate Normal Distribution

- Mahalanobis distance: $(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$

  measures the distance from ***x*** to ***μ*** in terms of ∑

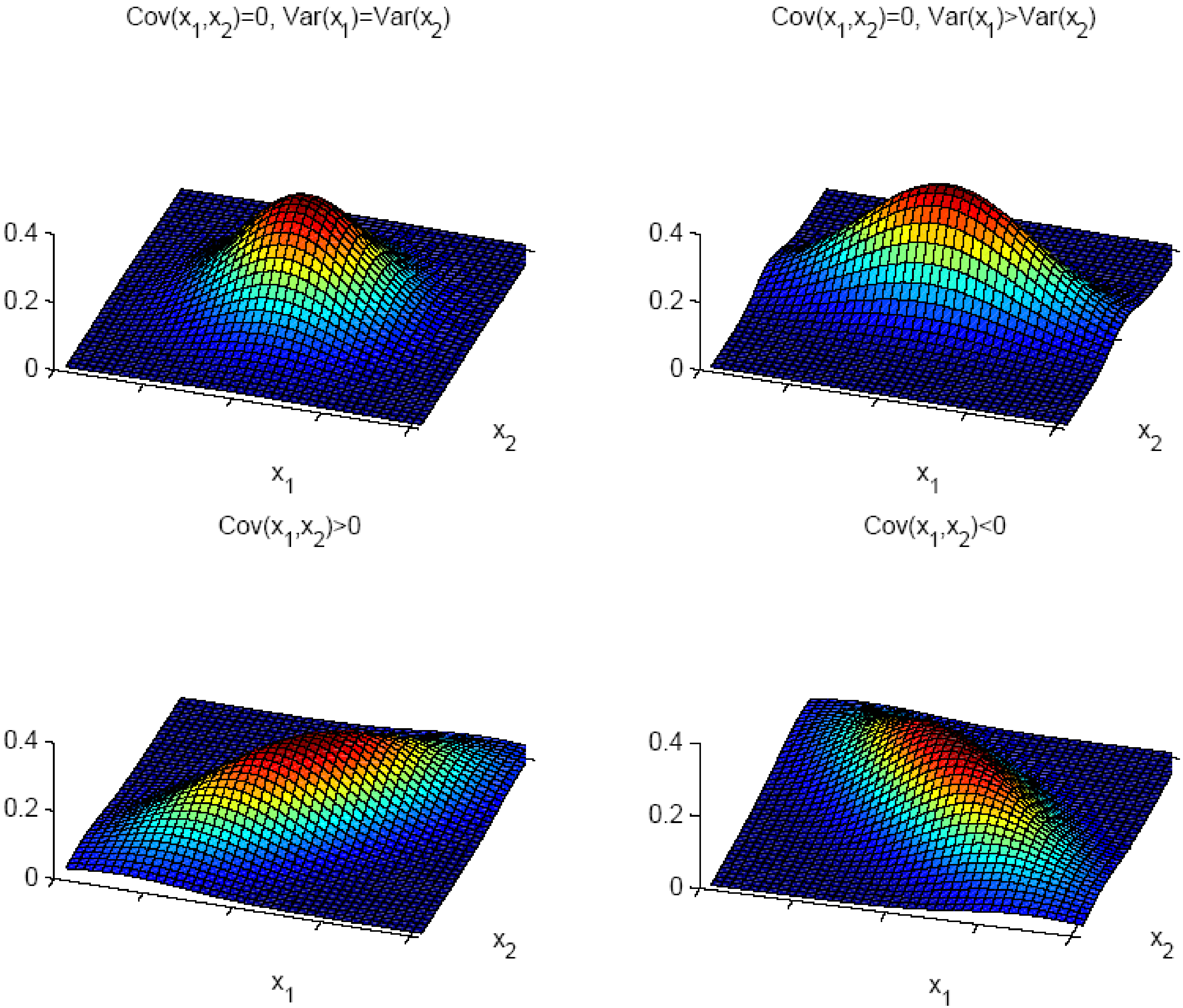
  normalizes for difference in variances and correlations

- Bivariate: $D = 2$

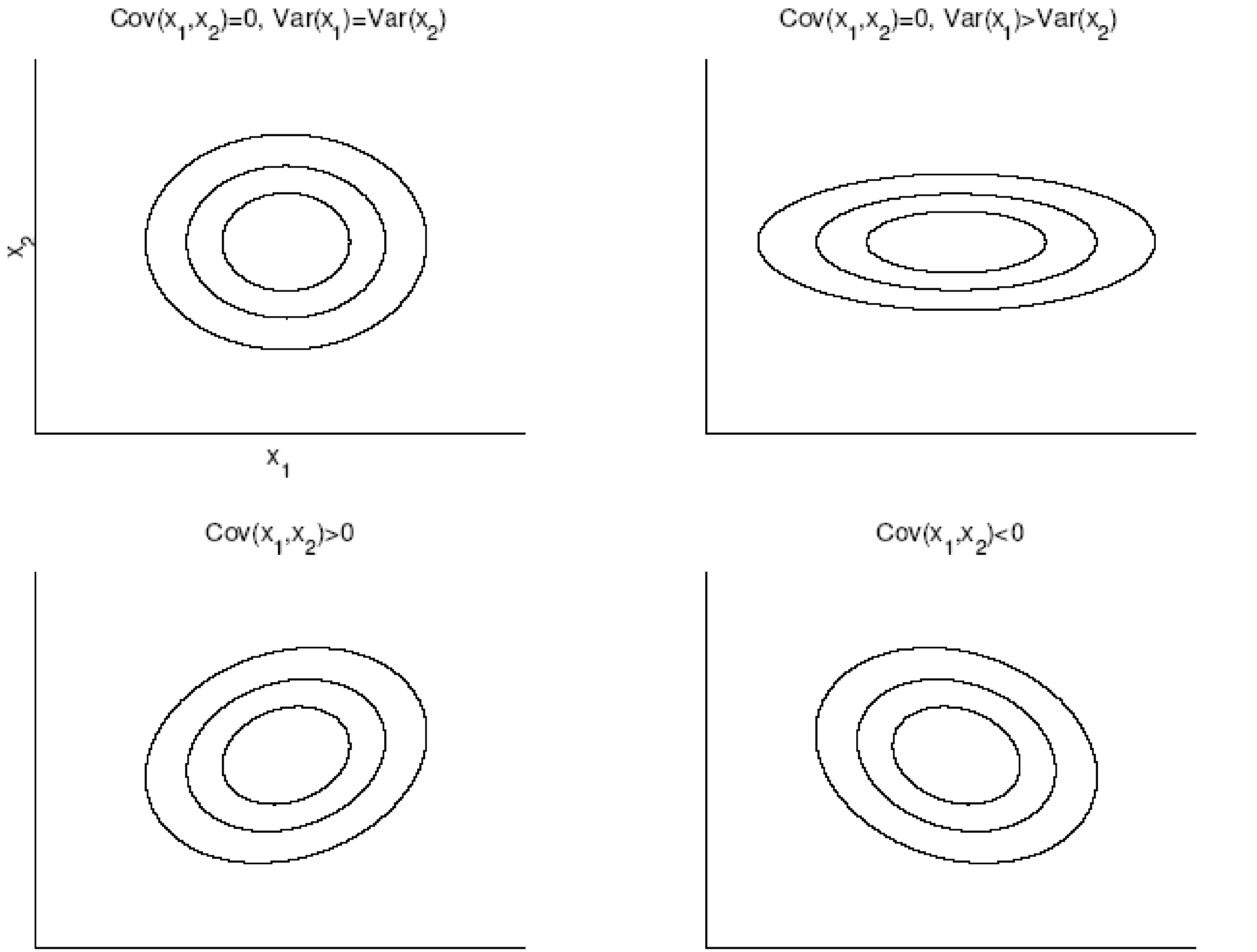
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(z_1 - 2\rho z_1 z_2 + z_2^2)\right]$$

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$


Cov(x1,x2)=0, Var(x1)=Var(x2)
Cov(x1,x2)=0, Var(x1)>Var(x2)
Cov(x1,x2)>0
Cov(x1,x2)<0
0.4
0.2
0
x1
x2

Bivariate Normal



Independent Inputs: Naive Bayes

- If x_i are independent
  - x_i are uncorrelated
  - offdiagonals of $\sum$ are 0
  - Mahalanobis distance reduces to weighted (by $1/\sigma_d$) Euclidean distance:

$$p(\boldsymbol{x}) = \prod_{d=1}^{D} p_d(x_d) = \frac{1}{(2\pi)^{D/2} \prod_{d=1}^{D} \sigma_d} \exp\left[-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_d - \mu_d}{\sigma_d}\right)^2\right]$$

- If variances are also equal => Euclidean distance

Revisit: Parametric Classification Ch 4.5

- Suppose classes $C_i, \quad i = 1, \ldots, K$
- Discriminative function for each class $g_i(\boldsymbol{x})$
- For sample x choose class k:

$$g_k(\boldsymbol{x}) = \max_i g_i(\boldsymbol{x})$$

- How to define discriminate function $g_i(\boldsymbol{x})$?
  - „Dumb“: class probabilty $P(C_i)$
  - Maximum Likelihood (ML) $p(\boldsymbol{x}|C_i)$
  - $\boxed{\text{Maximum a-posteriori (MAP) } p(\boldsymbol{x}|C_i)P(C_i)}$
  - Bayes $-R(\alpha_i|\boldsymbol{x})$

$$\log(p(\boldsymbol{x}|C_i)P(C_i)) = \log(p(\boldsymbol{x}|C_i)) + \log(P(C_i))$$

Parametric Classification

- If $(\boldsymbol{x}|C_i) \sim \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\Sigma}_i)$

$$p(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

- Discriminant functions (see Ch. 3.4)

$$\begin{aligned} g_i(\boldsymbol{x}) &= \log p(\boldsymbol{x}|C_i) + \log P(C_i) \\ &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) \\ &\quad + \log P(C_i) \end{aligned}$$

Estimation of Parameters

$$\mathbb{1}\left\{\boldsymbol{x}_k \in C_i\right\} = \begin{cases} 1 & , \ \boldsymbol{x}_k \in C_i \\ 0 & , \ \text{else} \end{cases}, \qquad N_i := \sum_{k=1}^{N} \mathbb{1}\left\{\boldsymbol{x}_k \in C_i\right\}$$

$$\widehat{P}(C_i) = \frac{N_i}{N}$$

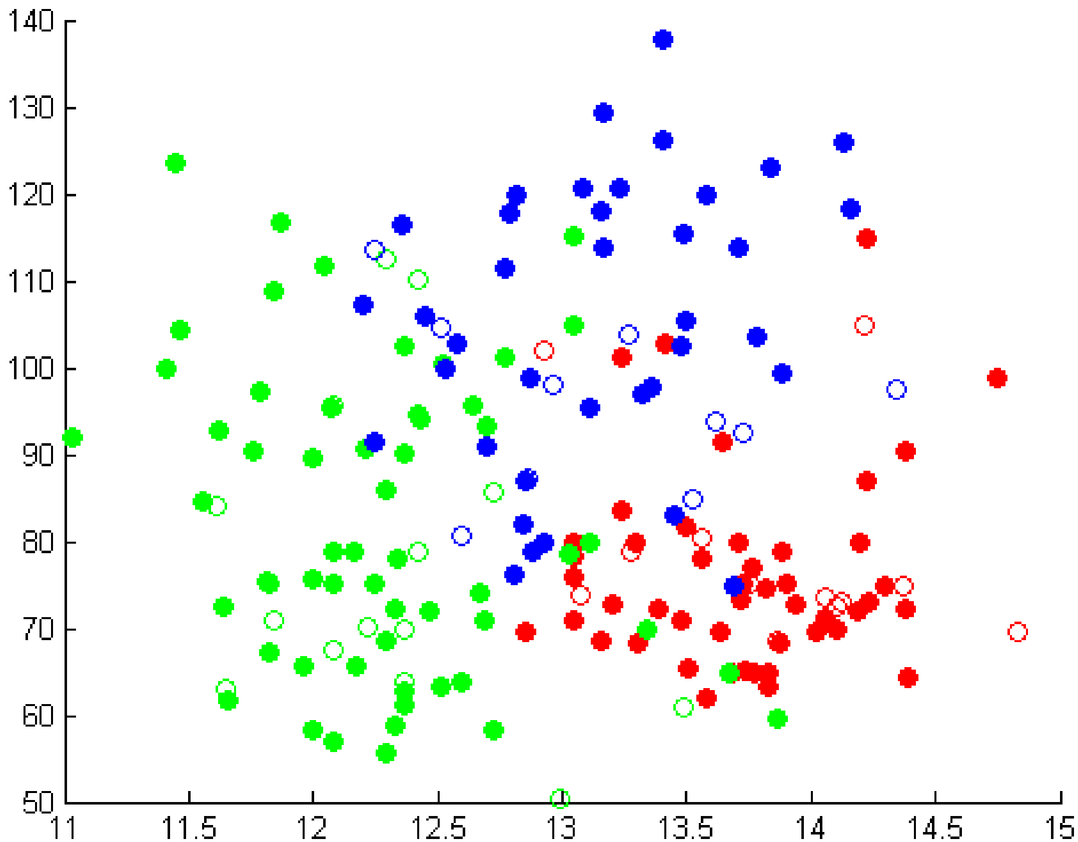
$$\boldsymbol{m}_i = \frac{1}{N_i} \sum_{k=1} \mathbb{1}\left\{\boldsymbol{x}_k \in C_i\right\} \boldsymbol{x}_k$$

$$\boldsymbol{S}_i = \frac{1}{N_i} \sum_{k=1}^{N} \mathbb{1}\left\{\boldsymbol{x}_k \in C_i\right\} (\boldsymbol{x}_k - \boldsymbol{m}_i)(\boldsymbol{x}_k - \boldsymbol{m}_i)^{\mathrm{T}}$$

$$g_i(\boldsymbol{x}) = -\frac{1}{2} \log |\boldsymbol{S}_i| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{m}_i)^{\mathrm{T}} \boldsymbol{S}_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i) + \log \widehat{P}(C_i)$$

Parametric Classification

- 3 classes
- Training and Test Data



Case 1: Different $\mathbf{S}_i$

Quadratic discriminant

$$\begin{aligned} g_i(\boldsymbol{x}) &= -\frac{1}{2}\log|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i) + \log\widehat{P}(C_i) \\ &= -\frac{1}{2}\log|\boldsymbol{S}_i| - \frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{m}_i + \boldsymbol{m}_i^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{m}_i\right) \\ &\quad + \log\widehat{P}(C_i) \end{aligned}$$

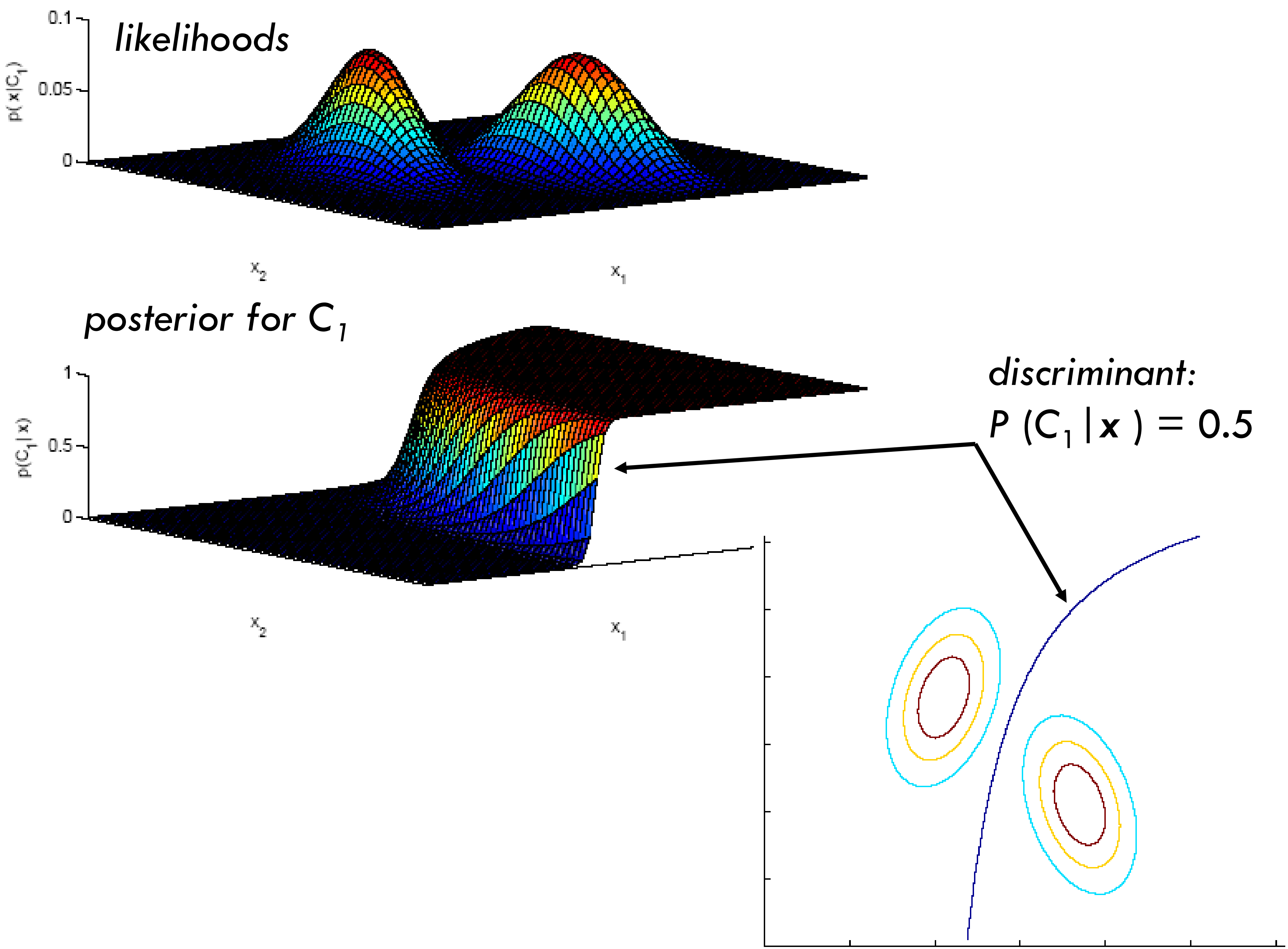
$$g_i(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{w}_{i0}$$

where

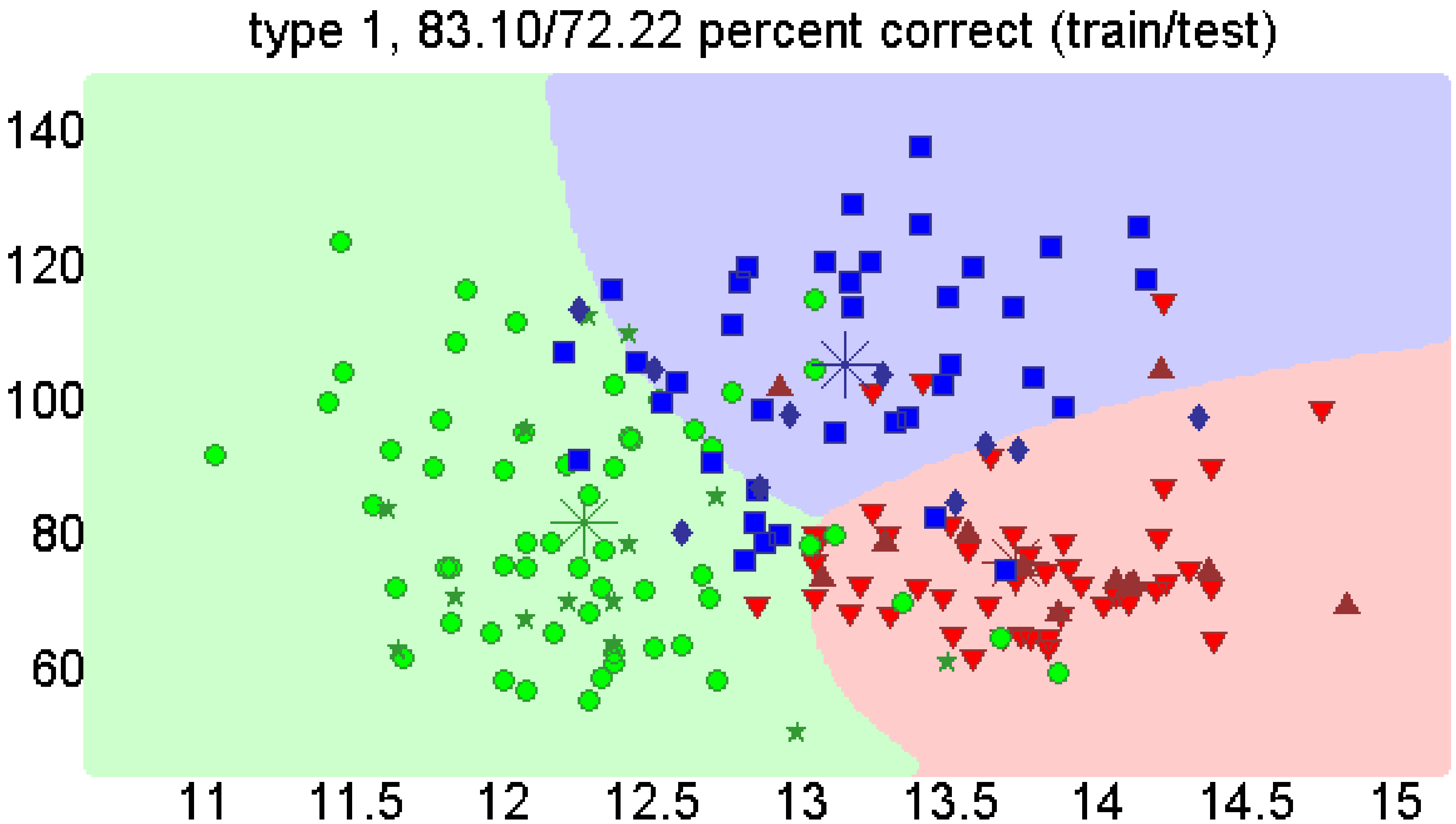
$$\boldsymbol{W}_i = -\frac{1}{2}\boldsymbol{S}_i^{-1}$$

$$\boldsymbol{w}_i = \boldsymbol{S}_i^{-1}\boldsymbol{m}_i$$

$$\boldsymbol{w}_{i0} = -\frac{1}{2}\boldsymbol{m}_i^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{m}_i - \frac{1}{2}\log|\boldsymbol{S}_i| + \log\widehat{P}(C_i)$$


likelihoods
0.1
0.05
0
p(x|C1)
x2
x1
posterior for C1
1
0.5
0
p(C1|x)
x2
x1
discriminant:
P (C1|x) = 0.5

Case 1: Different S_i



Case 2: Common Covariance Matrix **S**

- Shared common sample covariance **S**

$$\boldsymbol{S} = \sum_i \widehat{P}(C_i)\boldsymbol{S}_i$$

- Discriminant reduces to

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) + \log \widehat{P}(C_i)$$

which is a linear discriminant

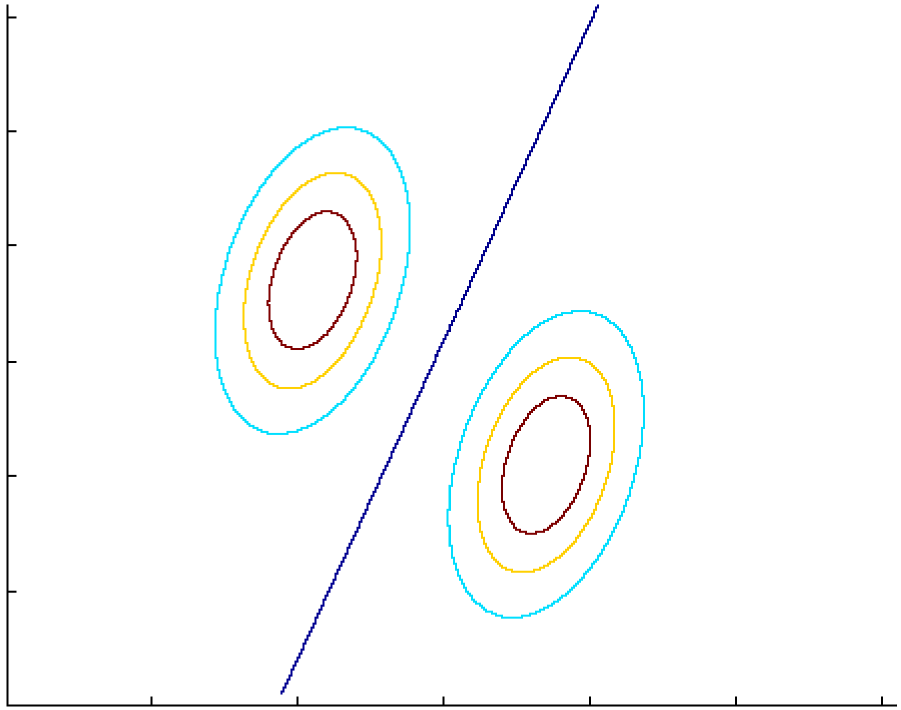
$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{w}_{i0}$$

where

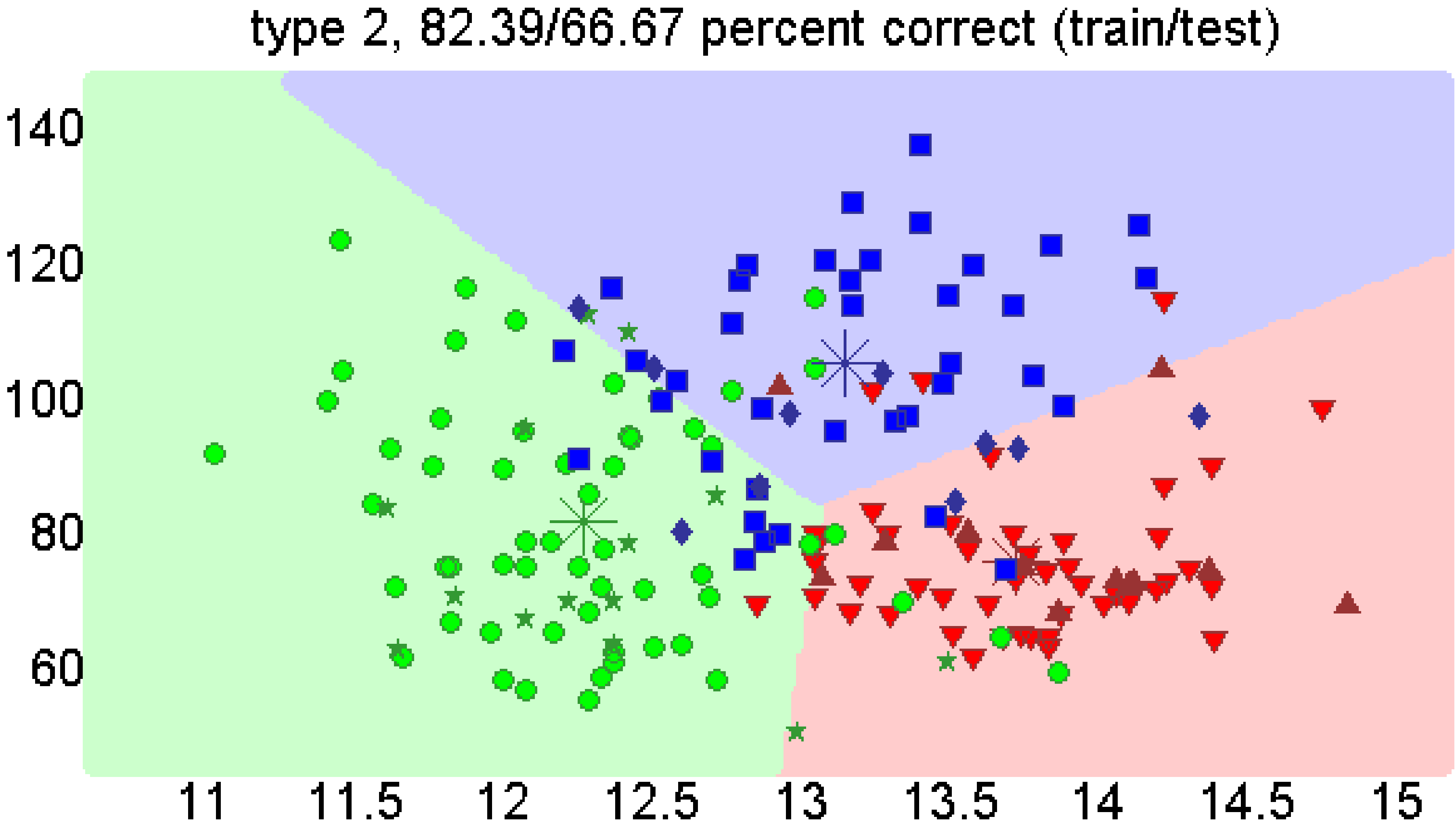
$$\boldsymbol{w}_i = \boldsymbol{S}^{-1}\boldsymbol{m}_i$$

$$\boldsymbol{w}_{i0} = -\frac{1}{2}\boldsymbol{m}_i^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{m}_i + \log \widehat{P}(C_i)$$

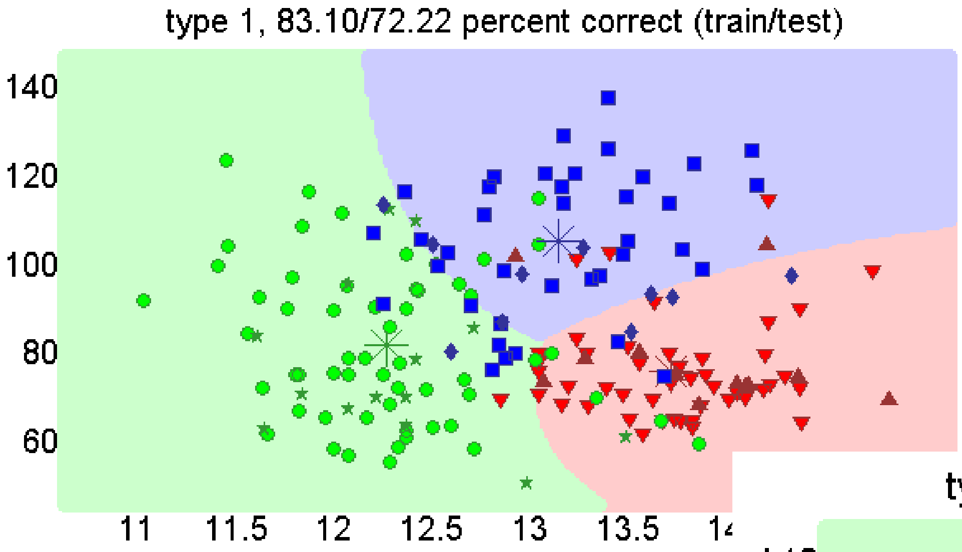
Case 2: Common Covariance Matrix **S**

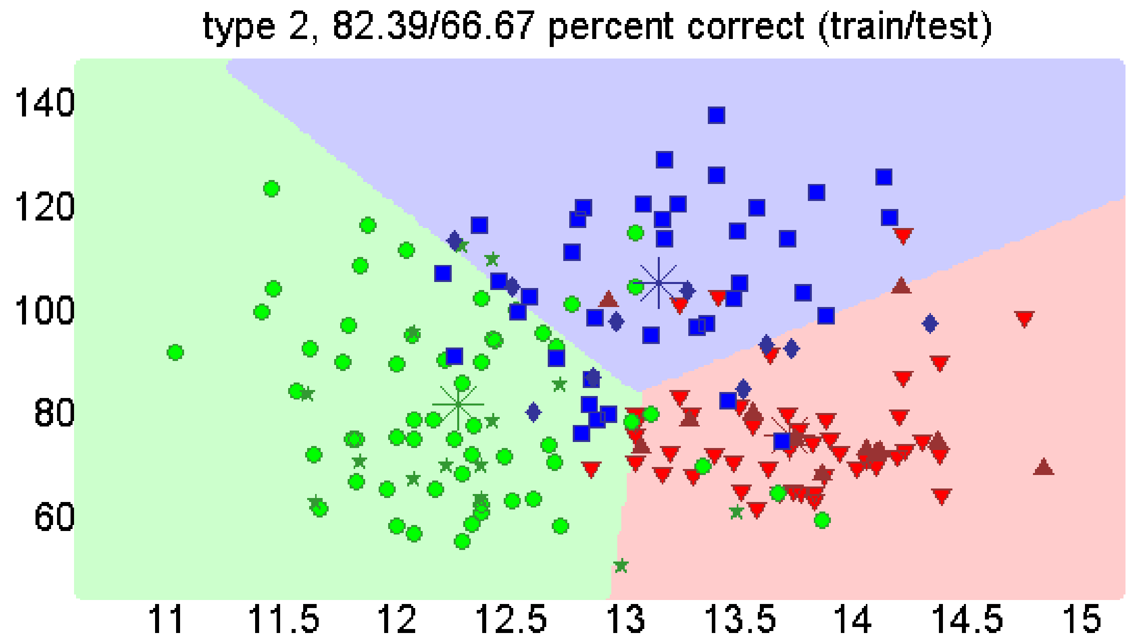


Case 2: Common Covariance Matrix S



Case 2: Common Covariance Matrix S





Case 3: Diagonal **S**

If x_d d =1,..., D are independent:

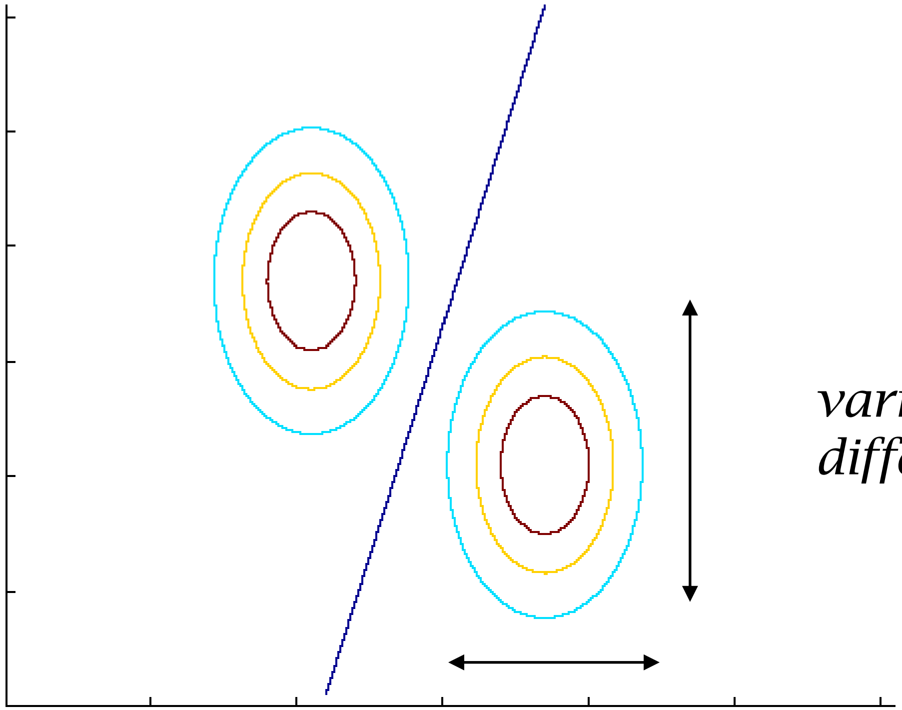
- $p(\boldsymbol{x}|C_i) = \prod_{d=1}^{D} p(x_d|C_i)$ (Naive Bayes' assumption)
- Covariance matrix is diagonal

$$g_i(\boldsymbol{x}) = -\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_d - m_{id}}{s_d}\right)^2 + \log \widehat{P}(C_i)$$

Classify based on

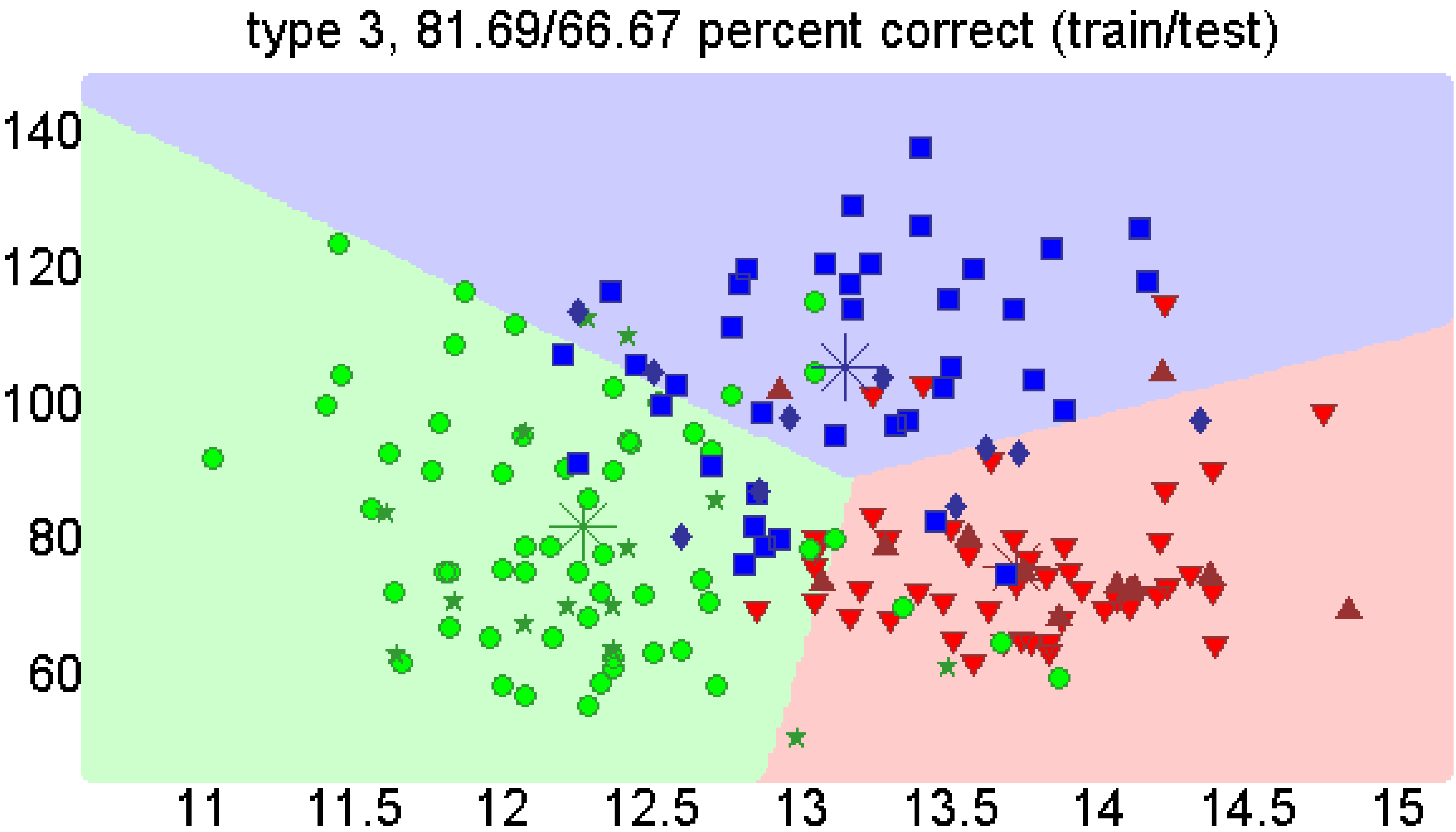
weighted Euclidean distance (in s_d units) to the nearest mean

Case 3: Diagonal **S**

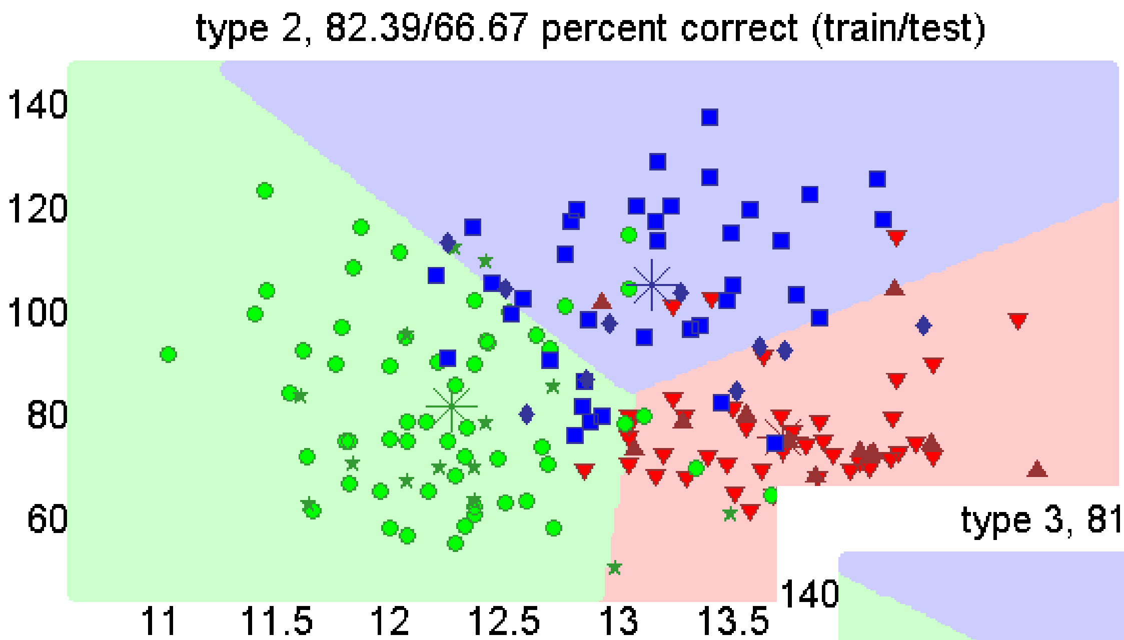


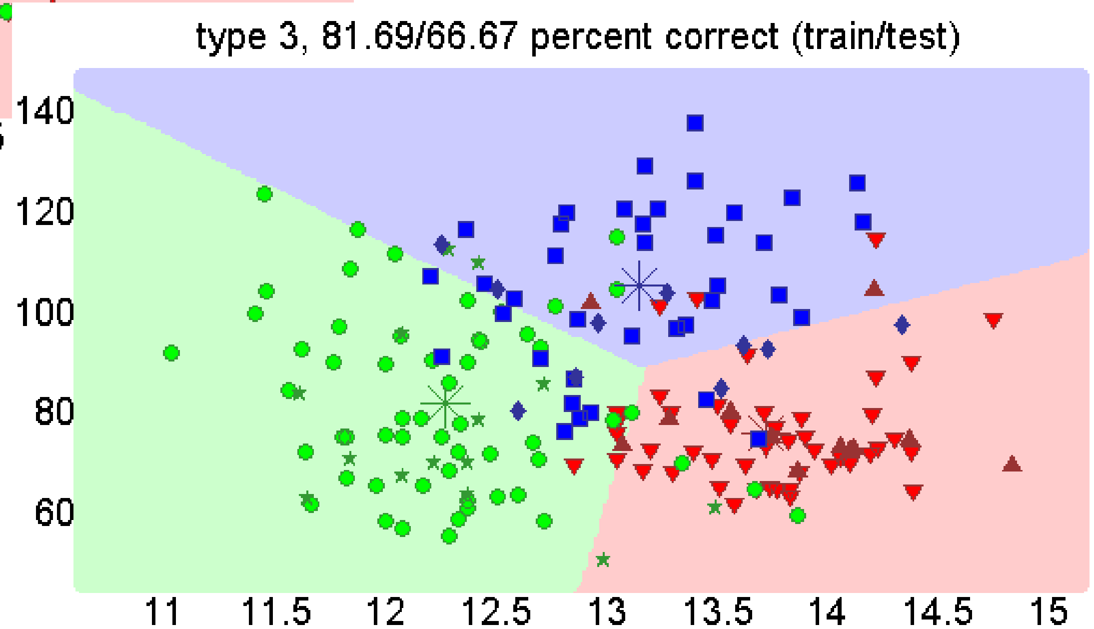
variances may be different

Case 3: Diagonal *S*



Case 3: Diagonal S





Case 4: Diagonal **S**, equal variances

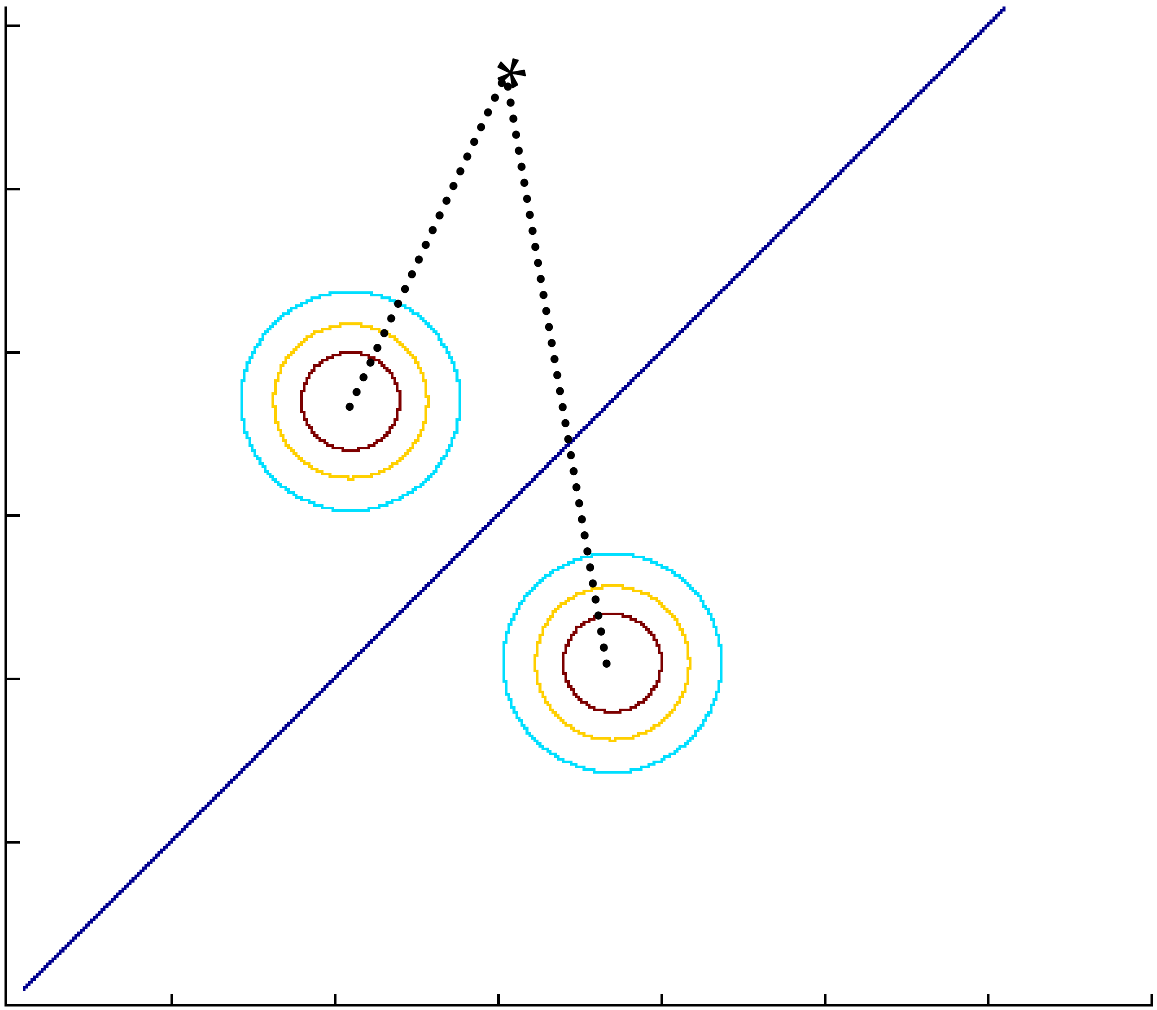
Nearest mean classifier:
Classify based on Euclidean distance to the nearest mean

$$g_i(\boldsymbol{x}) = -\frac{||\boldsymbol{x} - \boldsymbol{m_i}||^2}{2s^2} + \log \widehat{P}(C_i)$$

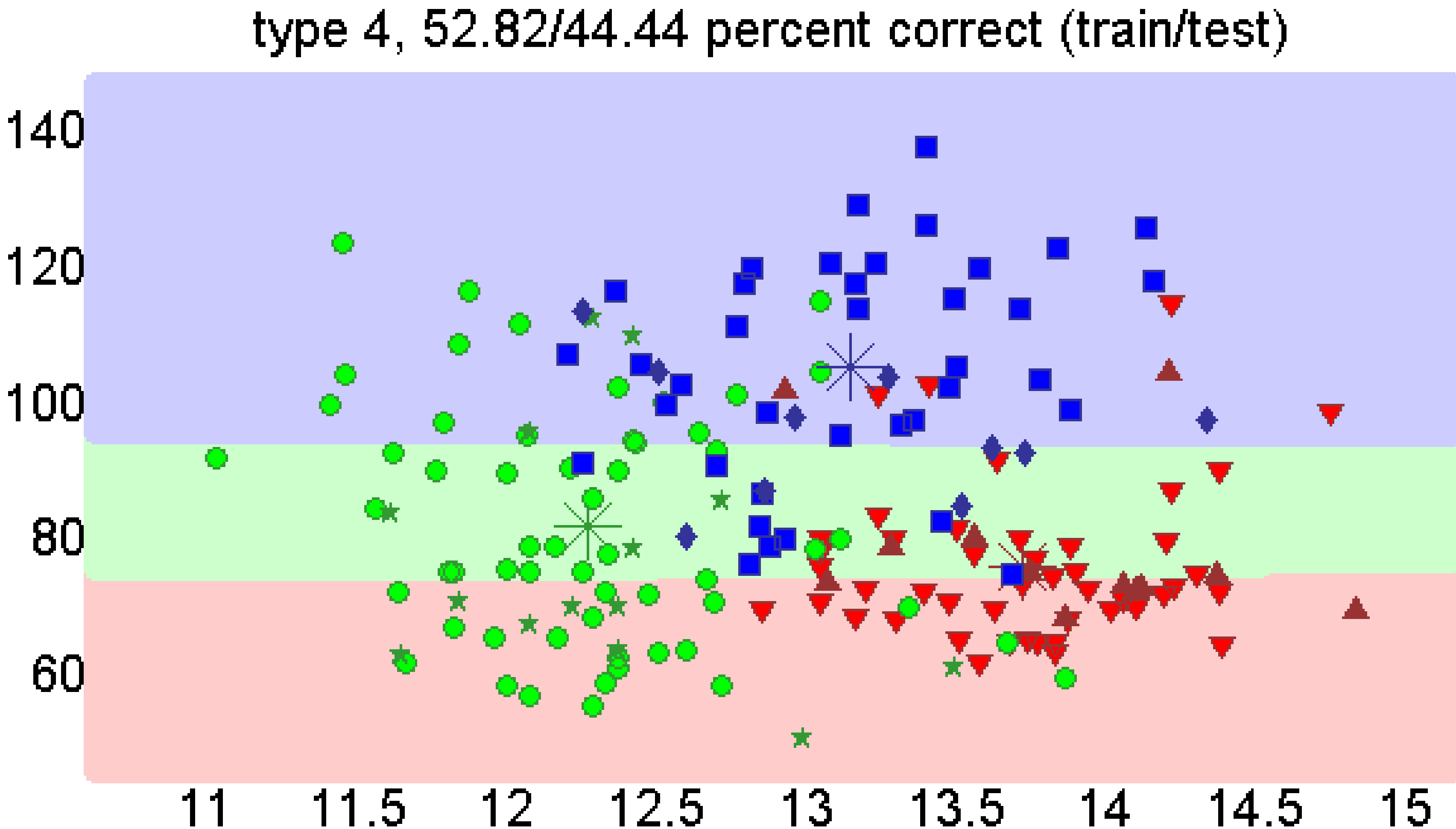
Each mean can be considered a prototype or template, hence this is **template matching**

Case 4: Diagonal **S**, equal variances

Assign the closest mean value



Case 4: Diagonal **S**, equal variances

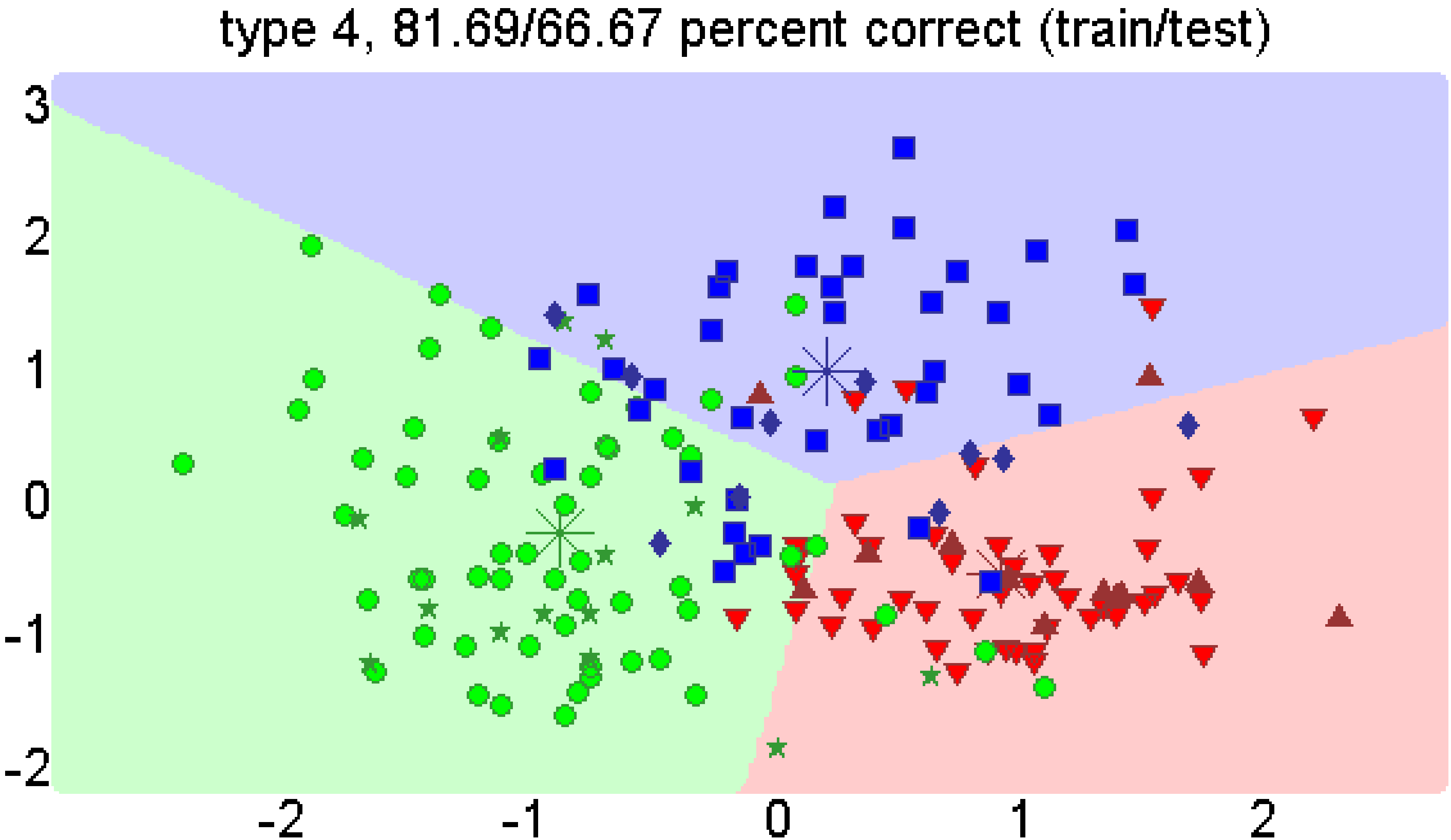


Why does it look so bad?

> No preprocessing: different variances (see axis)

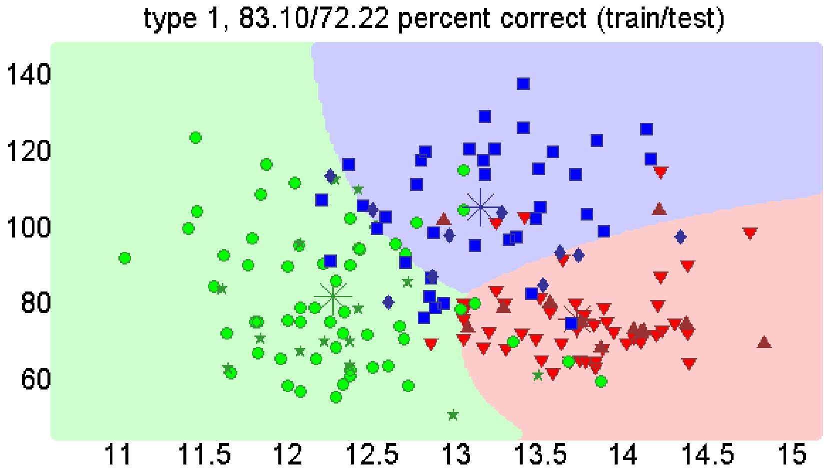
Case 4: Diagonal **S**, equal variances

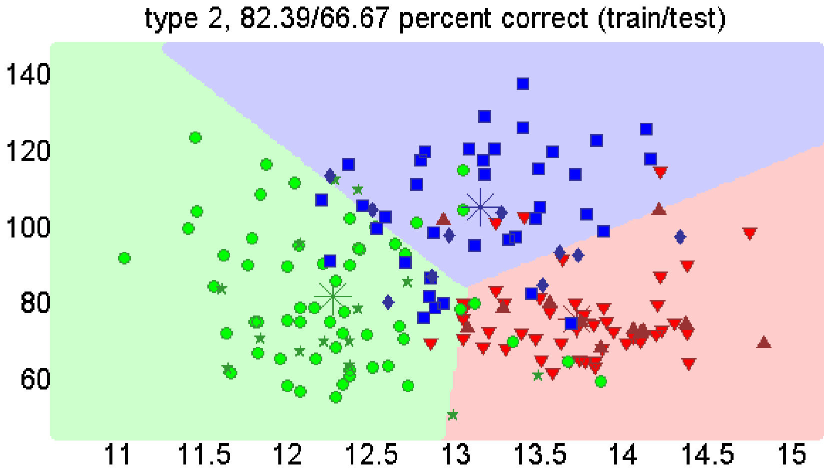
Result **with** preprocessing by z-normalization

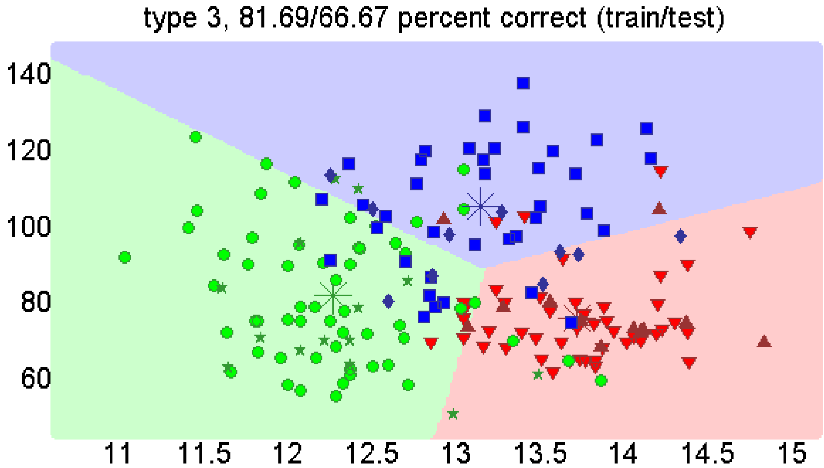


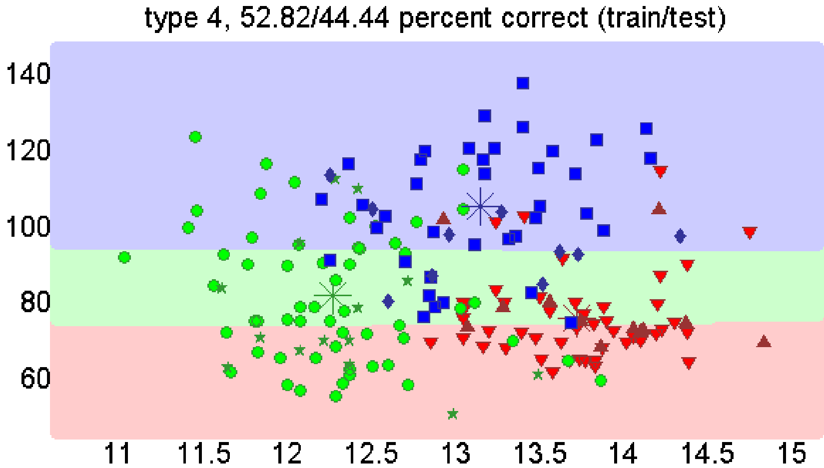
Parametric Classification

Result **without** preprocessing by z-normalization





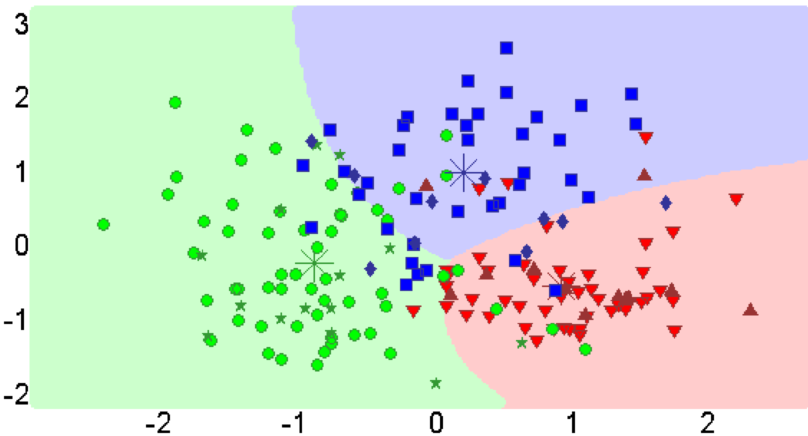




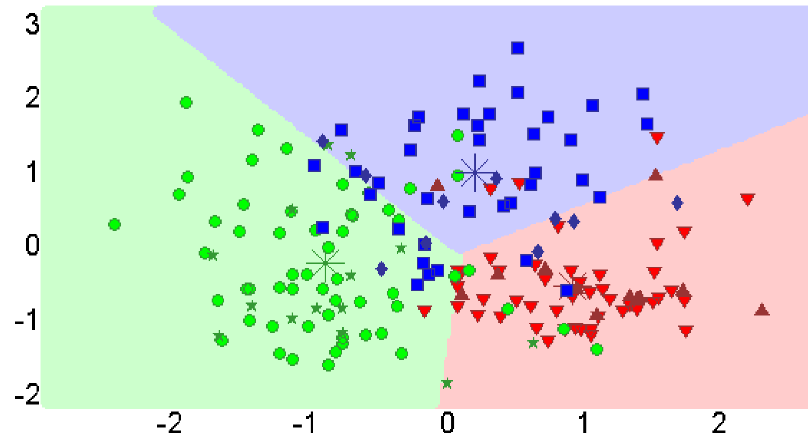
Parametric Classification

Result **with** preprocessing by z-normalization

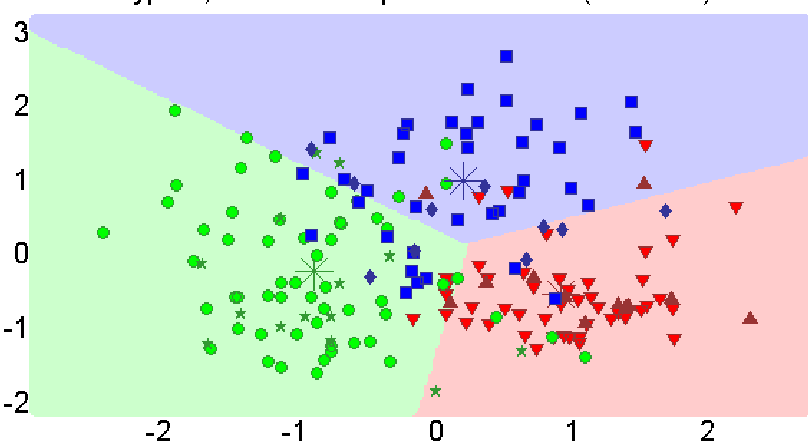
type 1, 83.10/72.22 percent correct (train/test)



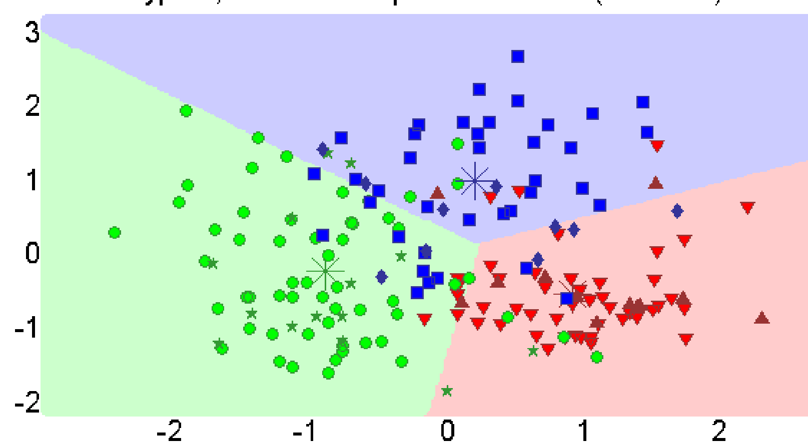
type 2, 82.39/66.67 percent correct (train/test)

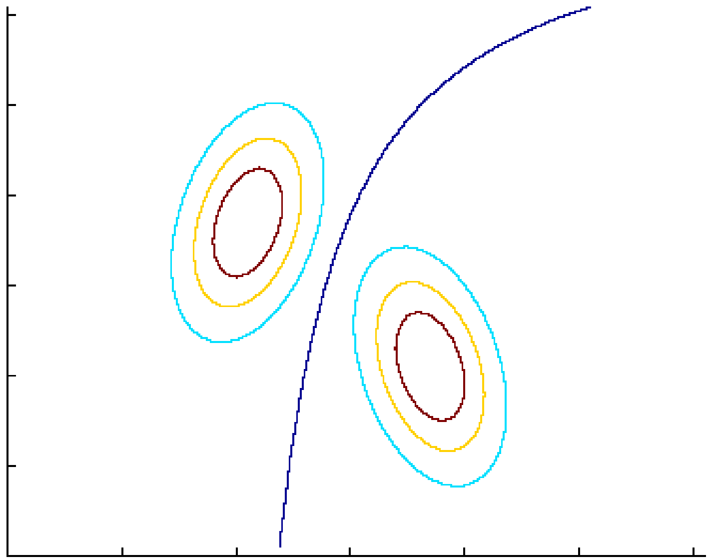


type 3, 81.69/66.67 percent correct (train/test)

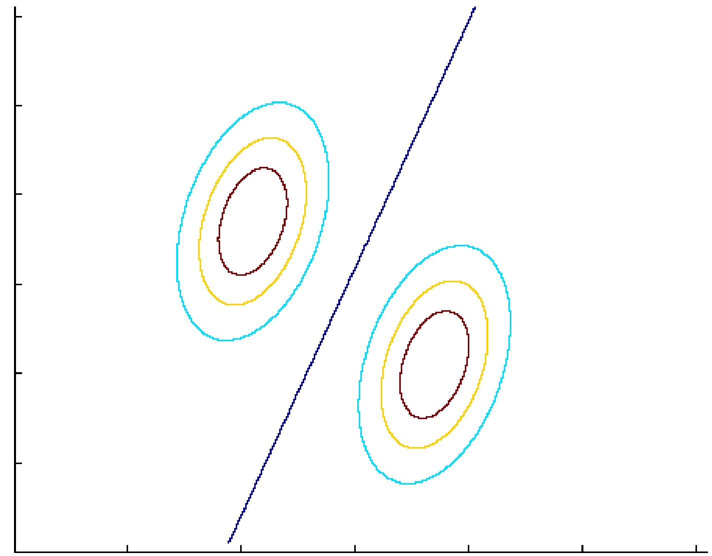


type 4, 81.69/66.67 percent correct (train/test)



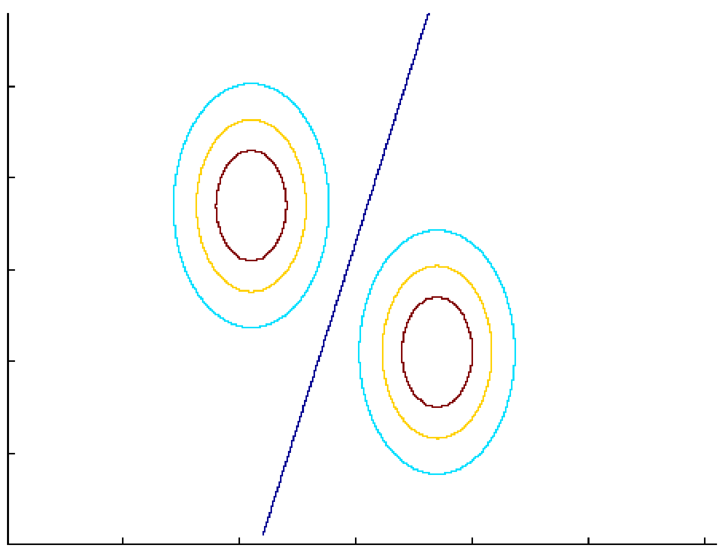


individual **S**$_i$

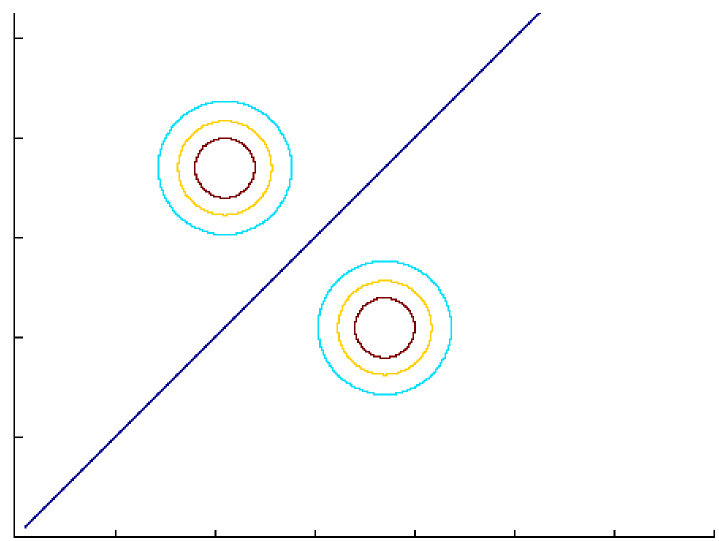


one **S** for all

diagonal **S**



diagonal **S** same variance



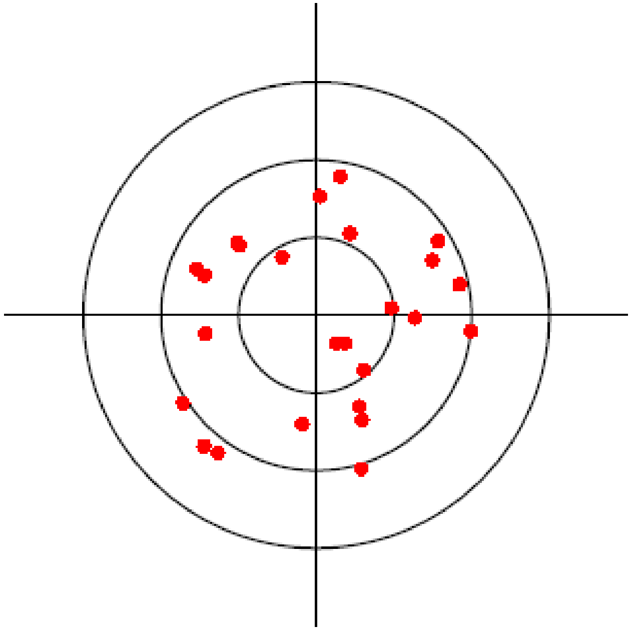
Model Selection

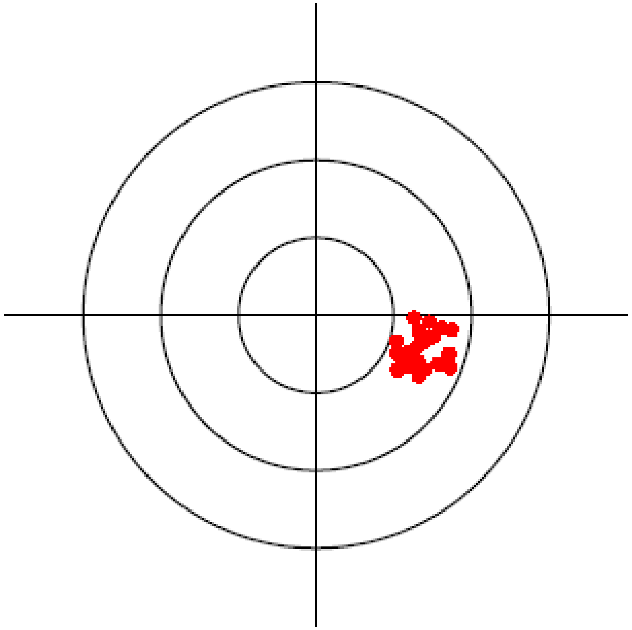
Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	D
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$D(D+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ D(D+1)/2$

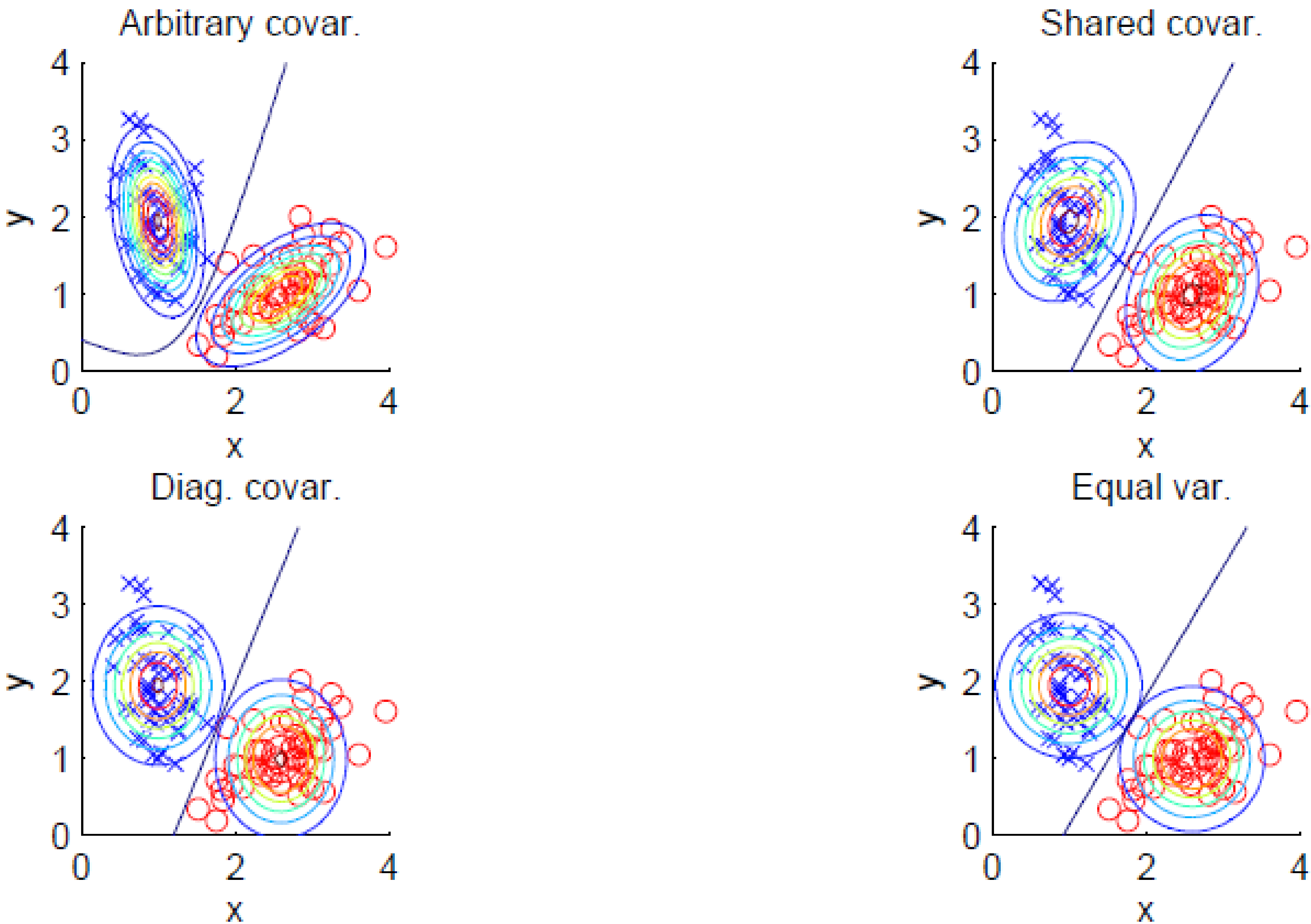
- As we increase complexity (less restricted **S**): bias decreases and variance increases
- Assume simple models (allow some bias) to control variance (regularization)

Model Selection

Variance vs. Bias






Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y

Model Selection

Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	D
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$D(D+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ D(D+1)/2$

- Weight the special cases:

$$\widetilde{\boldsymbol{S}}_i = \alpha\sigma^2\boldsymbol{I} + \beta\boldsymbol{S} + (1-\alpha-\beta)\boldsymbol{S}_i, \quad \alpha,\beta \in [0,1]$$

- Cross validation for selection

General Considerations

prior knowledge of the data

- (in)dependence
- Preprocessing is important
- How to define model quality?
  Depends on application

Overview of today

1) reminder and hint
2) multivariate data
   - descriptive statistics
   - classification
   - model selection
3) regression
   - 1D regression revisted
   - multivariate regression
   - model selection

Revisit: Linear Model for Regression

- Consider data: $(x_i, y_i)_{i=1}^N$
- linear model: $y_i = w_0 + w_1 x_i + \epsilon_i, \quad i = 1, \ldots, N$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- How to estimate parameters?

$$\min \frac{1}{2} \sum_{i=1}^N (w_0 + w_1 x_i - y_i)^2$$

$$f(w_0, w_1) = \frac{1}{2} \sum_{i=1}^N (w_0 + w_1 x_i - y_i)^2$$

Revisit: Linear Model for Regression

$$y_i = w_0 + w_1 x_i + \epsilon_i, \quad i = 1, \ldots, N, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\min \frac{1}{2} \sum_{i=1}^{N} (w_0 + w_1 x_i - y_i)^2$$

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{pmatrix}, \quad \boldsymbol{w} := \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$$

How to rewrite error in matrix-vector representation?

Revisit: Linear Model for Regression

$$y_i = w_0 + w_1 x_i + \epsilon_i, \quad i = 1, \ldots, N, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\min \frac{1}{2} \sum_{i=1}^{N} (w_0 + w_1 x_i - y_i)^2$$

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{pmatrix}, \quad \boldsymbol{w} := \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\min \frac{1}{2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

What are the estimated parameters?

Revisit: Linear Model for Regression

$$y_i = w_0 + w_1 x_i + \epsilon_i, \quad i = 1, \ldots, N, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$$\min \frac{1}{2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

$$f(\boldsymbol{w}) = \frac{1}{2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 = \frac{1}{2} \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right)^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right)$$

$$= \frac{1}{2} \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y}\right)$$

$$\nabla f(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} \overset{!}{=} 0$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}$$

Revisit: Linear Model for Regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

How is the estimator used to predict?

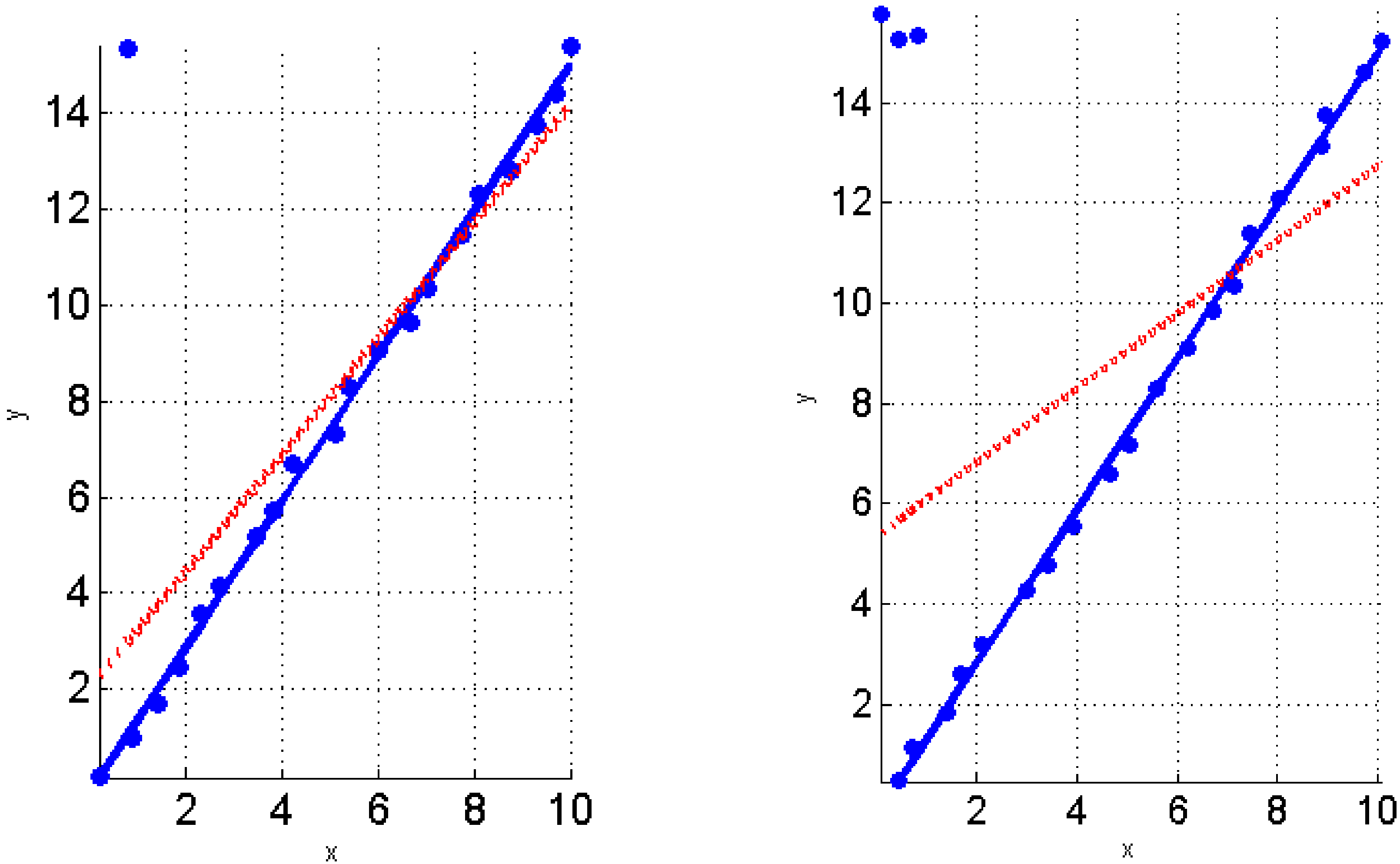
$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{w}}$$

$$\widehat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}}_{\boldsymbol{H}}\boldsymbol{y}$$

- H is called the **hat matrix**
- h_{ii} are **leverages** $[\frac{1}{n}, 1]$
- Encodes influence of input points in y

Revisit: Linear Model for Regression Outliers

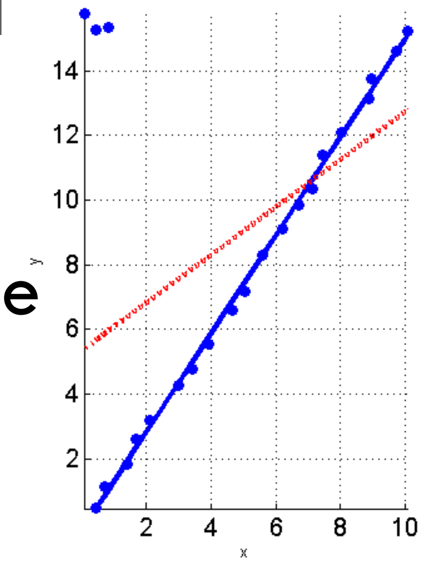
Regression line without (blue) vs. **with** outliers (red)

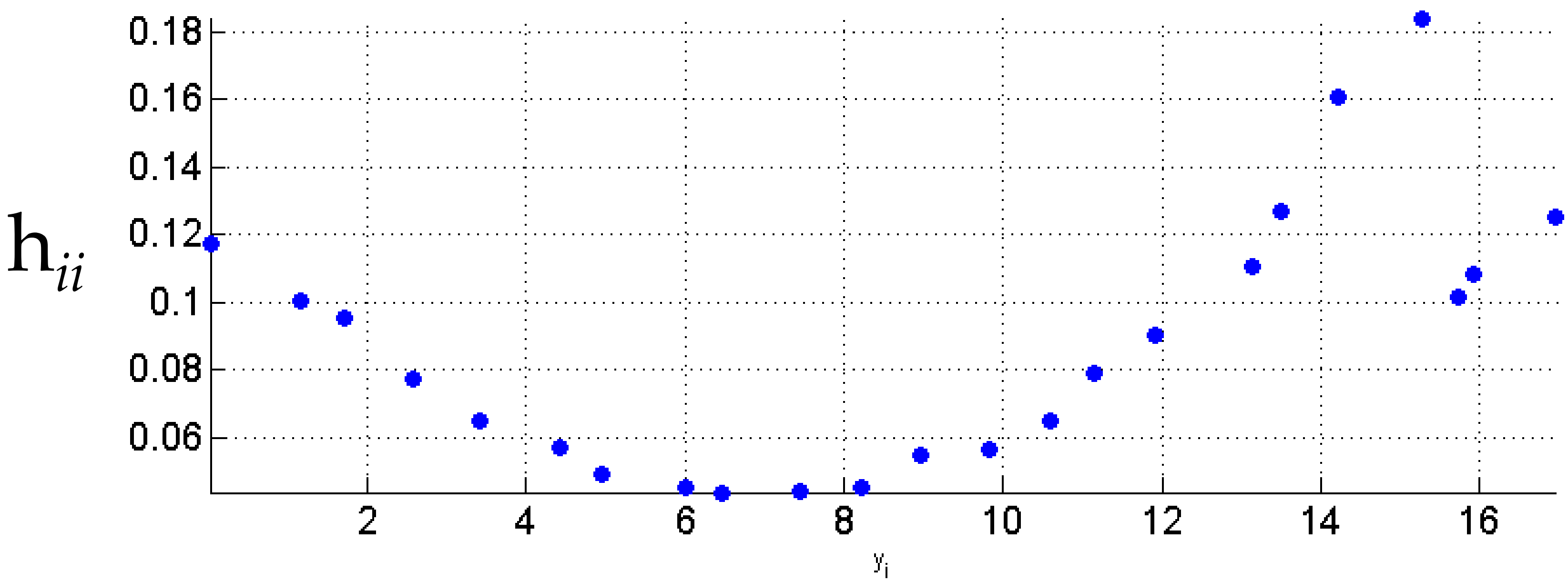


Revisit: Linear Model for Regression Diagnostic Insights

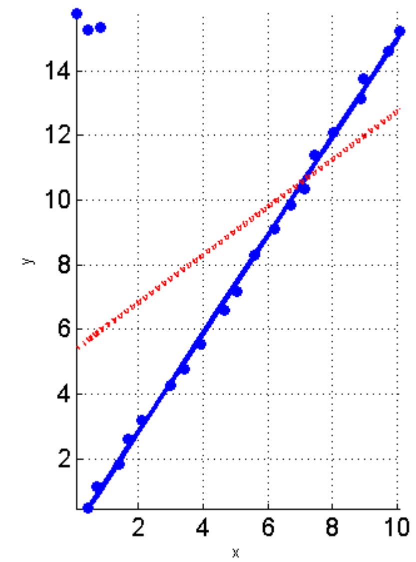
Influence as **leverage** of points:

How much does point i influence the estimate

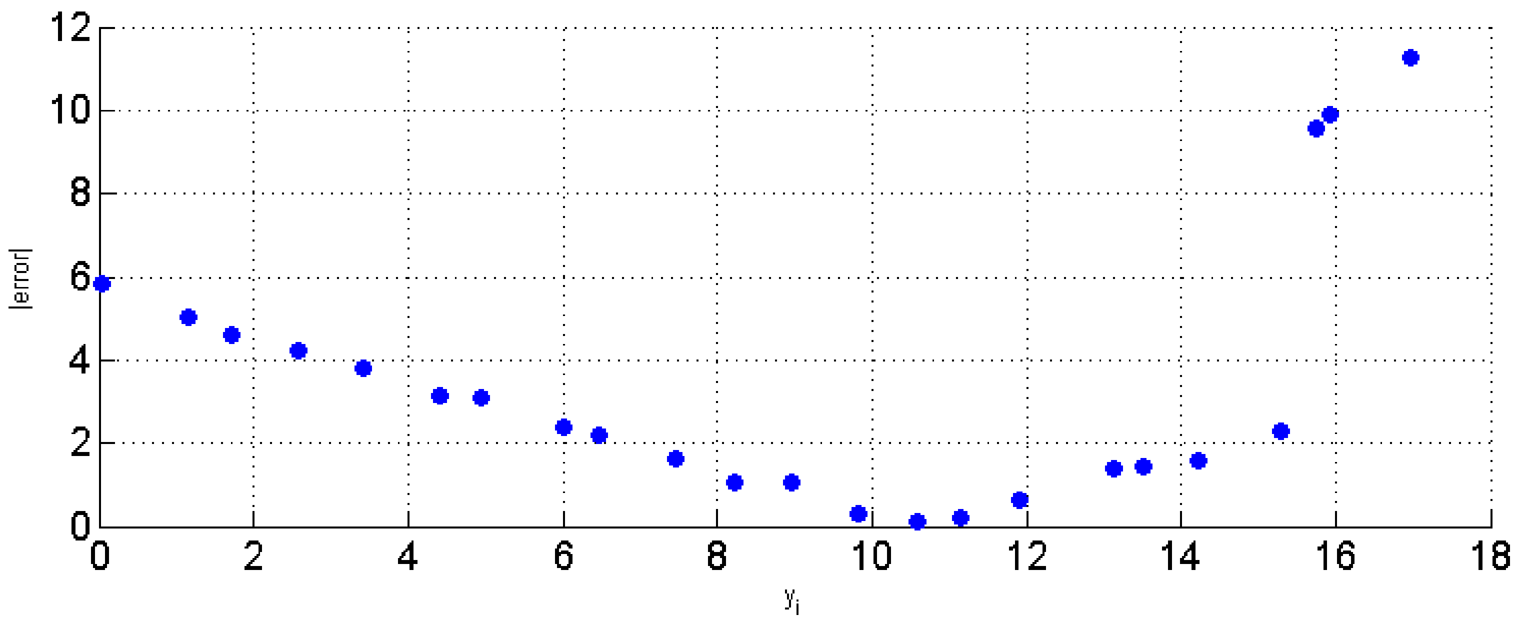




Revisit: Linear Model for Regression
Diagnostic Insights



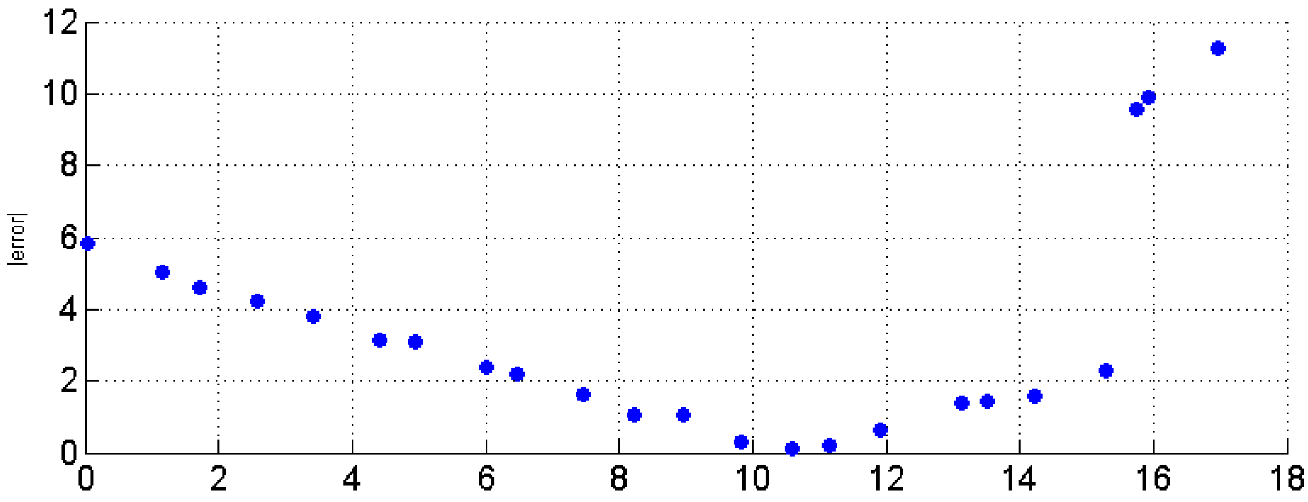
Absolute error $|\hat{y}_i - y_i|$



Revisit: Linear Model for Regression Diagnostic Insights

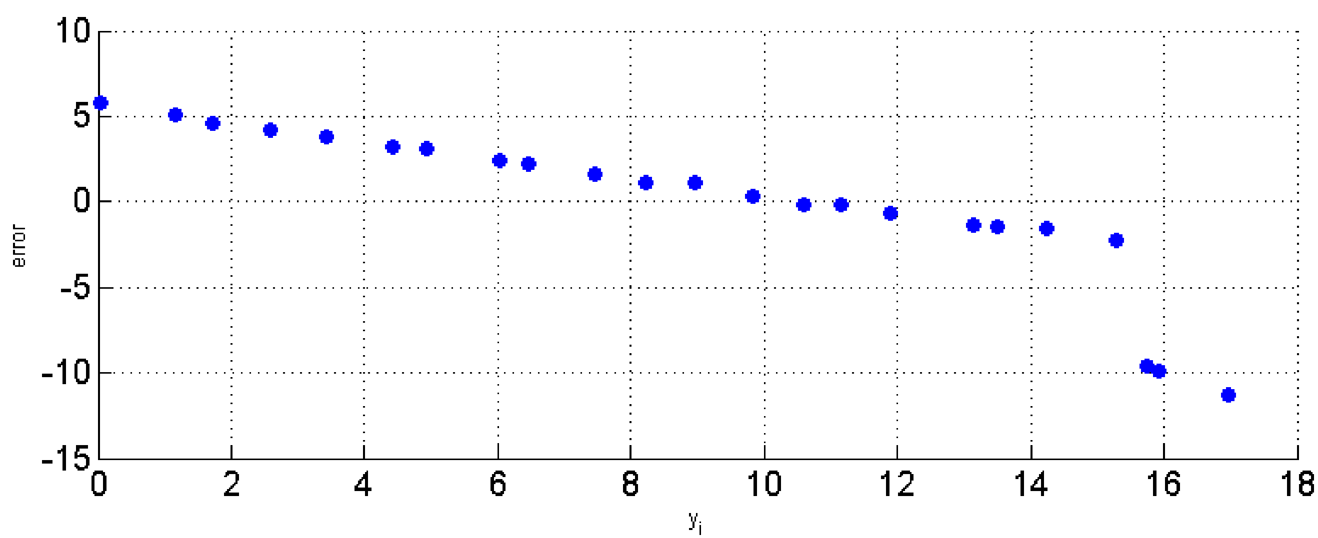
Absolute error

$$|\hat{y}_i - y_i|$$



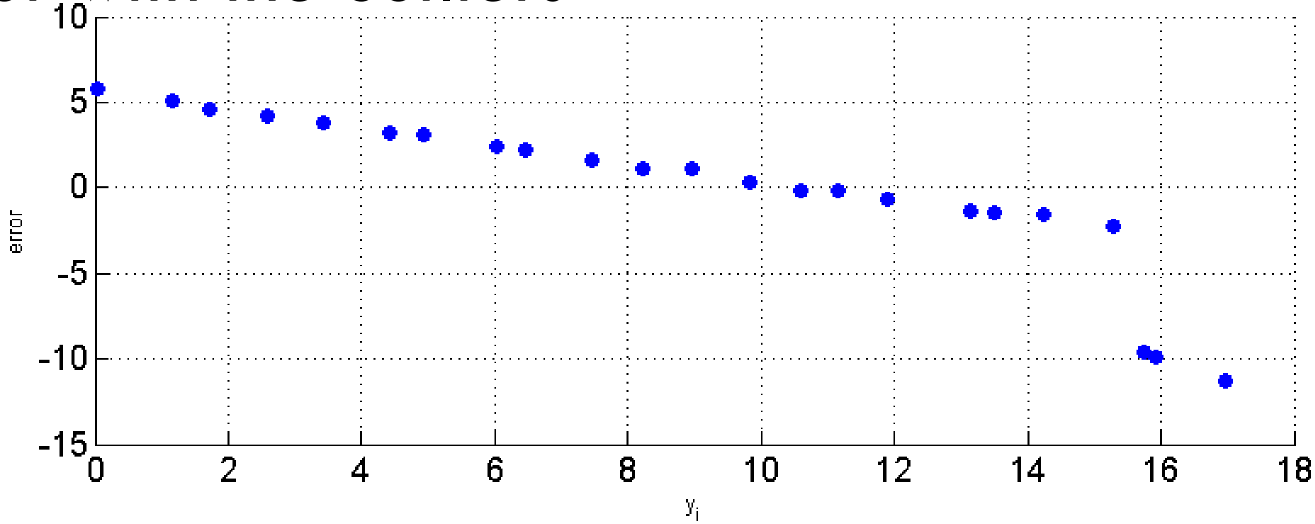
error with sign

$$\hat{y}_i - y_i$$



estimated error with the outliers

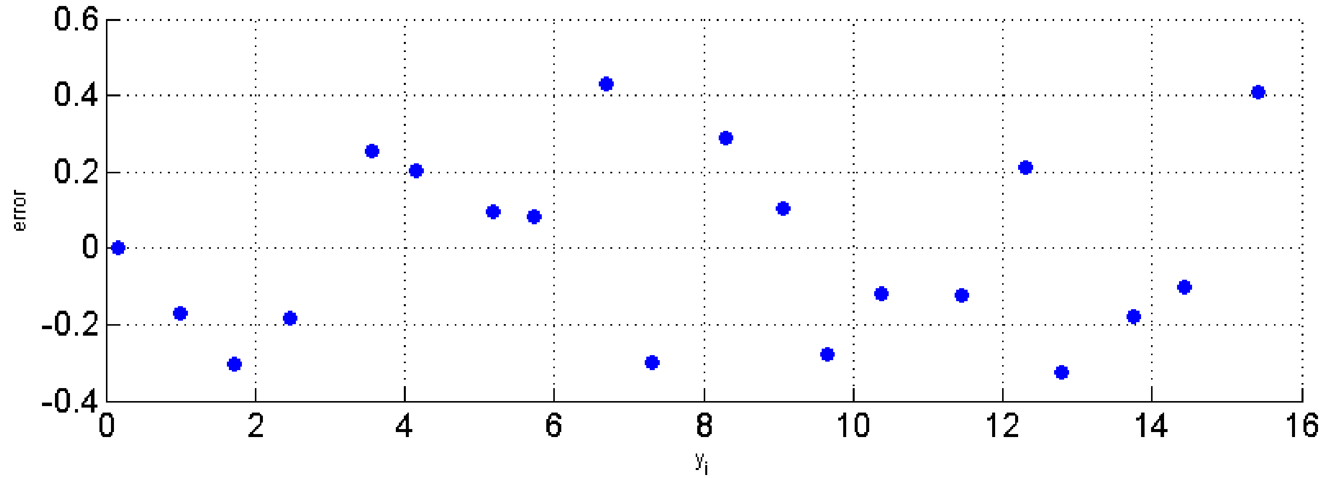
$\hat{y}_i - y_i$



$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

estimated error without the outliers

$\hat{y}_i - y_i$



Bottom conforms better to $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Measures for Model Quality

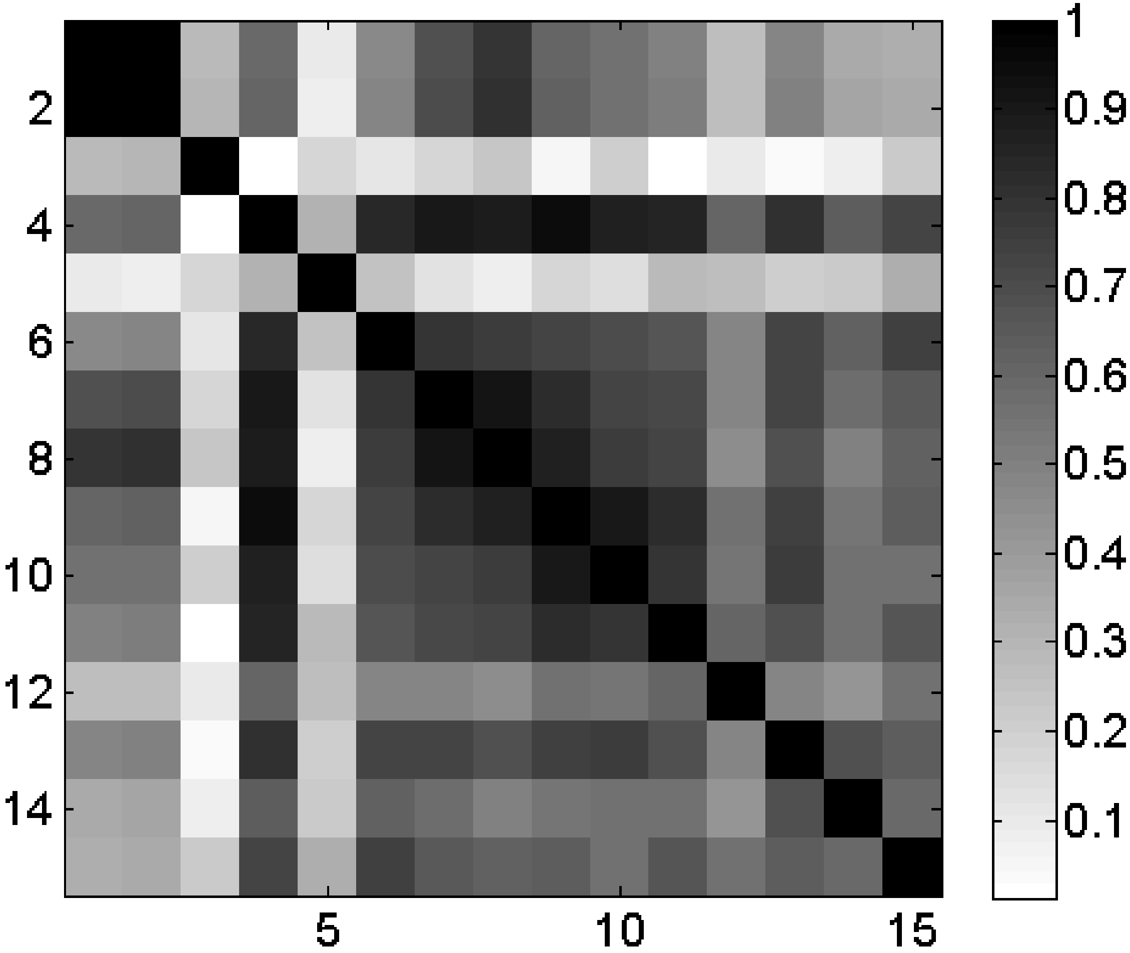
- Visual Inspection
- Mean Squared Error (MSE) $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$
- Root Mean Squared Error (RMSE) $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$
- Coefficient of determination

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}$$

- Number of Parameters
- Correlation (pairwise)
- Statistics of the Parameters:
  - Confidence Interval $\widehat{w}_i \pm \widehat{\sigma}_i, \quad \widehat{w}_i \pm 2\widehat{\sigma}_i$
- …

Correlation

Example: absolute values of correlation matrix



Diagnostic Insights

Improve model by

- Check model requirements
- Define outliers: there is no unique definition
- Estimate model on subset of points
- Inspect changes of
  - Errors (e.g. studentized residuals)
  - estimated parameters
- Consider changes of their values and statistics

Bias vs. variance

Multivariate Regression
What changes from one to many ?

$$y_i = w_0 + w_1 x_i + \epsilon_i, \quad i = 1, \ldots, N$$

$$\rightsquigarrow \quad y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots + w_D x_{iD} + \epsilon_i, \quad i = 1, \ldots, N$$

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} := \begin{pmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ & & \vdots & \\ 1 & x_{N1} & \cdots & x_{ND} \end{pmatrix}, \quad \boldsymbol{w} := \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\min \frac{1}{2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{w}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}}_{\boldsymbol{H}}\boldsymbol{y}$$

The same estimator as before can be used! ☺

Multivariate Regression

What is a „linear model“?

$$(1) \quad y_i = w_0 + w_1 x_i + \epsilon_i$$

$$(2) \quad y_i = w_0 + w_1 x_{1i} + w_2 x_{2i} + \epsilon_i$$

$$(3) \quad y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i$$

$$(4) \quad y_i = w_0 + w_1 x_i + w_2^2 x_i + \epsilon_i$$

$$(5) \quad y_i = w_0 + w_1 \log(x_i) + w_2 x_i + \epsilon_i$$

$$(6) \quad y_i = w_0 + w_1 \log(x_i) + w_2^3 x_i + \epsilon_i$$

$$\rightsquigarrow \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{w} + \boldsymbol{\epsilon}$$

1, 2, 3, 5 are linear with respect to the parameters $\boldsymbol{w}$

Multivariate Regression

- Multivariate linear model as

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2}^2 + \epsilon_i, \quad i = 1, \dots, N$$

$$y_i = w_0 + w_1 z_{i1} + w_2 z_{i2} + \epsilon_i, \quad i = 1, \dots, N$$

- Multivariate polynomial model

  with new higher-order variables
  (basis functions, kernel trick: Chapter 13)

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \ \boldsymbol{w} := \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix}, \ \boldsymbol{Z} := \begin{pmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{21} & x_{21}^2 \\ & \vdots & \\ 1 & x_{N1} & x_{N1}^2 \end{pmatrix} \in \mathbb{R}^{N \times M}$$

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{w} + \boldsymbol{\epsilon}$$

The same estimator as before can be used! ☺

Multivariate Regression

$$y_i = w_0 + w_1 z_{i1} + w_2 z_{i2} + \ldots + w_M z_{iM} + \epsilon_i, \quad i = 1, \ldots, N$$

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{w} := \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix}$$

$$\boldsymbol{Z} := \begin{pmatrix} 1 & x_{11} & x_{11}^2 & \sin(x_{11}) & \cdots & \log(x_{1P}) \\ 1 & x_{21} & x_{21}^2 & \sin(x_{21}) & \cdots & \log(x_{2P}) \\ & & \vdots & & & \\ 1 & x_{N1} & x_{N1}^2 & \sin(x_{N1}) & \cdots & \log(x_{NP}) \end{pmatrix} \in \mathbb{R}^{N \times M}$$

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{w} + \boldsymbol{\epsilon}$$

The same estimator as before can be used! ☺

$$\widehat{\boldsymbol{w}} = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{Z}\widehat{\boldsymbol{w}} = \boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}$$

Linear Models: Best Practice Tipps

- Look at the data!
- Investigate correlation
- Correlation is not causality
- Investigate errors: is there a visible trend?
- Outlier: there is no unique definition
- Point with high influence is not necessarily an outlier
- Consider different measures for model selection
  Not only the smallest error

Summary

- Multivariate Data
- Multivariate Normal Distribution
- Correlation vs. Dependence
- Classification with Discriminant Functions

- Model Quality
- Model Selection
- Preprocessing is important
- Linear Models for Regression

Books

- *Pattern Classification*
  Duda, Hart, Stork
- *Pattern Recognition And Machine Learning*
  - Bishop
- *Regression: Models, Methods and Applications*
  - Fahrmeir, Kneib, Lang

APPENDIX

Covariance / Correlation

$$\boldsymbol{x} \in \mathbb{R}^D \quad \mathrm{Cov}(x_i, x_j) = \sigma_{ij} \qquad \mathrm{Corr}(x_i, x_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \in [-1, 1]$$

$$\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_D]^{\mathrm{T}} \in \mathbb{R}^D$$

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \\
&= \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
&= \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - 2\boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
&= \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - 2\boldsymbol{\mu}\mathrm{E}[\boldsymbol{x}]^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - 2\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}
\end{aligned}$$

Revisit: Linear Model for Regression
Diagnostic Insights

Improve model quality

- Estimate model on subset of points
- Inspect changes of
  - Errors (studentized residuals)
  - estimated parameters
- Consider changes of their values and statistics
  - Confidence intervals $\widehat{\epsilon}_i \sim \mathcal{N}(0, \widehat{\sigma}^2)$

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow P(|\frac{X-\mu}{\sigma}| < 2\sigma) = 95.5\%$$
$$\Rightarrow P(|\frac{X-\mu}{\sigma}| < 3\sigma) = 99.7\%$$

Model Quality for Selection

Confusion matrix

In general:

Positive = identified

negative = rejected

	is positive	is negative
Predicted postive	TP	FP (Type I error)
Predicted negative	FN (Type II error)	TN

Therefore:

- True positive (TP) = correctly identified
- False positive (FP) = incorrectly identified
- True negative (TN) = correctly rejected
- False negative (FN) = incorrectly rejected