Statistical Learning in a Nutshell

Yangyao CHEN
cyy12345678@163.com

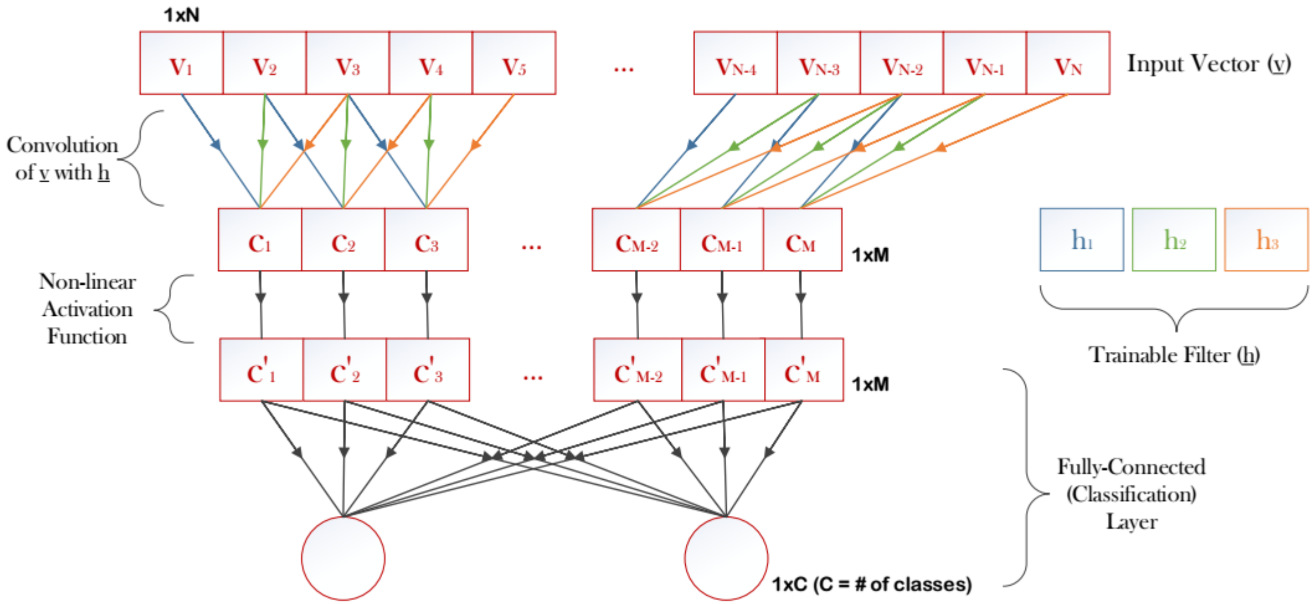
A1. Astrophysicists work with them …

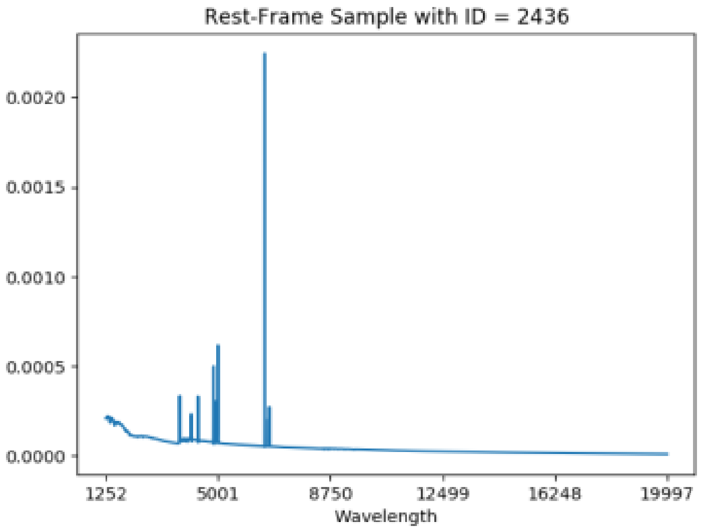
Examples

- Spectral redshift determination
- Galaxy morphological type
- Density initial condition reconstruction

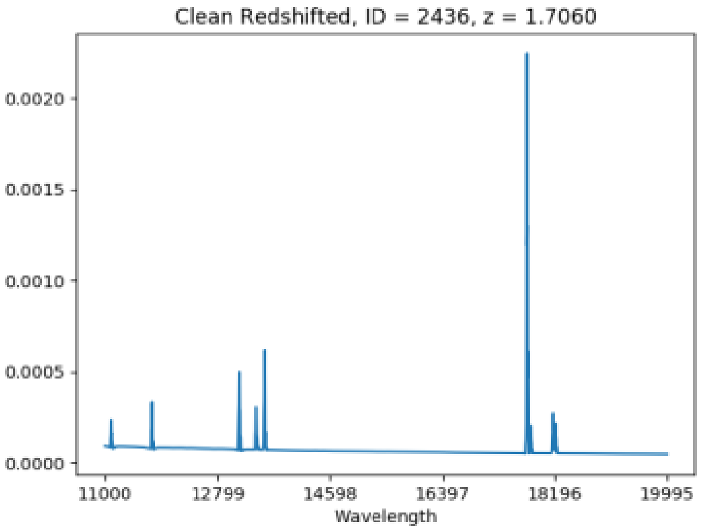
Convolutional Neural Networks for Spectroscopic Redshift Estimation on Euclid Data

Radamanthys Stivaktakis[1,2], Grigorios Tsagkatakis[2], Bruno Moraes[3], Filipe Abdalla[3,4], Jean-Luc Starck[5], Panagiotis Tsakalides[1,2]
Department of Computer Science - University of Crete, Greece[1]
Institute of Computer Science - Foundation for Research and Technology (FORTH), Greece[2]
Department of Physics & Astronomy, University College London, UK[3]
Department of Physics and Electronics, Rhodes University, South Africa[4]
Astrophysics Department - CEA Saclay, Paris, France[5]

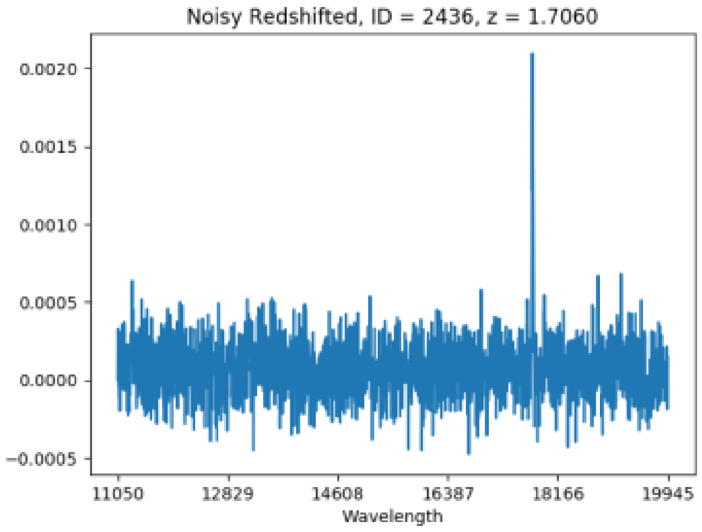




(a) Clean Rest-Frame Spectral Profile

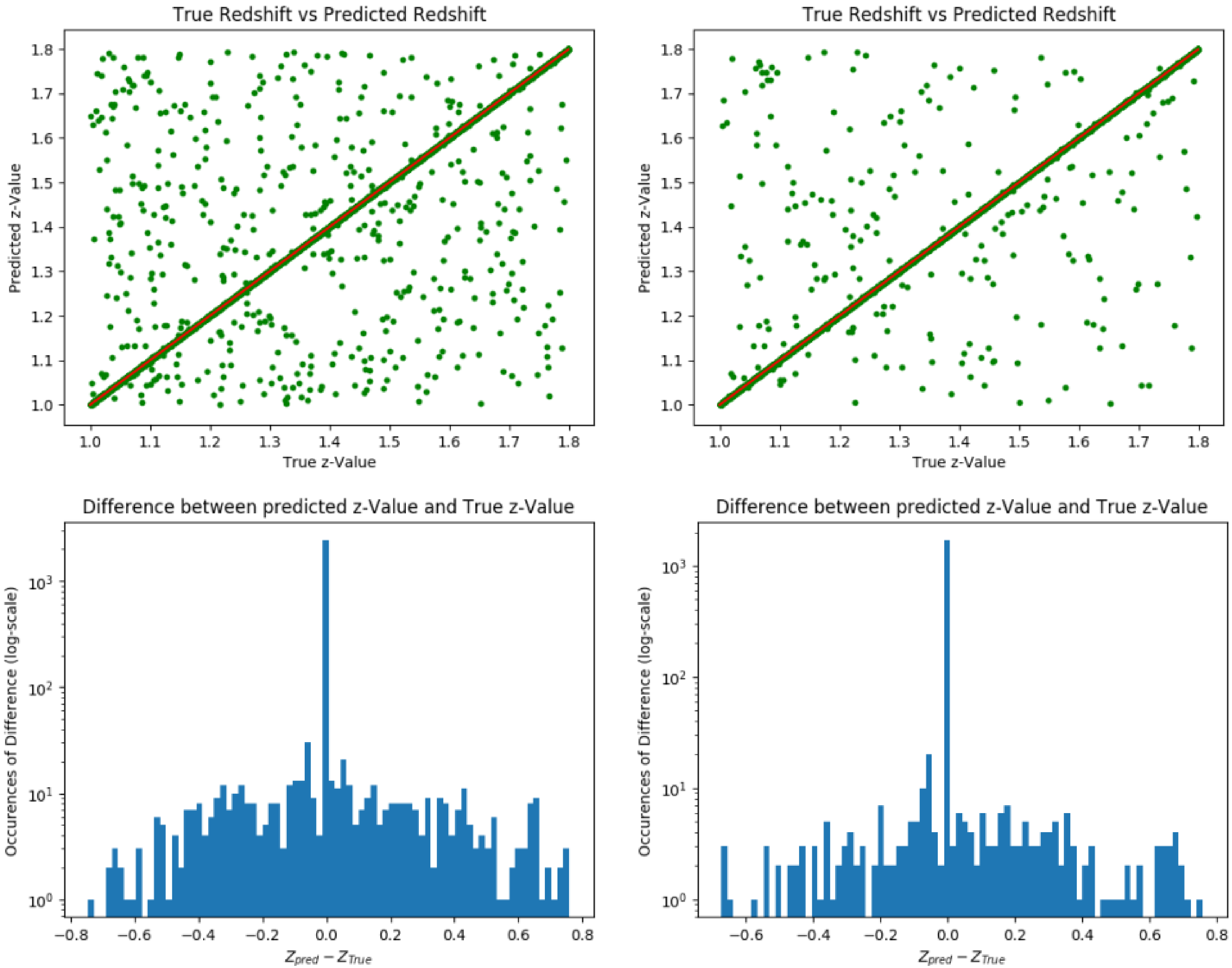


(b) Clean (Randomly) Redshifted Equivalent



(c) Noisy Redshifted Equivalent

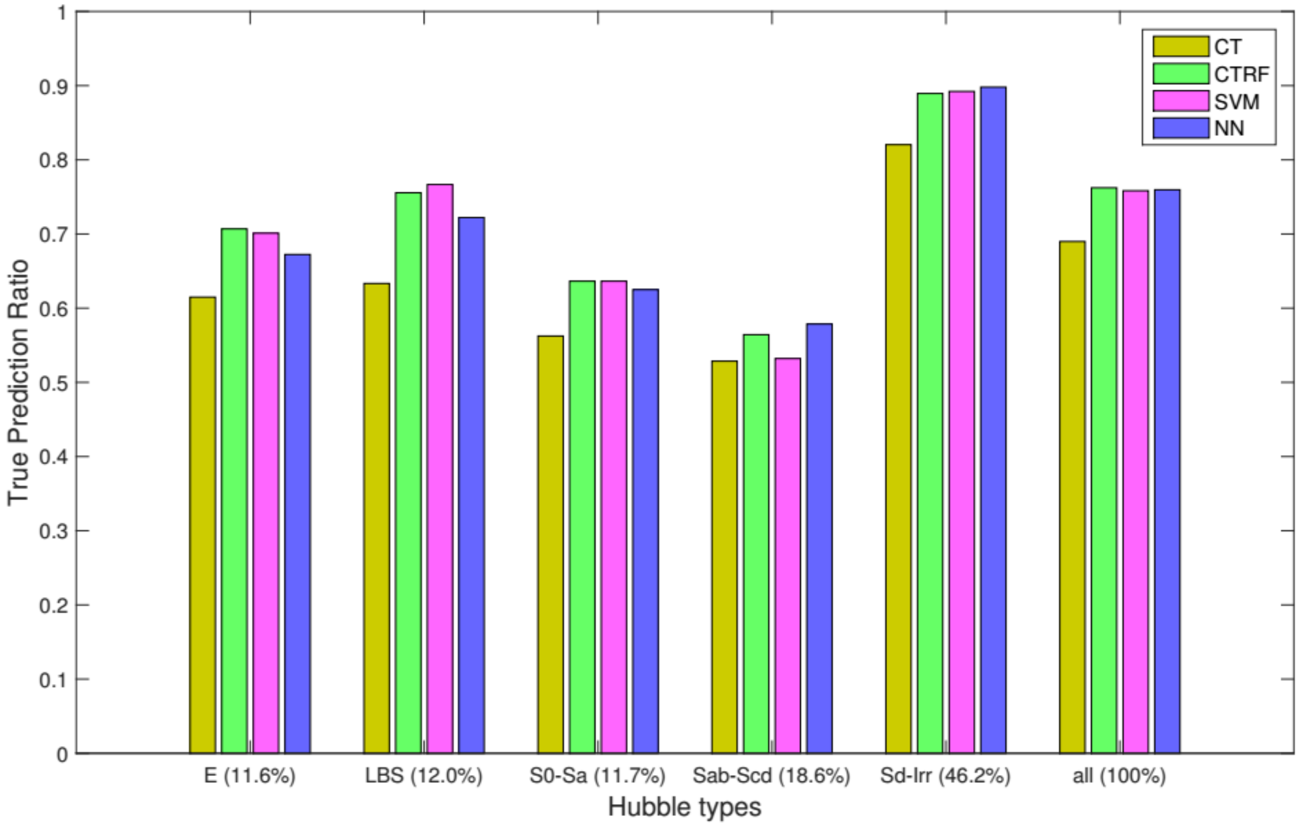
A. Purpose of this mini-talk

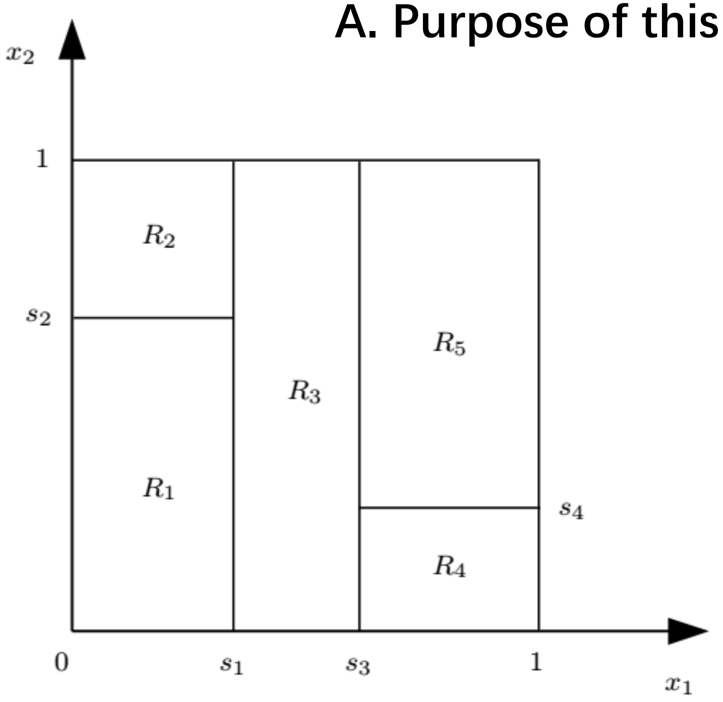


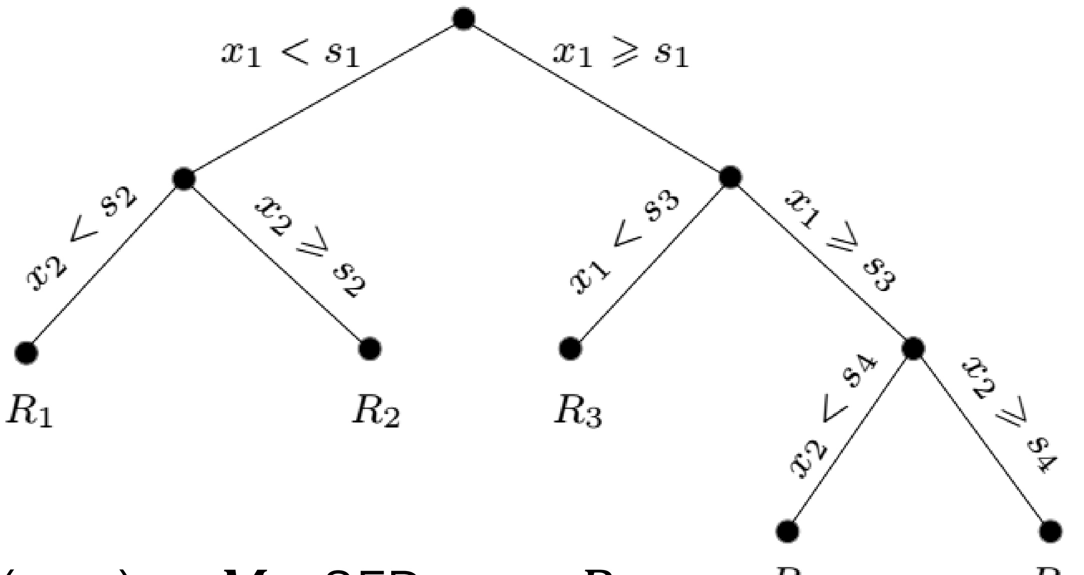
A. Purpose of this mini-talk

Galaxy And Mass Assembly: Automatic Morphological Classification of Galaxies Using Statistical Learning

Sreevarsha Sreejith,[1] Sergiy Pereverzyev Jr.,[2] Lee S. Kelvin,[1,3] Francine Marleau,[1] Markus Haltmeier,[2] Judith Ebner,[2] Joss Bland-Hawthorn,[4] Simon P. Driver,[5,6] Alister W. Graham,[7] Benne W. Holwerda,[8] A. M. Hopkins,[9] J. Liske,[10] Jon Loveday,[11] Amanda J. Moffett,[12] K. A. Pimbblet,[13,14] Edward N. Taylor,[7] Lingyu Wang,[15,16] Angus H. Wright[17]







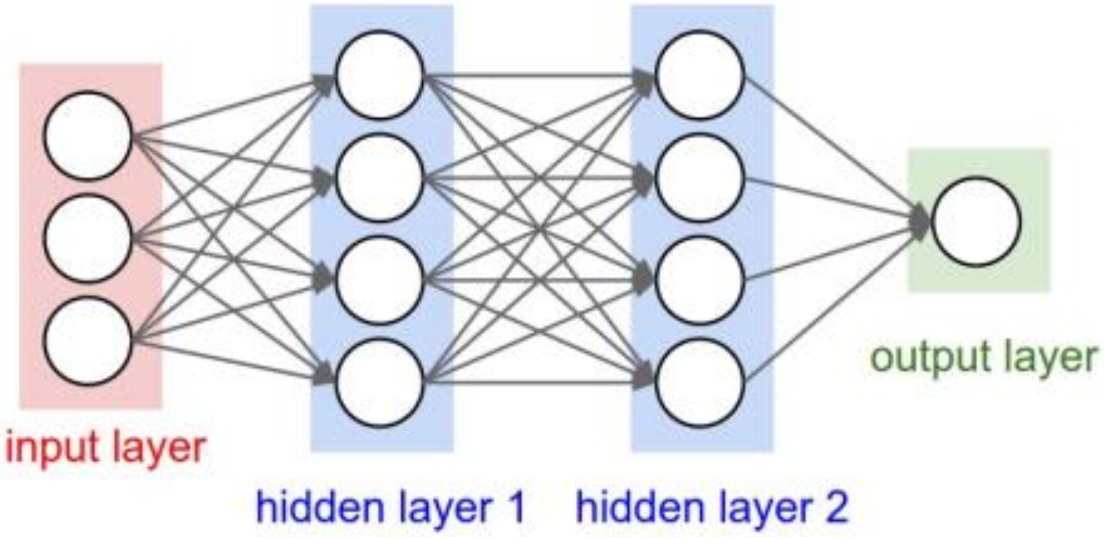
(x, …) = M_*, SFR, u-r, R_e, …

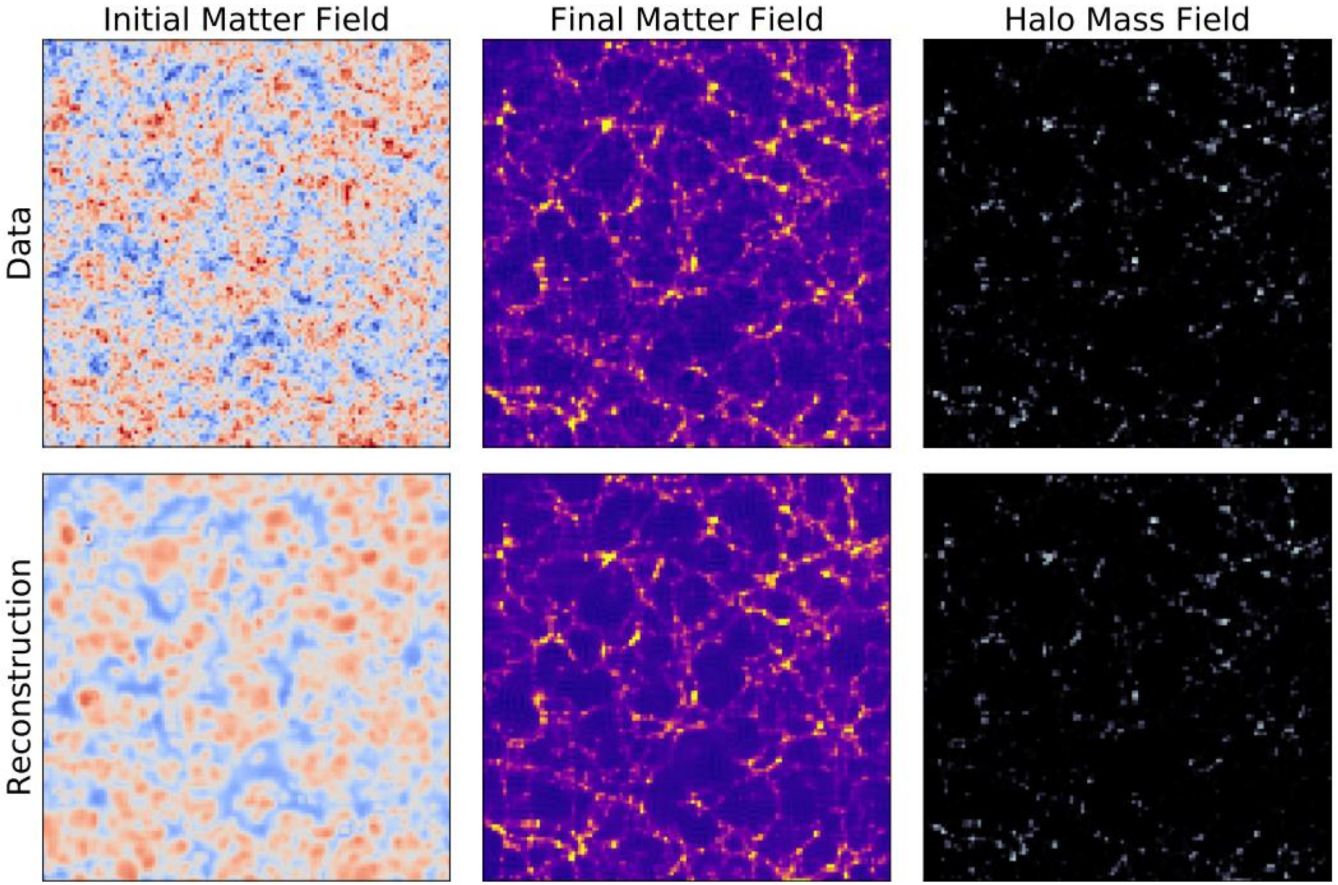
Cosmological Reconstruction From Galaxy Light: Neural Network Based Light-Matter Connection

Chirag Modi,[a] Yu Feng,[a] Uroš Seljak[a,b]

[a]Berkeley Center for Cosmological Physics and Department of Physics, University of California, Berkeley, CA 94720

[b]Physics Division, Lawrence Berkeley National Laboratory, Cyclotron Rd, Berkeley, CA 94720





A2. We are interested in them …

Regression

- Convolution Neural Network for regression (Xiaosheng Zhao)

Classification

- Decision Tree + Random Forest for classification(Cheng Cheng)
- SVM for classification (Kai Wang)

Clustering

- K-Means for clustering(Yangyao Chen)
- Hierarchical Algorithm for clustering(Kai Wang)
- Model Mixture for clustering(Kai Wang)

Sampling

- MCMC for sampling(Kai Wang)

A3. Purpose of this mini-talk

An introduction to the Framework of ML
Procedure in doing ML
Most important things to be considered

Hope that

Find the exact position of topic we have presented/in literature
Know what are missing in our previous talk
Pick up good ML model in future works

B1. But, what is machine learning?

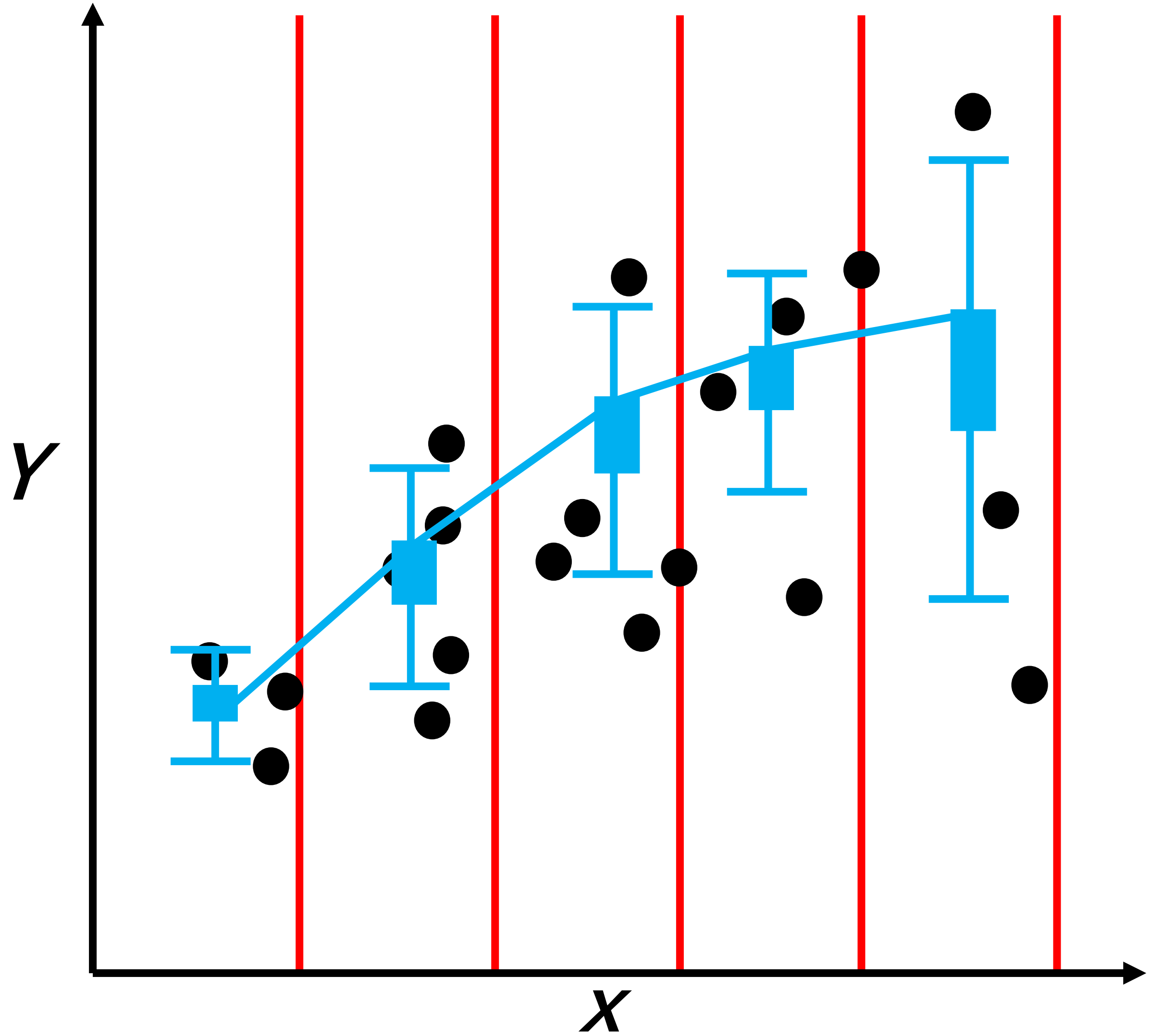
Data + Model + objective
+ optimization
+ Prediction/Interpretation

Anything that saves
our brain force

- Artificial Intelligence (AI)
  - Machine Learning (ML)
    - Neural Networks (NNs)
      - Ordinary NNs
      - Deep NNs (CNNs, RNNs, …)
    - Linear/Polynomial Regr (with LS, MLE, MAP), …
      LDA/QDA, logistic Regr, Fisher discriminant, SVMs, KNN, …
      K-Means, Mixture Gaussiam, Hierarchicals, …
    - …
  - Preprocessors(cpp, cuda, RDMBS embed, …)
    Compilers(cc, fort, javac)
    Interpreters (shell, Py, R, …)
    TCP/IP protocol stacks (ssh, sftp, https, …)
    IDEs (vim, VS, Spyder, …)
    Text soft (word/latex)
    Paper managements (Endnote, Mendeley, …)
  - …

B2. ML is everywhere …

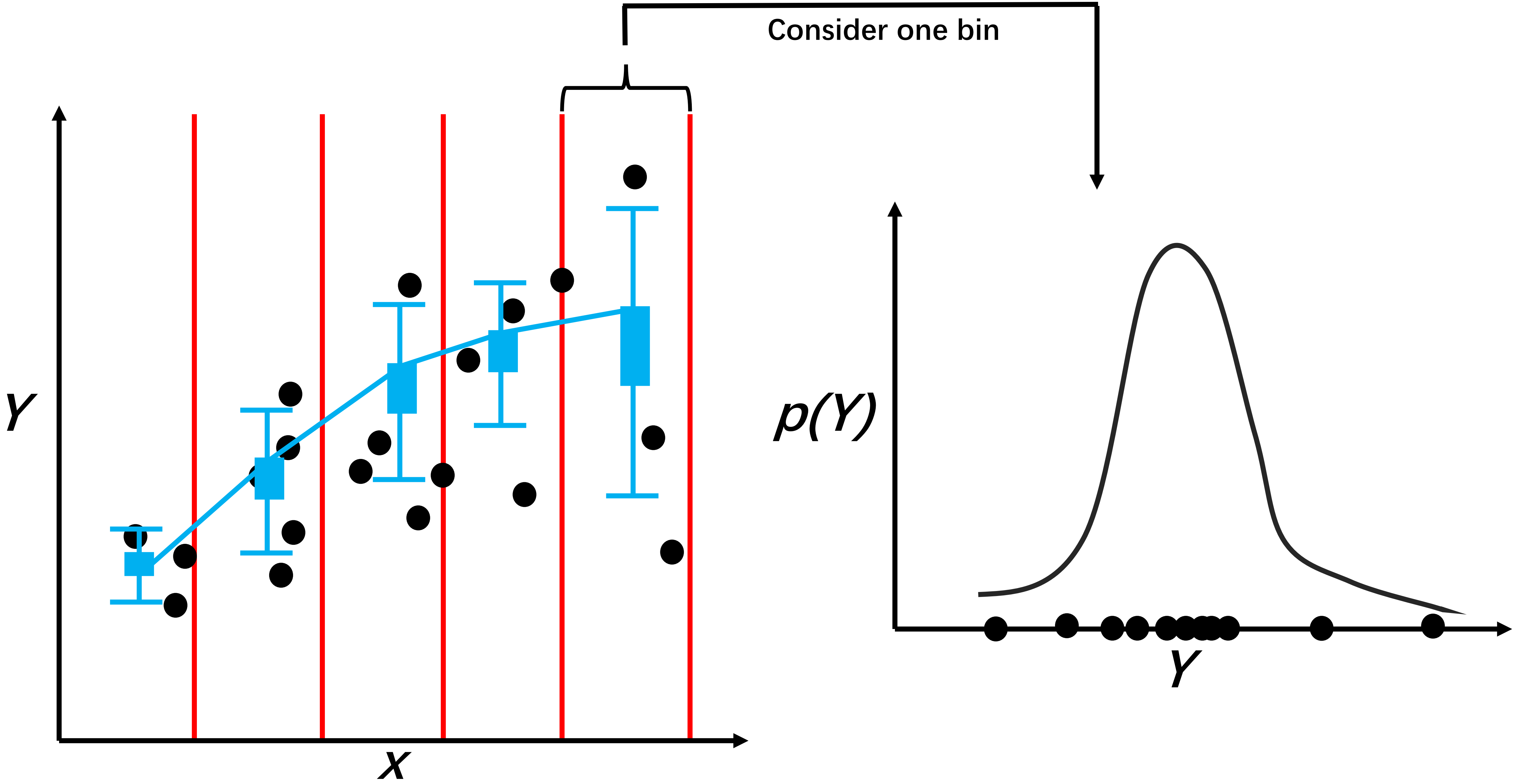
We are always doing this:



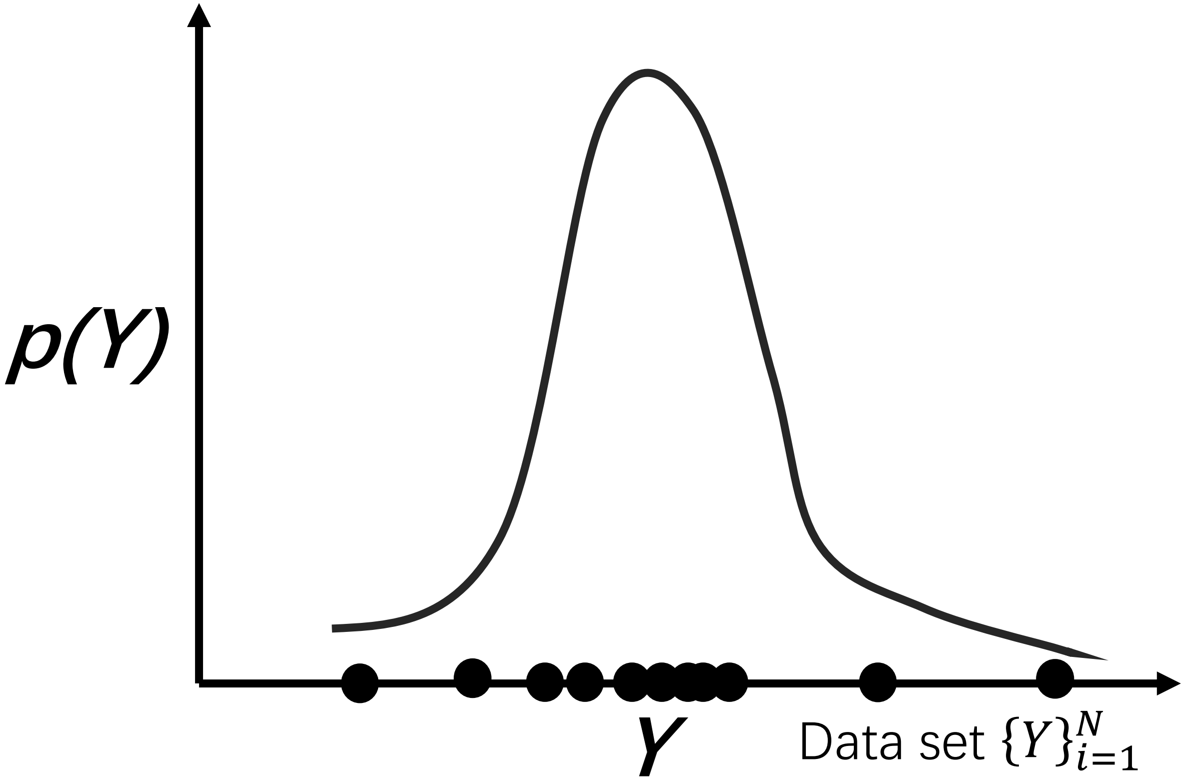
Is this a ML process?

What is the data set $\{X, Y\}_{i=1}^{N}$?
What is the model to be built?
Where is objective function?
What is the learning algorithm (to optimize the objective)?
What is the prediction / interpretation?

B2. ML is everywhere ...



B2. ML is everywhere ...



Find a alternative definition of mean

Consider the square-sum of offset

$$\mathcal{O}(\boldsymbol{y}_{\mathrm{c}}|\{Y\}) \coloneqq \sum_{i=1}^{N} \|\boldsymbol{y}_{\mathrm{i}} - \boldsymbol{y}_{\mathrm{c}}\|_2^2$$

Intuition

If $\boldsymbol{y}_{\mathbf{c}}$ is closer to the 'central' of $\{Y\}$, $\mathcal{O}$ is smaller.

Theorem

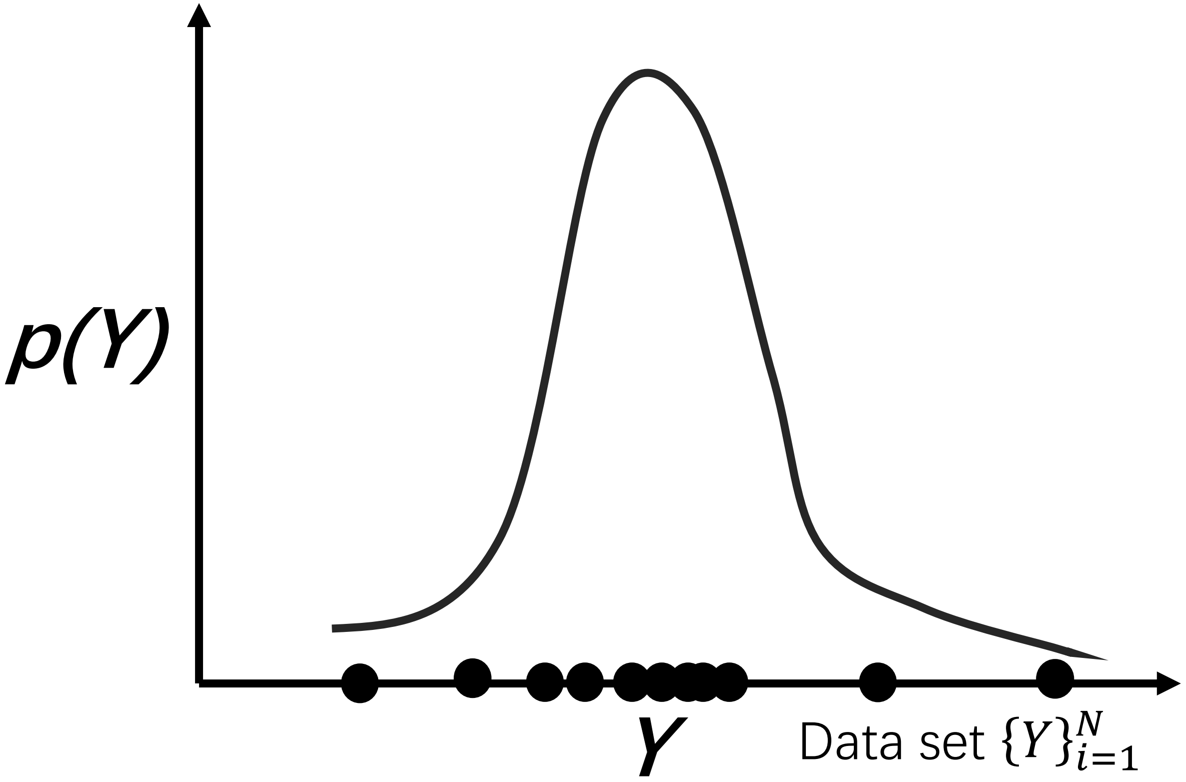
minimizing $\mathcal{O}$ gives sample mean:

$$\widehat{\boldsymbol{y}_{\mathrm{c}}} = \mathrm{argmin}_{\mathbf{y}_{\mathrm{c}}}\{\ \mathcal{O}(\boldsymbol{y}_{\mathrm{c}}|\{Y\})\ \} = \text{sample mean}$$

This description seems redundant, but

- can be easily extended
- raises important questions: why use 2-norm? why use square-sum?

B2. ML is everywhere ...



Extend the objective function

Consider $\boldsymbol{p}$ norm and power $\boldsymbol{q}$

$$\mathcal{O}(\boldsymbol{y}_{\mathrm{c}}|\{Y\},p,q) \coloneqq \sum_{i=1}^{N} \|\boldsymbol{y}_{\mathrm{i}} - \boldsymbol{y}_{\mathrm{c}}\|_p^q$$

Try minimizing the objective

$$\widehat{\boldsymbol{y}_{\mathrm{c}}}(\{Y\},p,q) = \mathrm{argmin}_{\mathbf{y}_{\mathbf{c}}}\{\ \mathcal{O}(\boldsymbol{y}_{\mathrm{c}}|\{Y\},p,q)\ \}$$

Theorem

- $\widehat{\boldsymbol{y}_{\mathrm{c}}}(\{Y\},2,2)$ = mean
- $\widehat{\boldsymbol{y}_{\mathrm{c}}}(\{Y\},2,1)$ = median (That is, 50 % quantile)
- $\widehat{\boldsymbol{y}_{\mathrm{c}}}(\{Y\},2,0)$ = mode

Extension

- From 2-norm to p-norm
- From p-norm to other types of metric, e.g. correlation-based distance, hamming distance

B2. General Procedure of building a ML Algorithm

Data set $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i])\}_{i=1}^{N}$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i])\}_{i=1}^{N}$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C1. Variable Coding Issues

Data set $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i]\}_{i=1}^N$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- Variable coding (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ that is compatible with observations $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i]\}_{i=1}^N$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

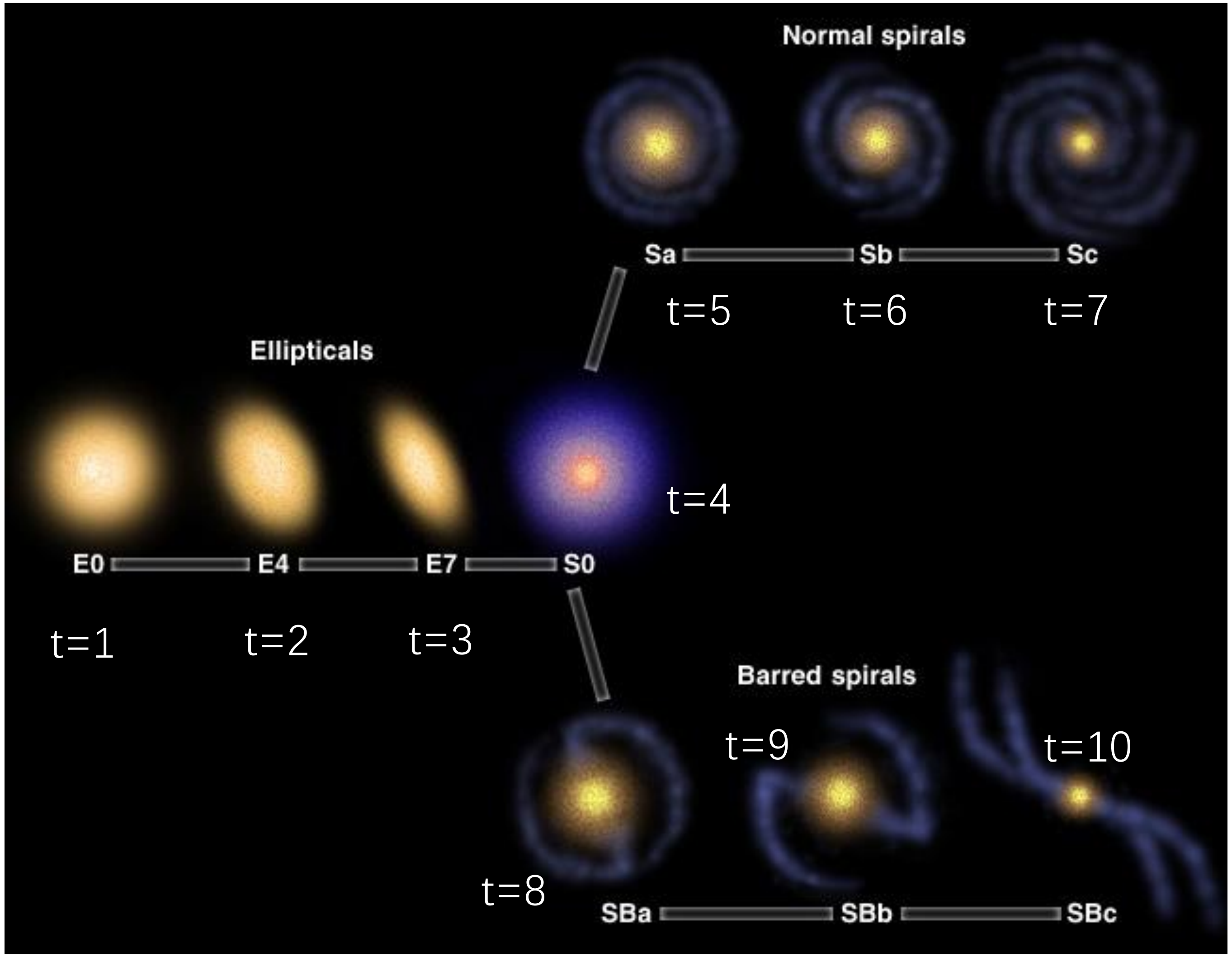
Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C1. Variable Coding Issues



Suppose we have different galaxy types (e.g., E, S, SB, …), we want to study the relation between physical quantities (e.g. SFR) and galaxy type

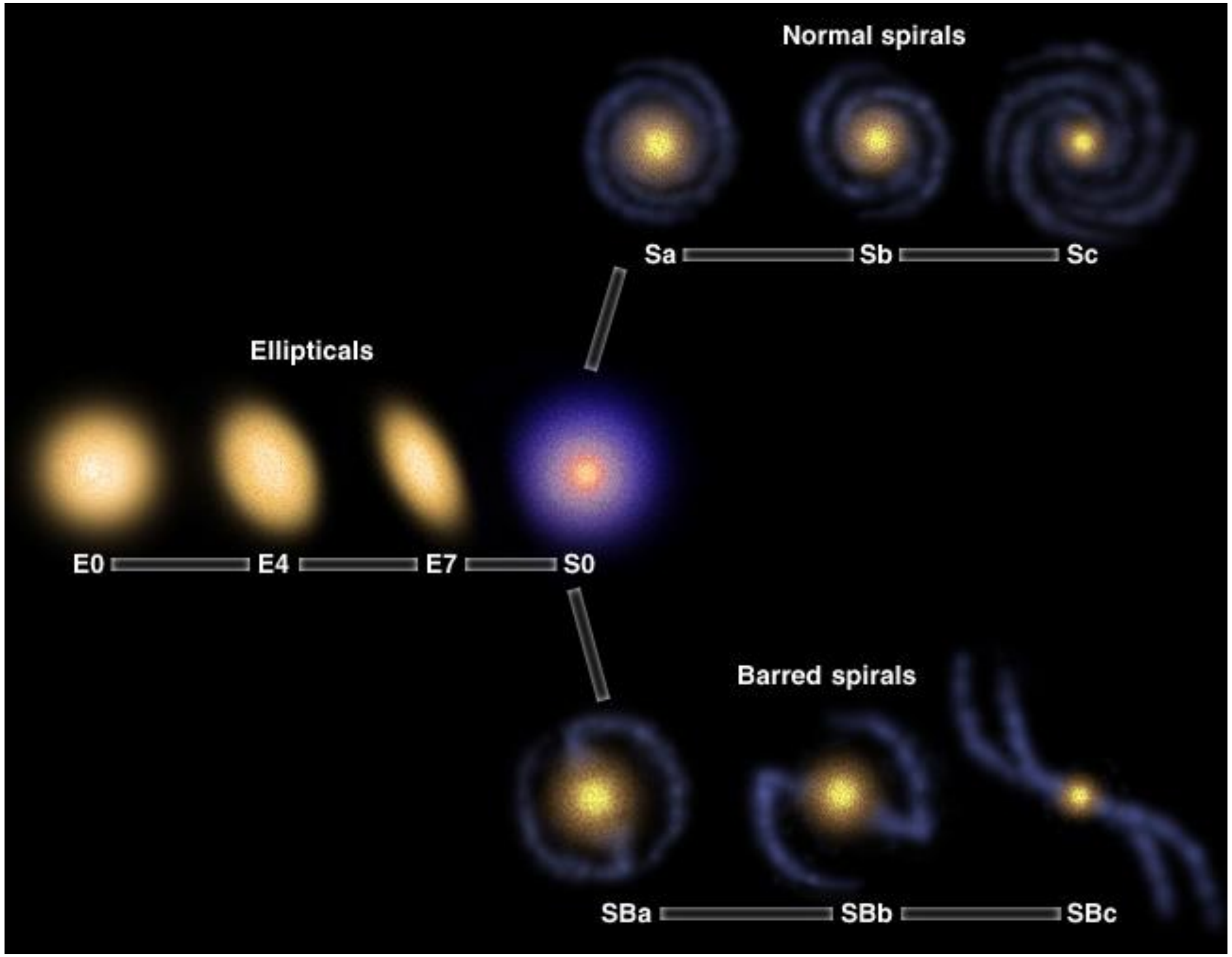
- Just a regression problem of SFR on discrete variables, say, t

  Simplest Model: $SFR = w \times t + b$
- But, how to code the galaxy types

Two considerations

- The order of types is not clear
- The spacing between two types is not fixed

C1. Variable Coding Issues



Better solution: 1-of-K coding

$$E_0 \rightarrow \mathbf{t} = (1,0,0,0,0,0,\dots)$$
$$E_1 \rightarrow \mathbf{t} = (0,1,0,0,0,0,\dots)$$
$$E_2 \rightarrow \mathbf{t} = (0,0,1,0,0,0,\dots)$$
$$\dots$$
$$Sa \rightarrow \mathbf{t} = (0,0,0,1,0,0,\dots)$$
$$Sb \rightarrow \mathbf{t} = (0,0,0,0,1,0,\dots)$$
$$\dots$$
$$SBa \rightarrow \mathbf{t} = (0,0,0,0,0,1,\dots)$$

And we do the regression on a higher dimensional space

$$\mathrm{SFR} = \mathbf{w}^T\mathbf{t} + b$$

Pros and cons

- The order and spacing between variables are disappeared
- But, a higher dimension feature space is introduced.

C2. Error Function

Data set $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i]\}_{i=1}^N$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(\boldsymbol{x}_i, [\boldsymbol{y}_i]\}_{i=1}^N$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C3. The Bias Variance Trade-off

Data set $\{(x_i, [y_i]\}_{i=1}^N$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(x_i, [y_i]\}_{i=1}^N$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

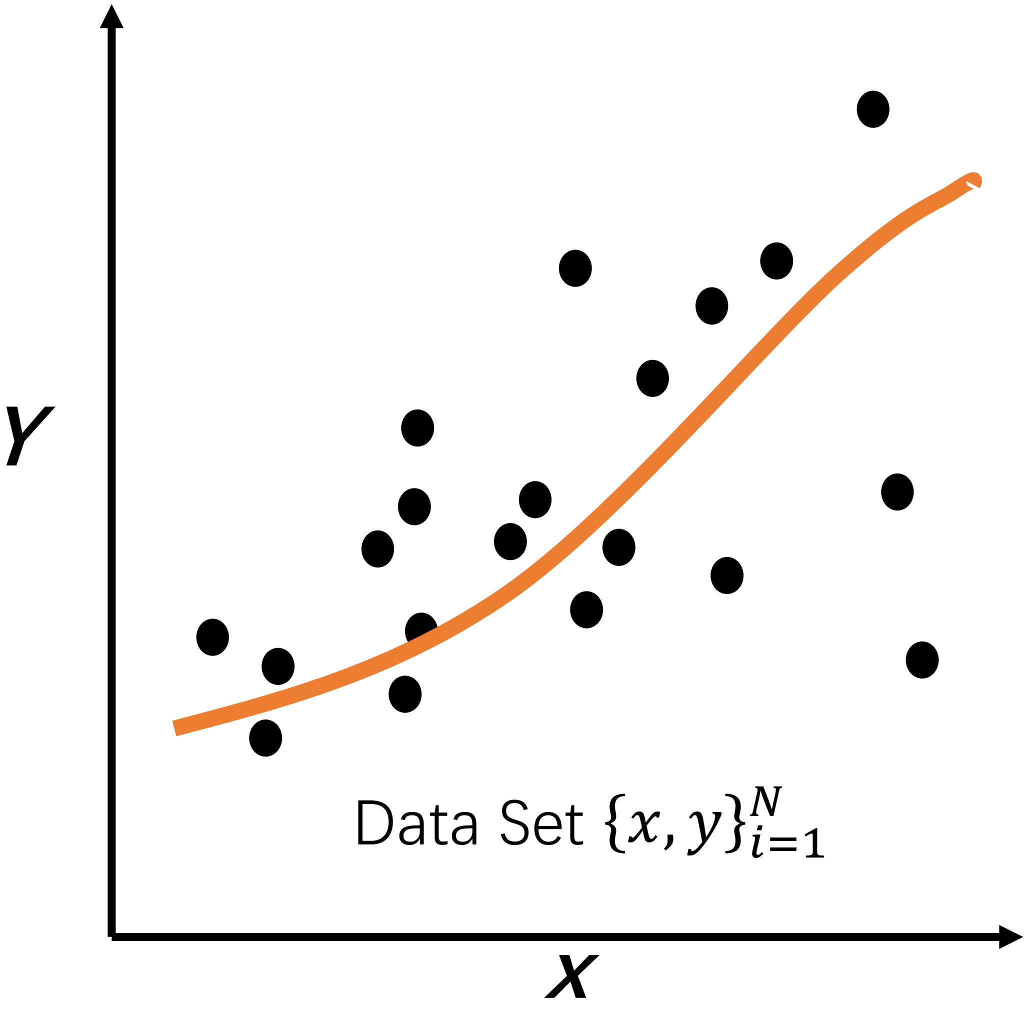
- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C3. The Bias Variance Trade-off

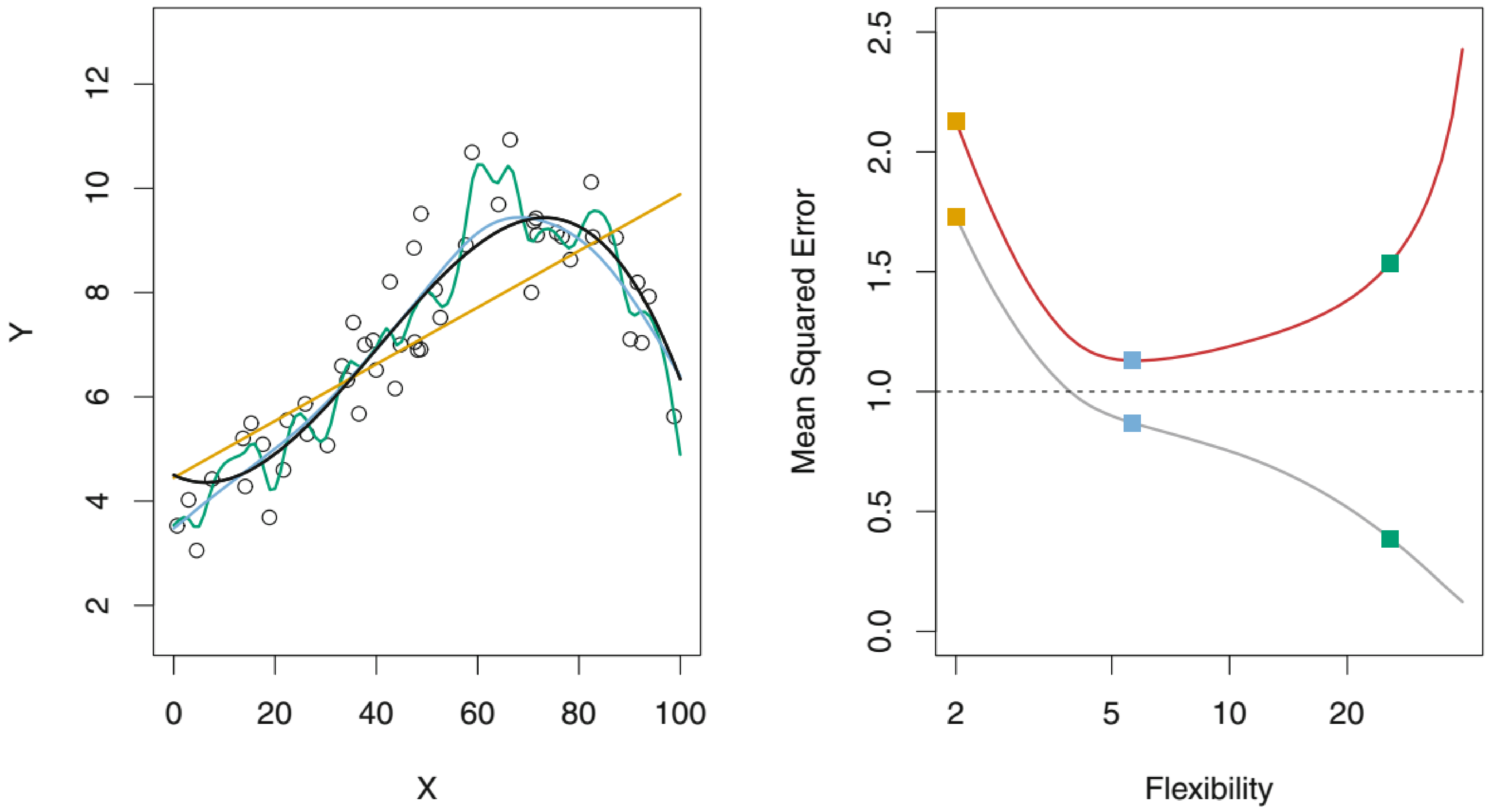
Review: Least Square Linear Regression

- Model = (Generalized) Linear model, probabilistic discriminative
$$p(y|x) = N(t(x;\mathbf{w})|\mu,\sigma^2)$$
$$t(x;\mathbf{w}) = \sum_{m=0}^{p} w_m x^m$$
- Objective = RSS error function
$$\mathcal{O}(\mathbf{w}|\,\{x,y\}_{i=1}^N) := \sum_i \{y_i - t(x_i;\mathbf{w})\}^2$$
- Optimization = analytical
$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}}\{\,\mathcal{O}(\mathbf{w}|\,\{x,y\}_{i=1}^N)\}$$
- Goodness of fitting = test MSE
$$MSE = \frac{1}{\#(\text{test set})}\sum_{j\in\text{test set}}\{y_j - t(x_j;\hat{\mathbf{w}})\}^2$$

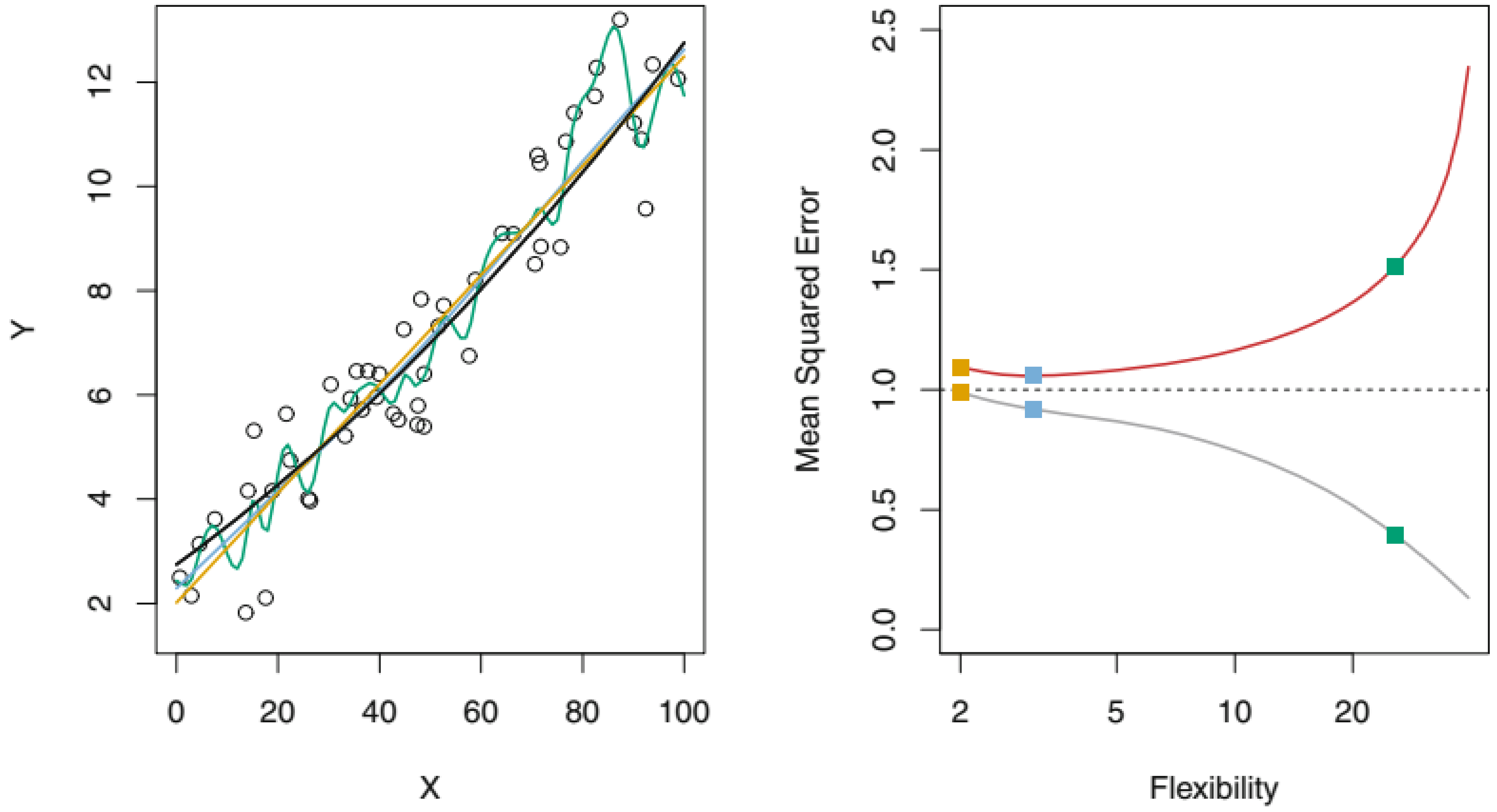
Problem: How to choose $\boldsymbol{p}$?



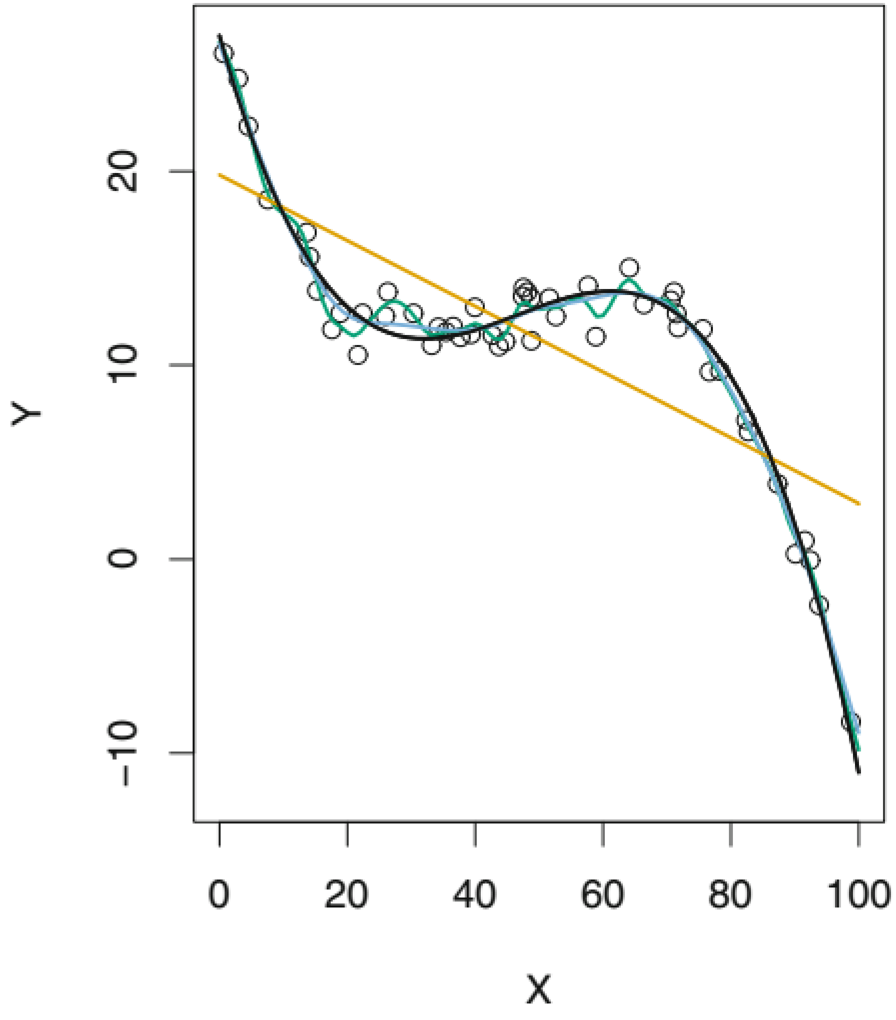
C3. The Bias Variance Trade-off

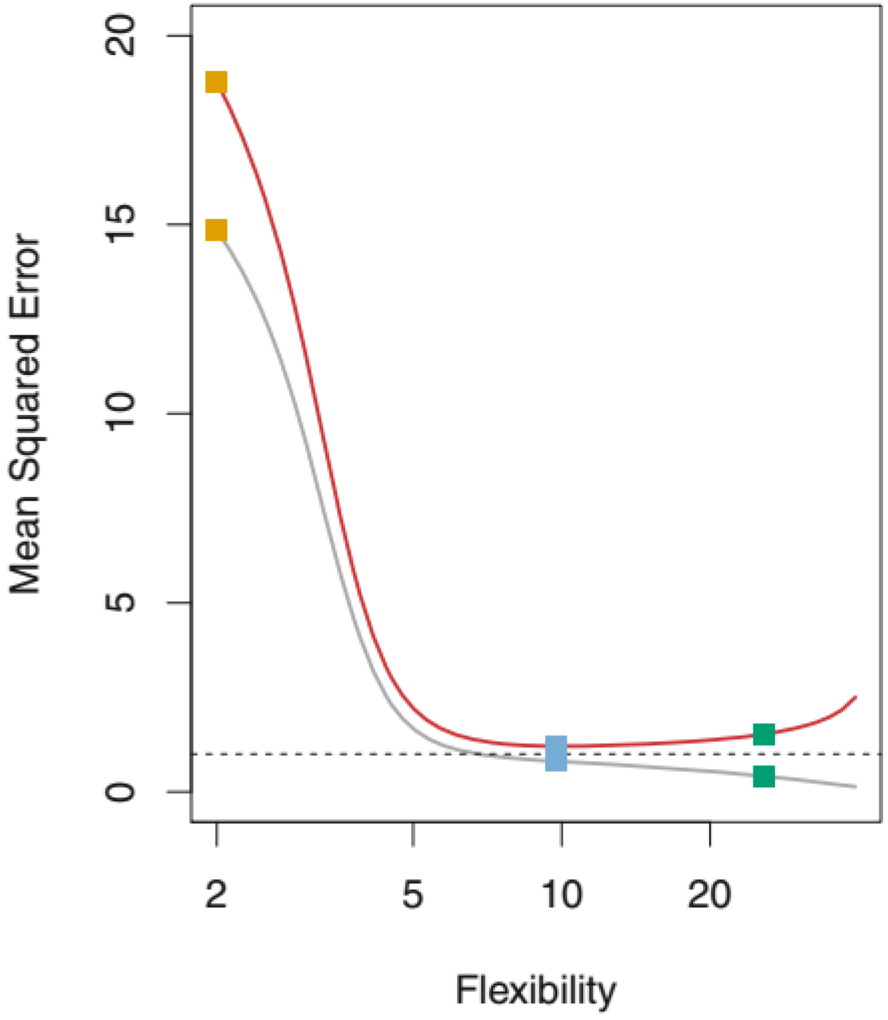


C3. The Bias Variance Trade-off



C3. The Bias Variance Trade-off



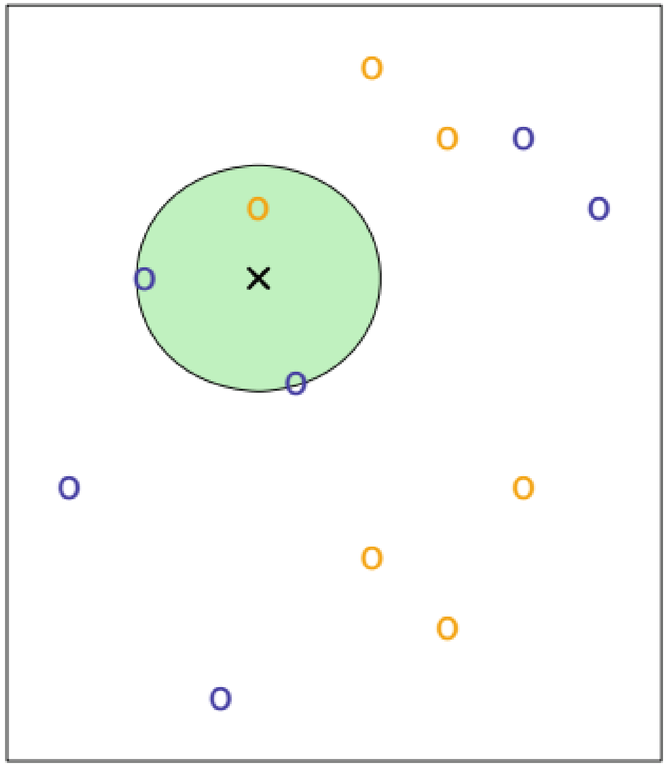


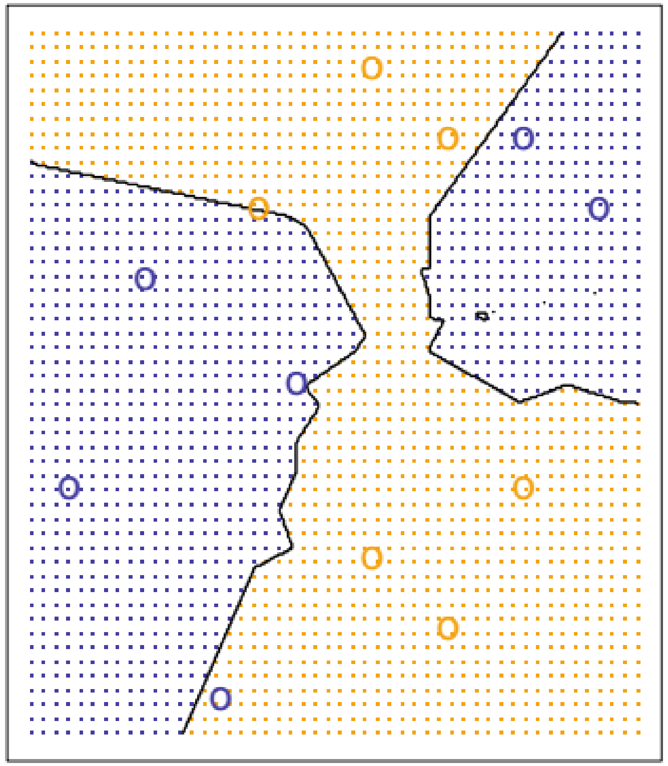
C3. The Bias Variance Trade-off

Review: KNN classification

- Model: Nonlinear, Discriminant
  $$y\left(\mathbf{x}\middle|\{\mathbf{x},y\}_{i=1}^{N}\right) = \operatorname{argmax}_{C_k} \sum_{j\in \text{Neighbor}_{\mathbf{x}}} I(y_j = C_k)$$
- Goodness of fitting = test Error Rate
  $$Error\ Rate = \frac{1}{\#(\text{test set})} \sum_{j\in \text{test set}} I\left\{y\left(\mathbf{x}_j\middle|\{\mathbf{x},y\}_{i=1}^{N}\right) \neq y_j\right\}$$

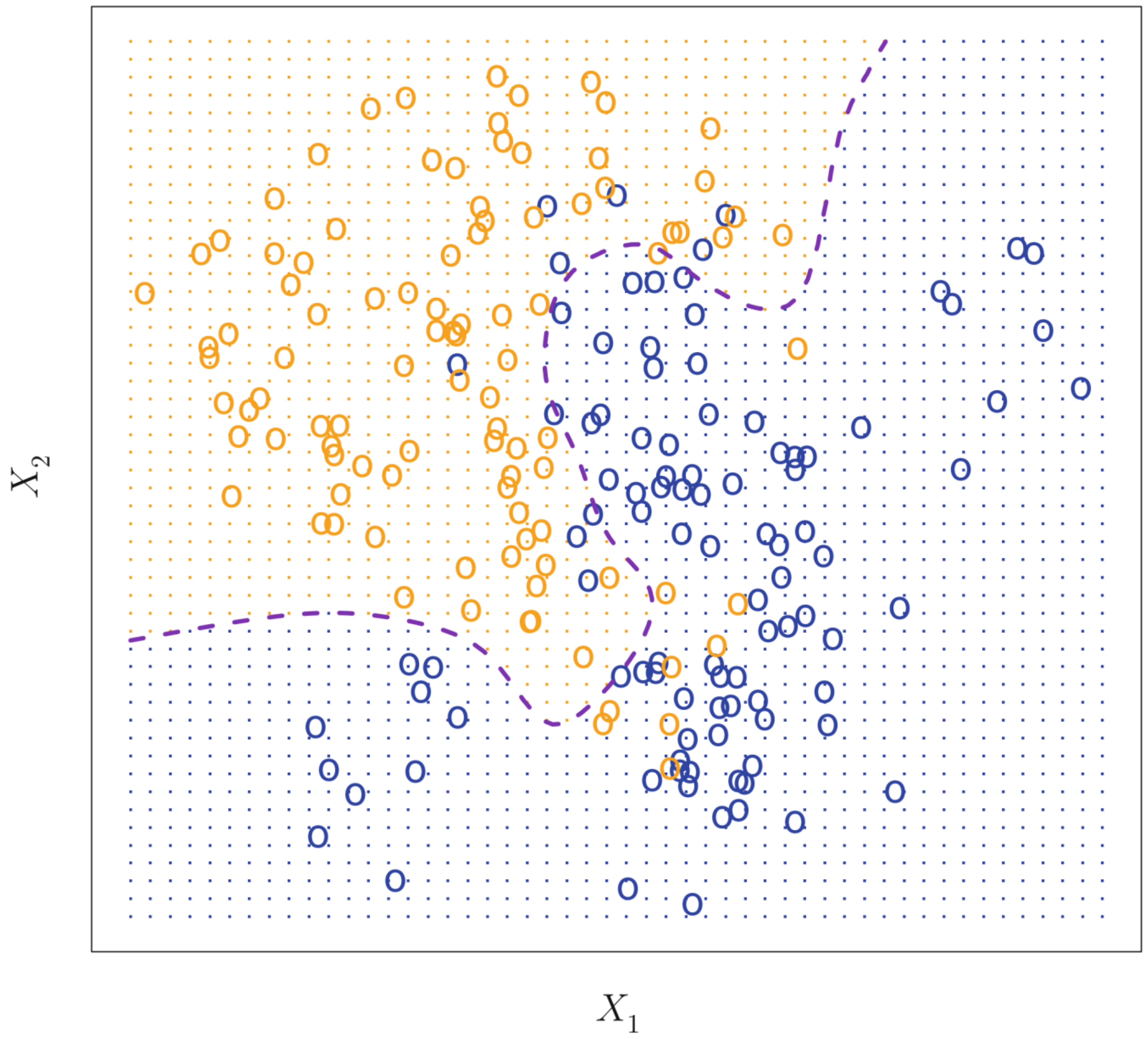
Problem: How to choose k ?



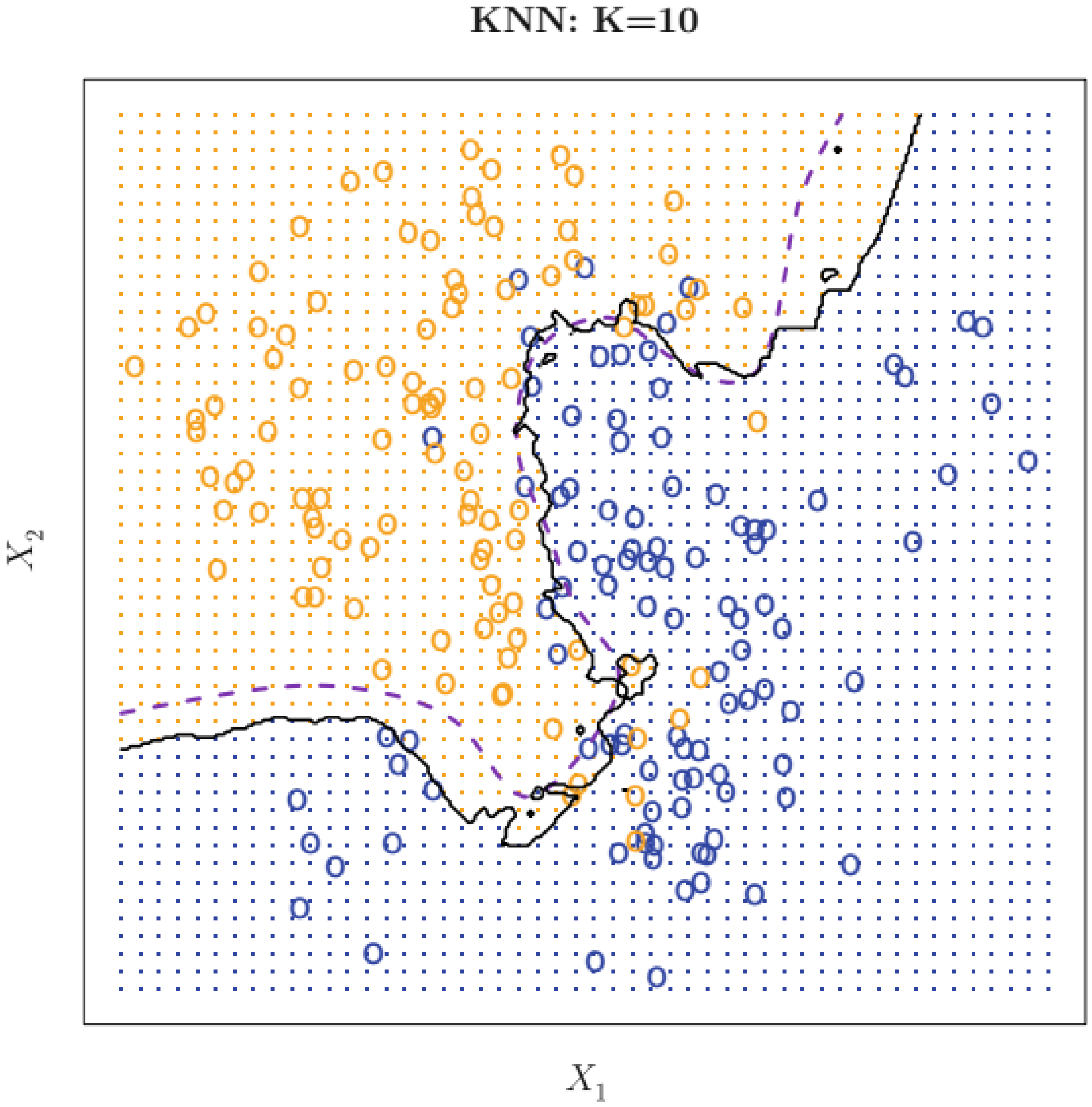


Data Set $\{\mathbf{x},y\}_{i=1}^{N}$
$y \in \{C_1, C_2, \dots\}$

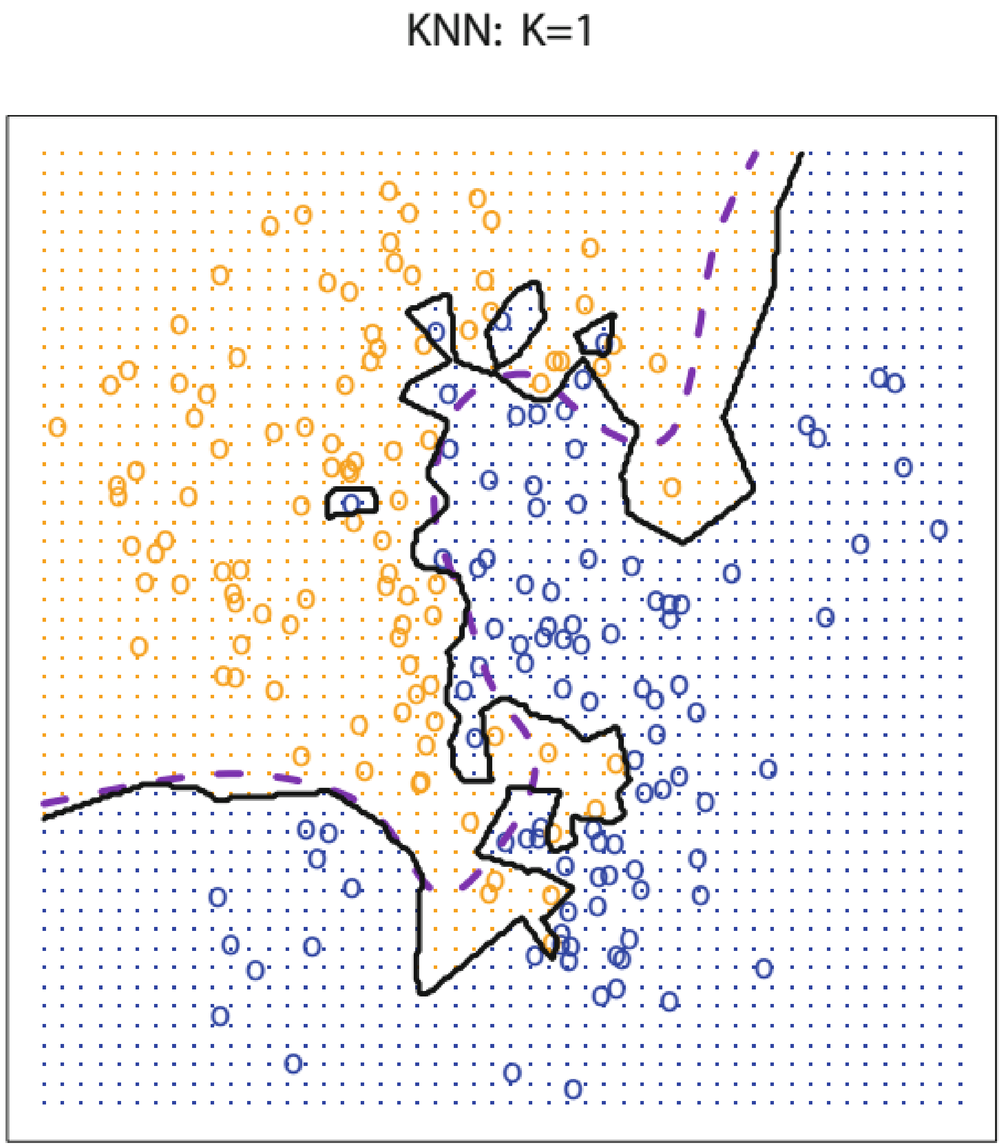
C3. The Bias Variance Trade-off

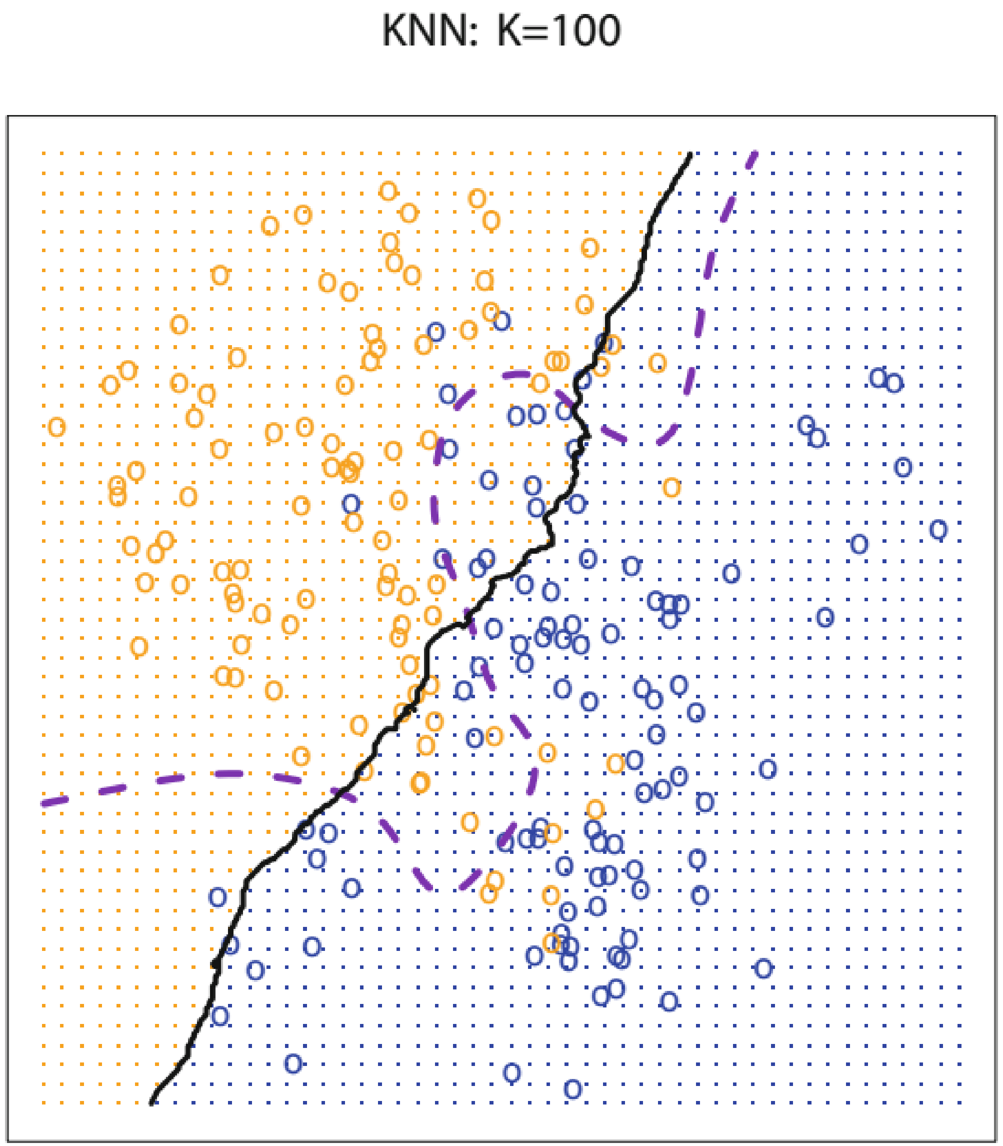


C3. The Bias Variance Trade-off

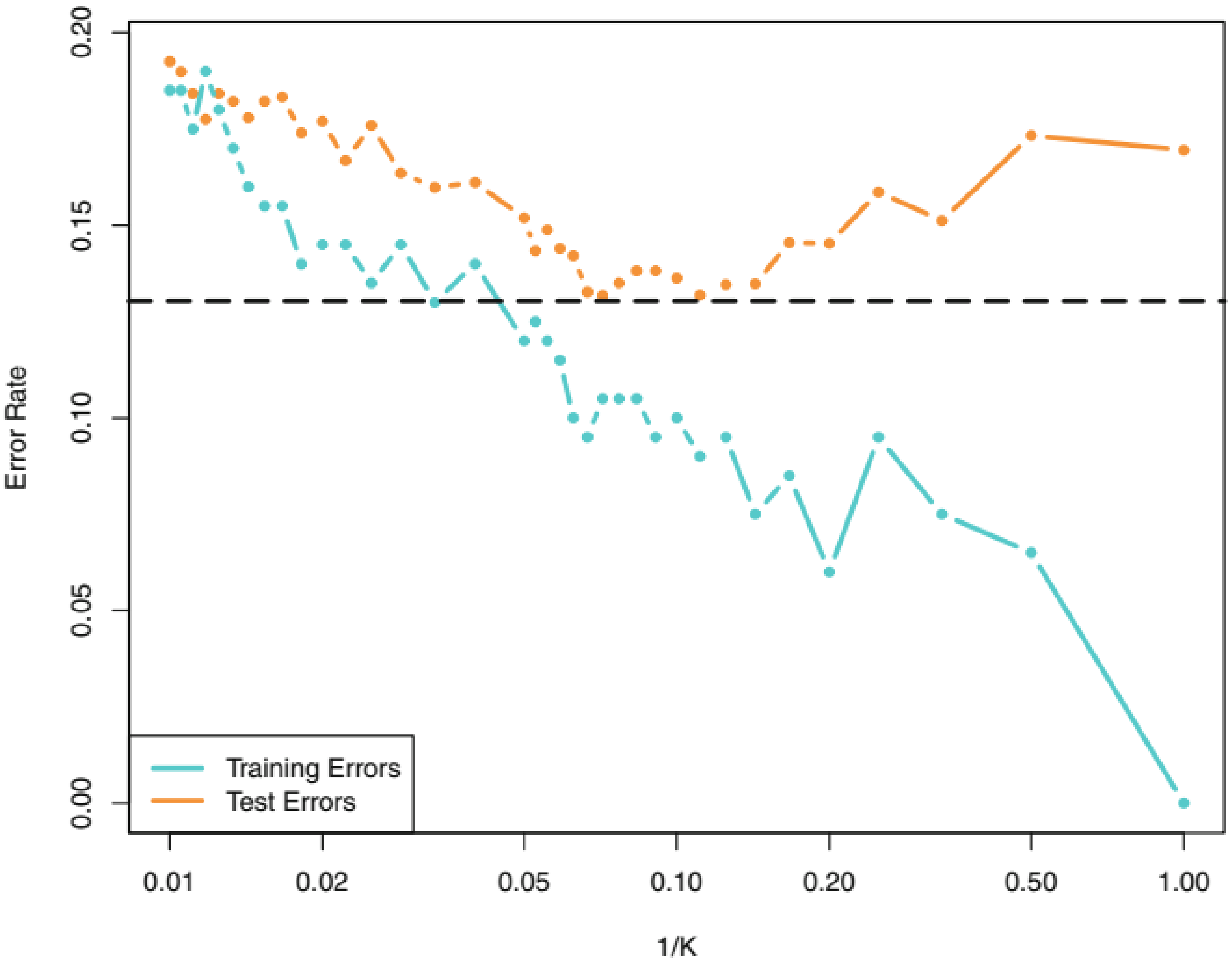


C3. The Bias Variance Trade-off





C3. The Bias Variance Trade-off



C3. The Bias Variance Trade-off

Model set

$$F = \{t(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\mathbf{Dom}}$$

e.g. for generalized linear regression

$$F = \{t(\mathbf{x};\mathbf{w})\}_{\mathbf{w}\in\Re^p}$$
$$= \text{all p order polynomials}$$

After trained with data set D, we pick a model from the set to minimize to training error

$$\hat{t}\,(\,\boldsymbol{\theta}(D))$$

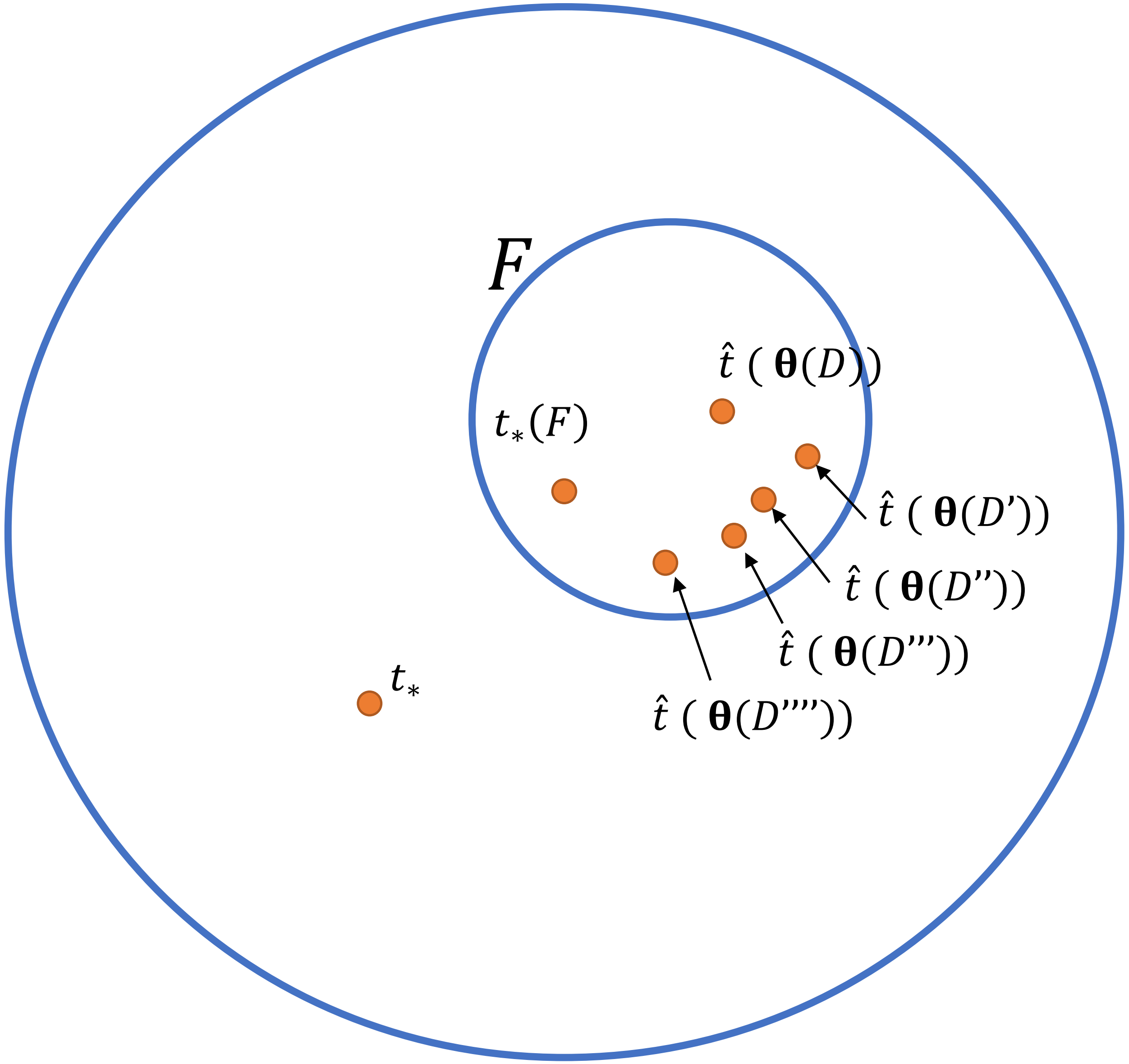
Consider minimizing the test error $E\{t\}$(e.g., MSE for regr., Error rate for classif.)

Best model in set F

$$t_*(F) = \text{argmin}_{t\in F}E\{t\}$$

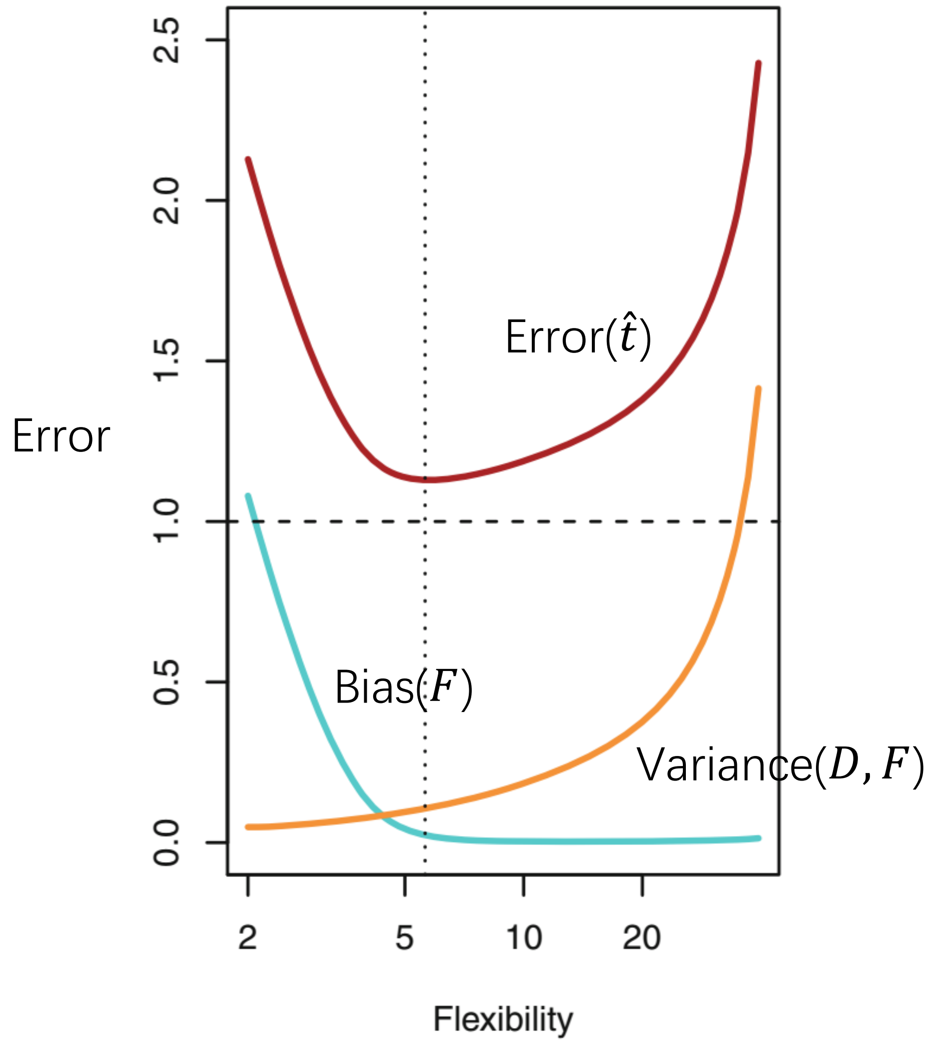
Best model in all possible model (i.e., the Model Universe)

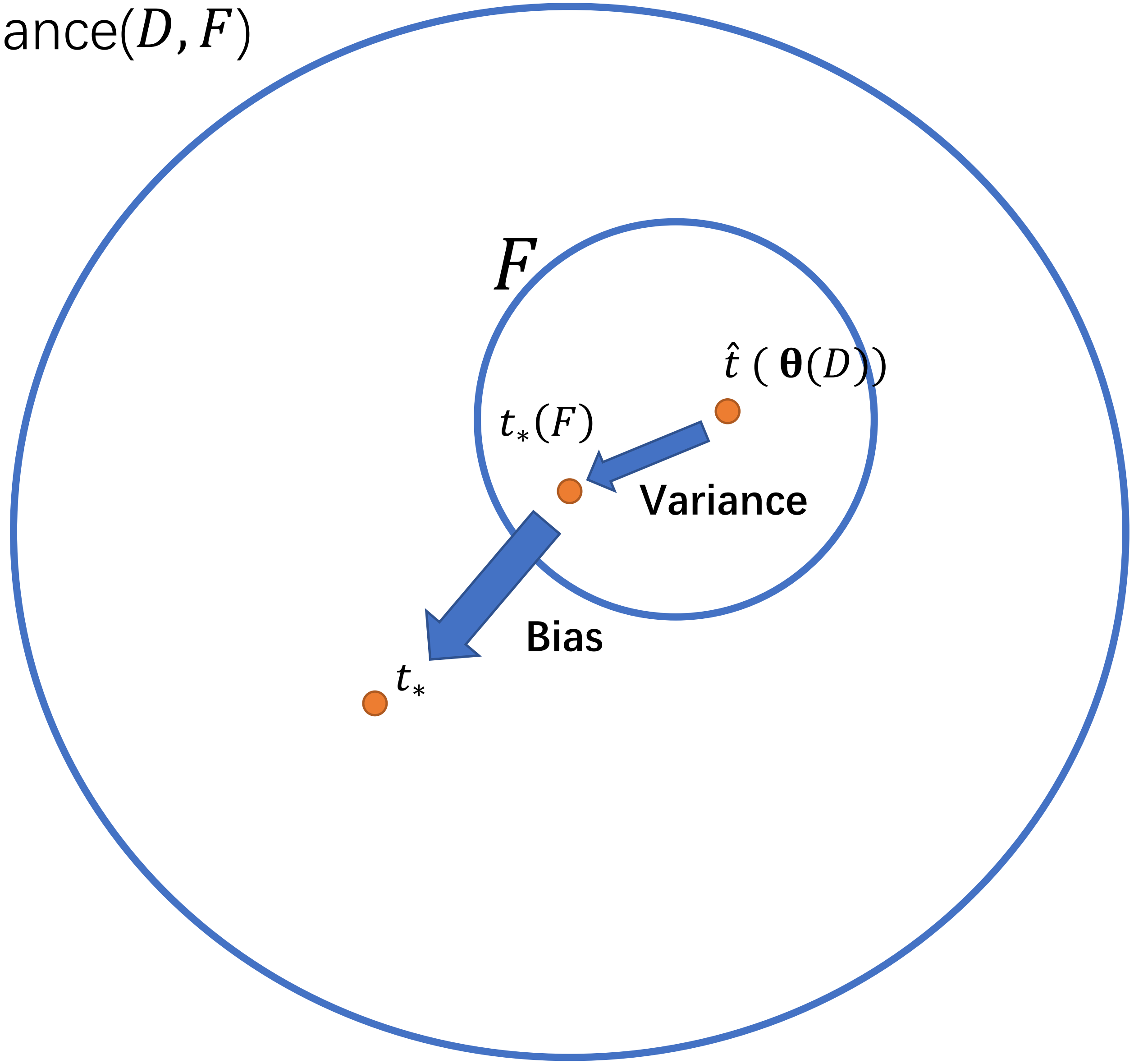
$$t_* = \text{argmin}_t E\{t\}$$



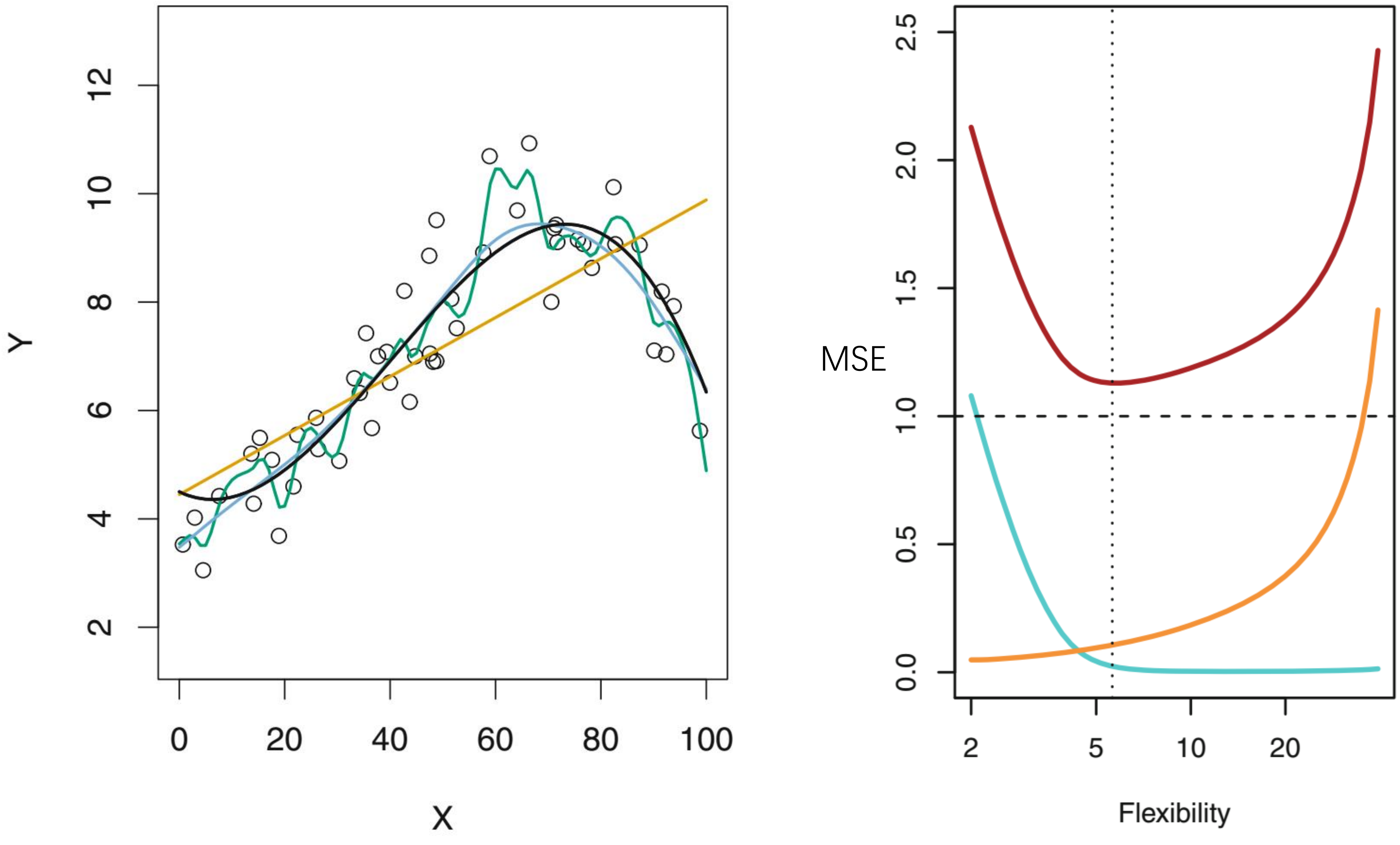
C3. The Bias Variance Trade-off

$$\text{Error}(\hat{t}) = \text{Bias}(F) + \text{Variance}(D, F)$$

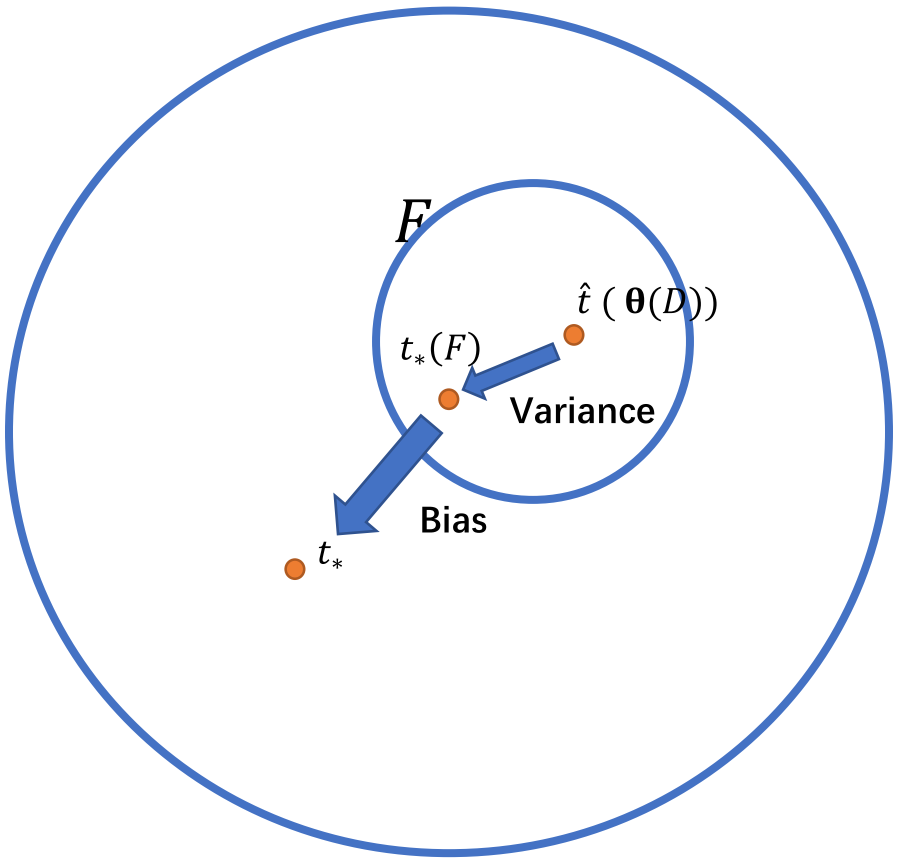




C3. The Bias Variance Trade-off



C3. The Bias Variance Trade-off

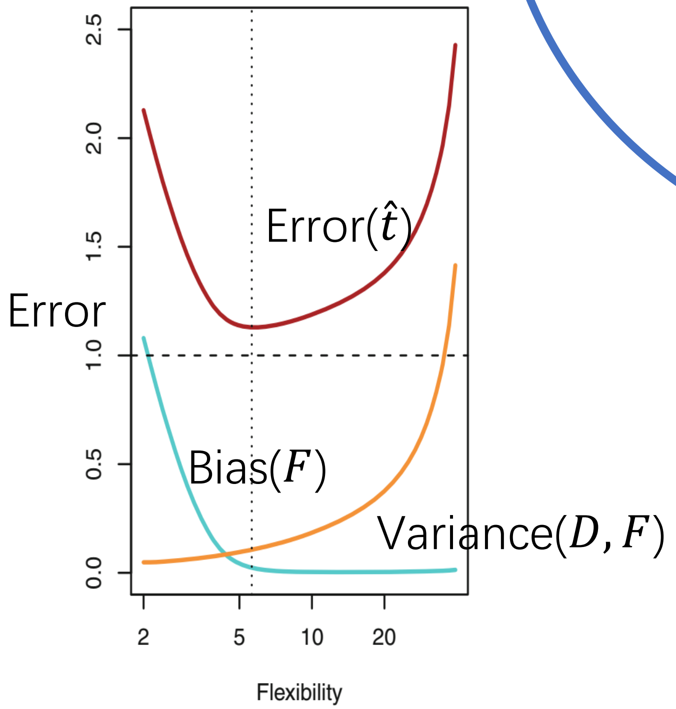


Always remember

- The competition between bias and variance makes the error a U-shape curve
- When the variance dominates (i.e., a large model is chosen), over-fitting emerges
- Large data set does not guarantee better fitting, if the model set F is wrong.

Many method can be used to find the approximate minimum of the U-curve

- Analytical based methods (p-value)
- Analytical approximation (AIC, BIC, …)
- Sampling partition (cross validation)
- Model Regularization (Bayes prior, MLE regularization)



$$\text{Error}(\hat{t}) = \text{Bias}(F) + \text{Variance}(D, F)$$

C3. The Prediction-interpretation Trade-off

Data set $\{(\boldsymbol{x_i}, [\boldsymbol{y_i}]\}_{i=1}^{N}$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ **that is compatible with observations** $\{(\boldsymbol{x_i}, [\boldsymbol{y_i}]\}_{i=1}^{N}$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

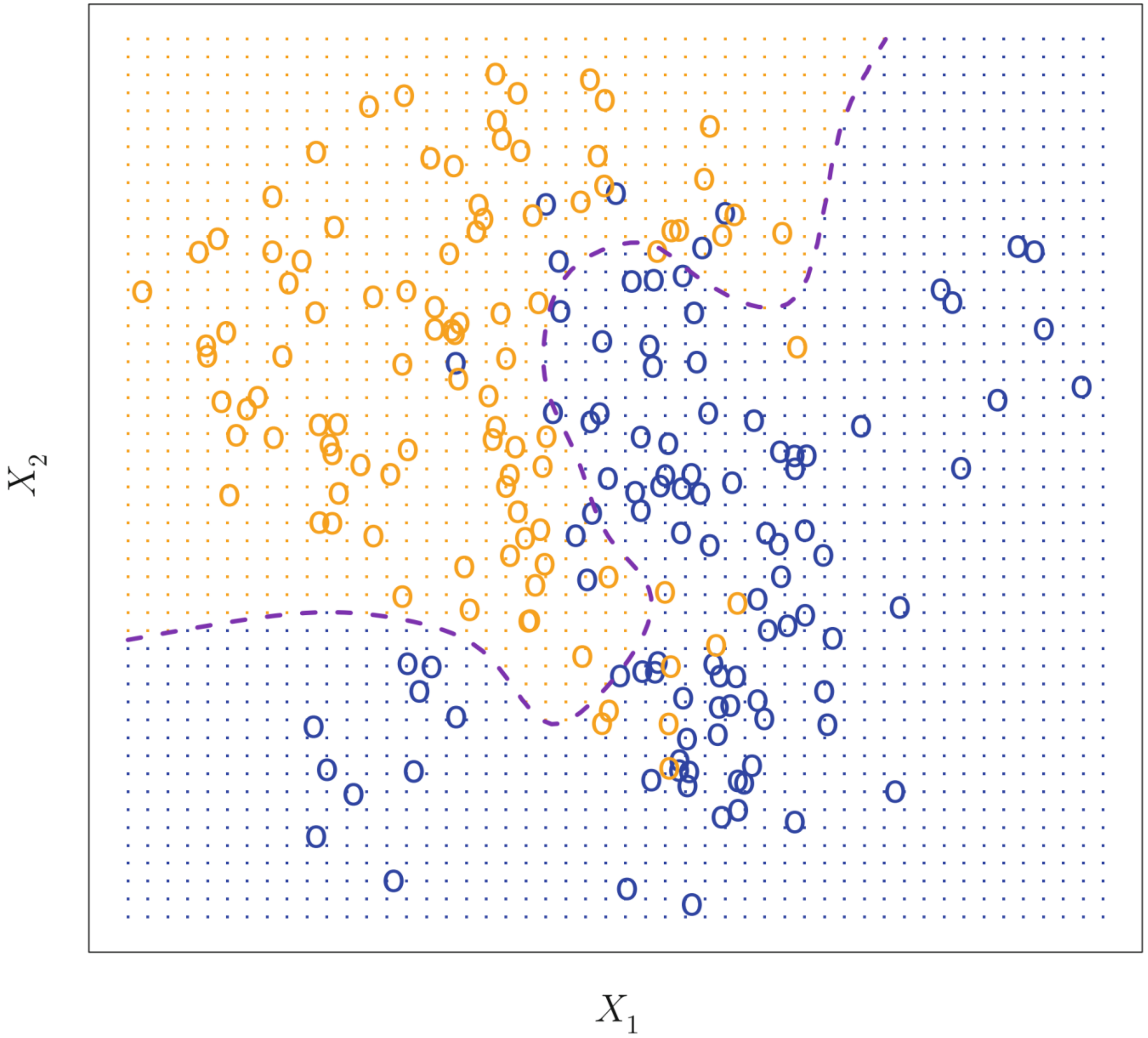
Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C3. The Prediction-interpretation Trade-off

Recall the classification problem, we have three types of methods

- Generative model (LDA, QDA, …)
- Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
- Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)



C3. The Prediction-interpretation Trade-off

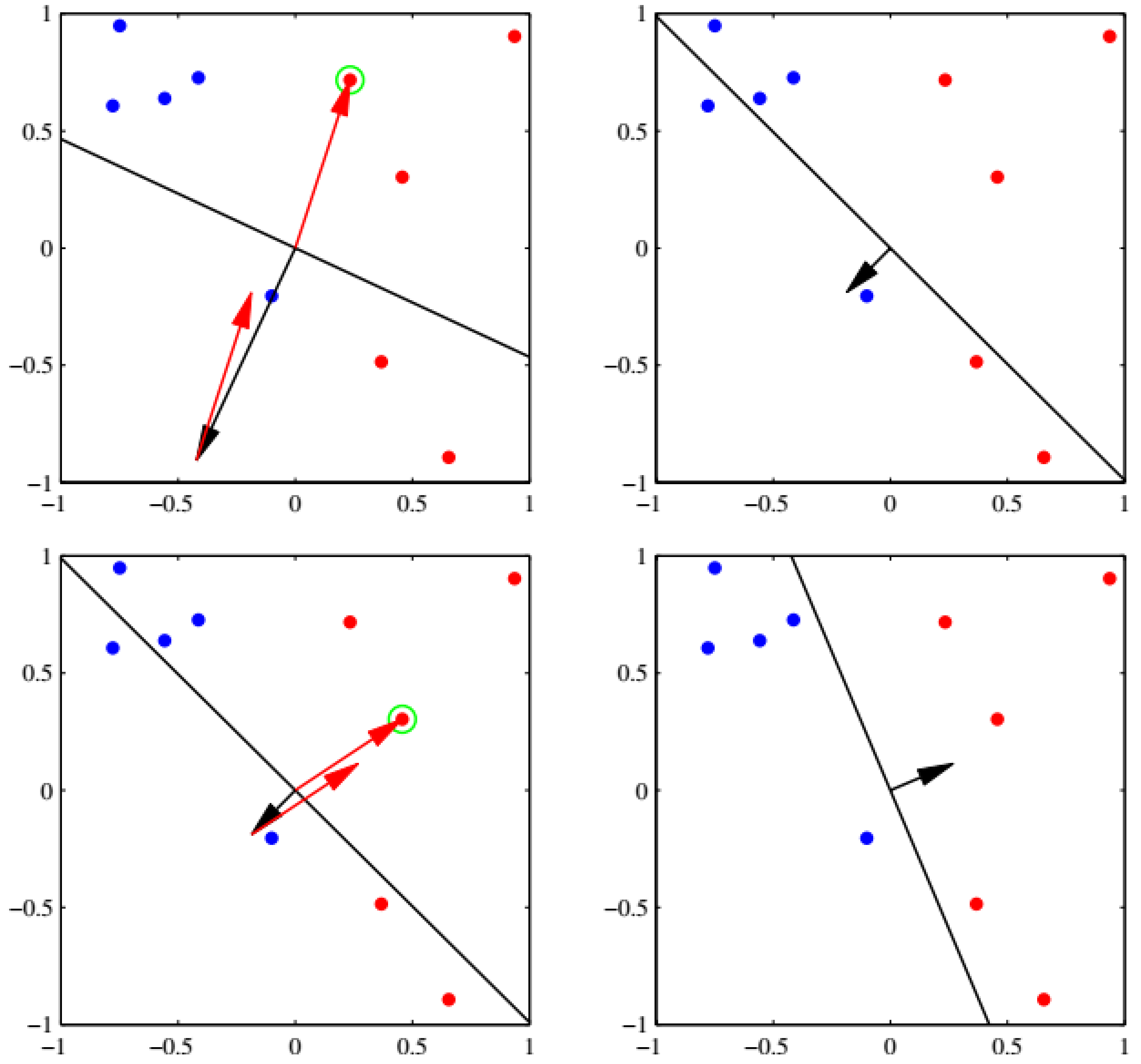
Single layer perceptron

Model: Linear Decision Boundary

$$y = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$$

Objective: Sum of point-plane distance of error-assigned points

Optimizing: stochastic gradient descent



C3. The Prediction-interpretation Trade-off

Logistic Regression

Model: conditional class probability

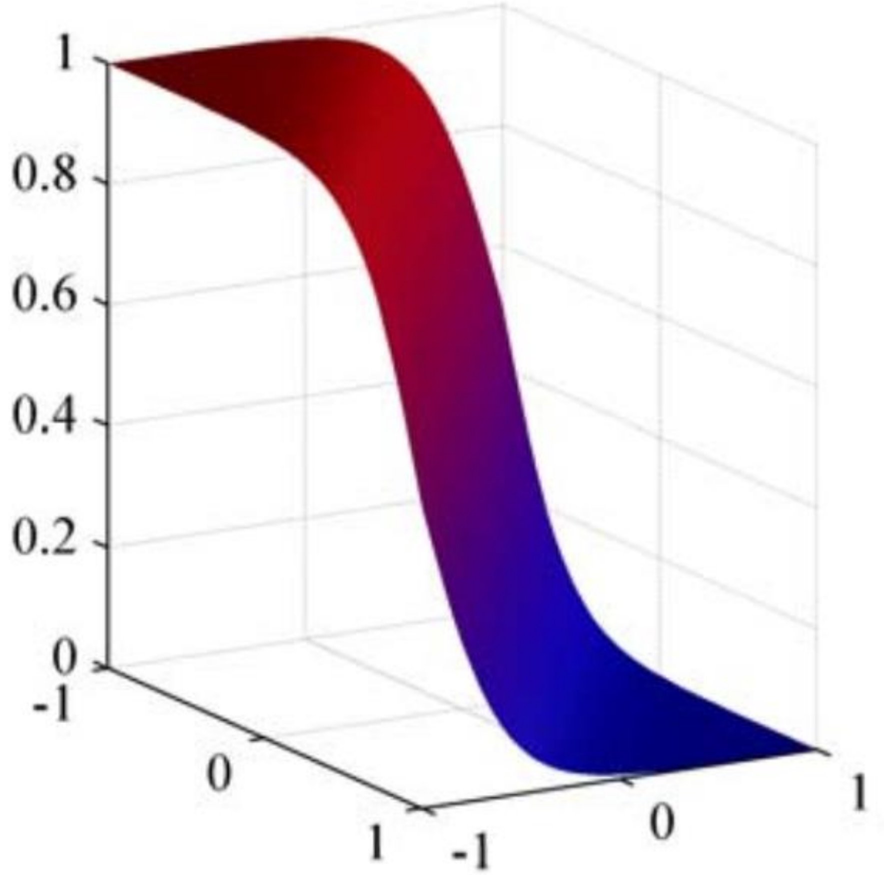
$$p(y = C_1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$p(y = C_2|\mathbf{x}) = 1 - p(y = C_1|\mathbf{x})$$

Objective: Likelihood function

Optimizing: numeric

Advantage: Allow decision



C3. The Prediction-interpretation Trade-off

Linear Discriminative Analysis (LDA)

Model: Posterior distribution through Bayes

$$p(y = C_k|\mathbf{x}) = \frac{p(y = C_k)p(\mathbf{x}|y = C_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}|\mathbf{y} = C_k) = N(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{k}}, \Sigma\ )$$

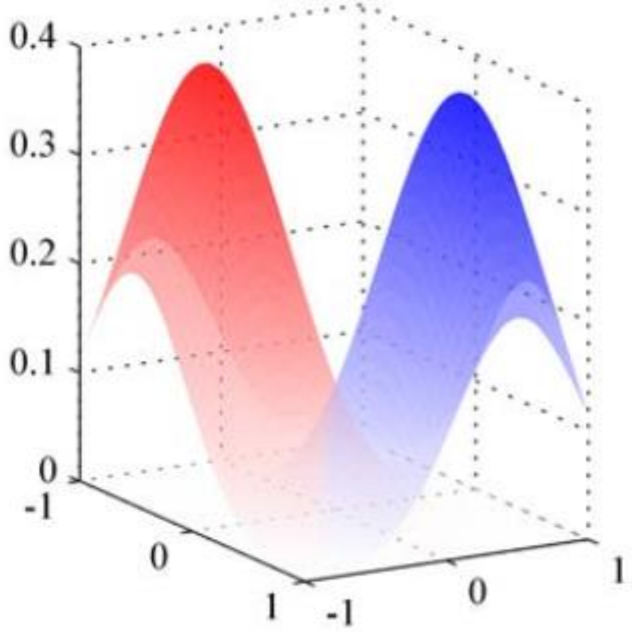
Objective: product of posteriors

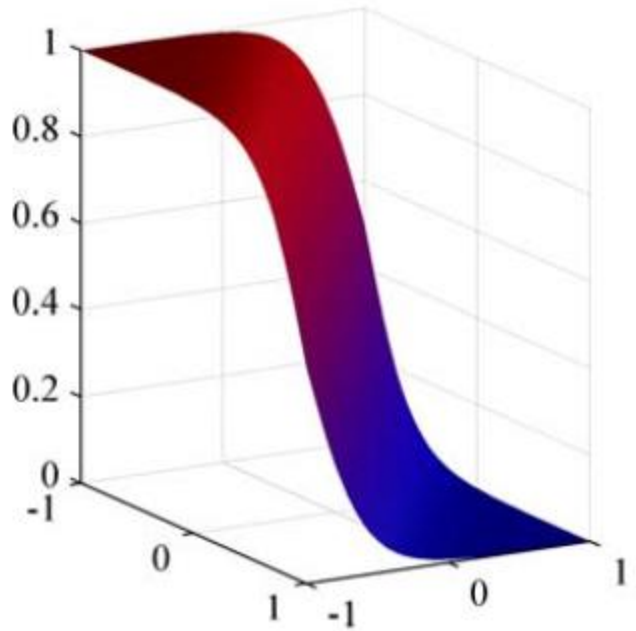
Optimizing: analytical

Prediction: $y = C_k$ if $p(y = C_k|\mathbf{x})$ is largest

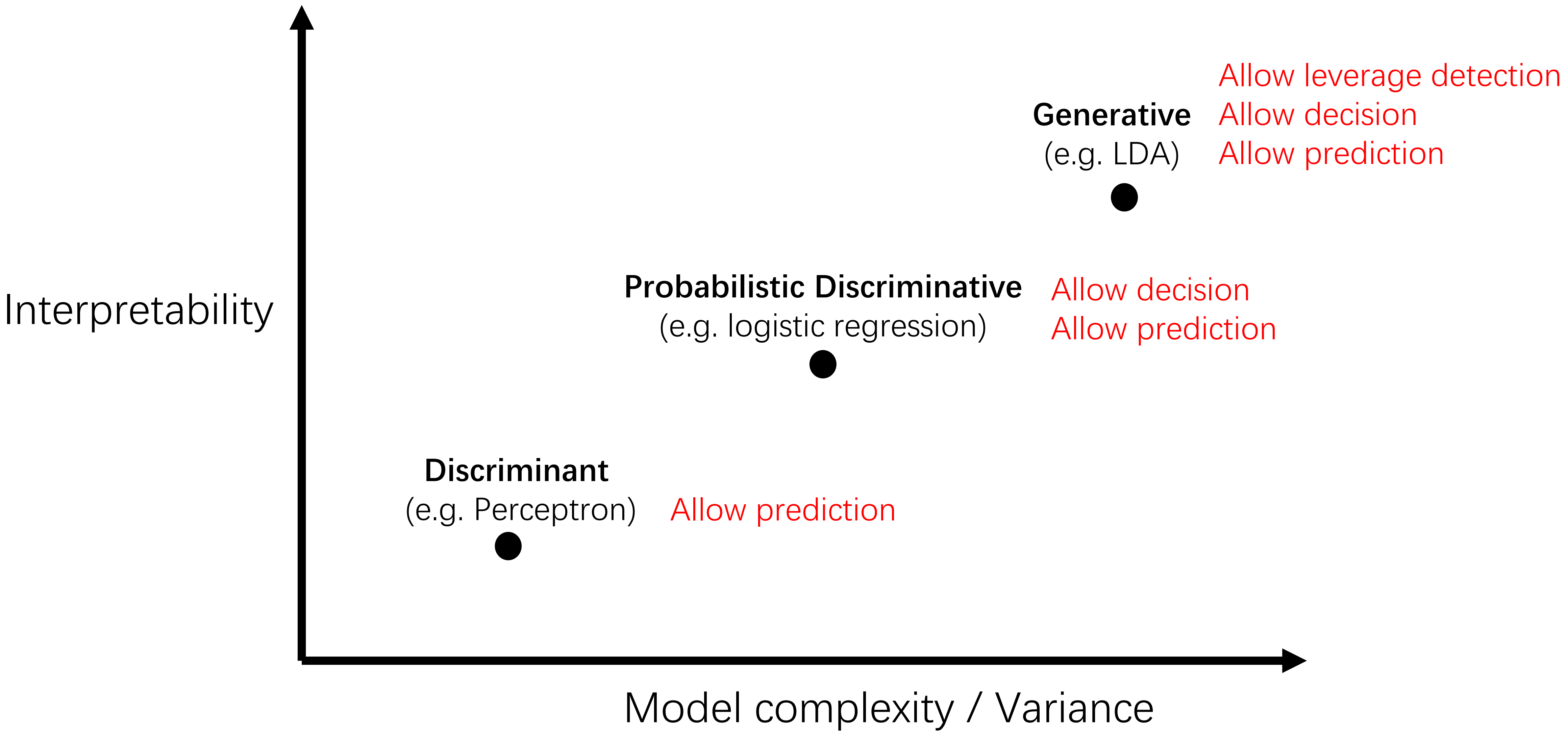
Advantage:

Allow leverage detection





C3. The Prediction-interpretation Trade-off



C4. Non-linearity

Data set $\{(x_i, [y_i]\}_{i=1}^{N}$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(x_i, [y_i]\}_{i=1}^{N}$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

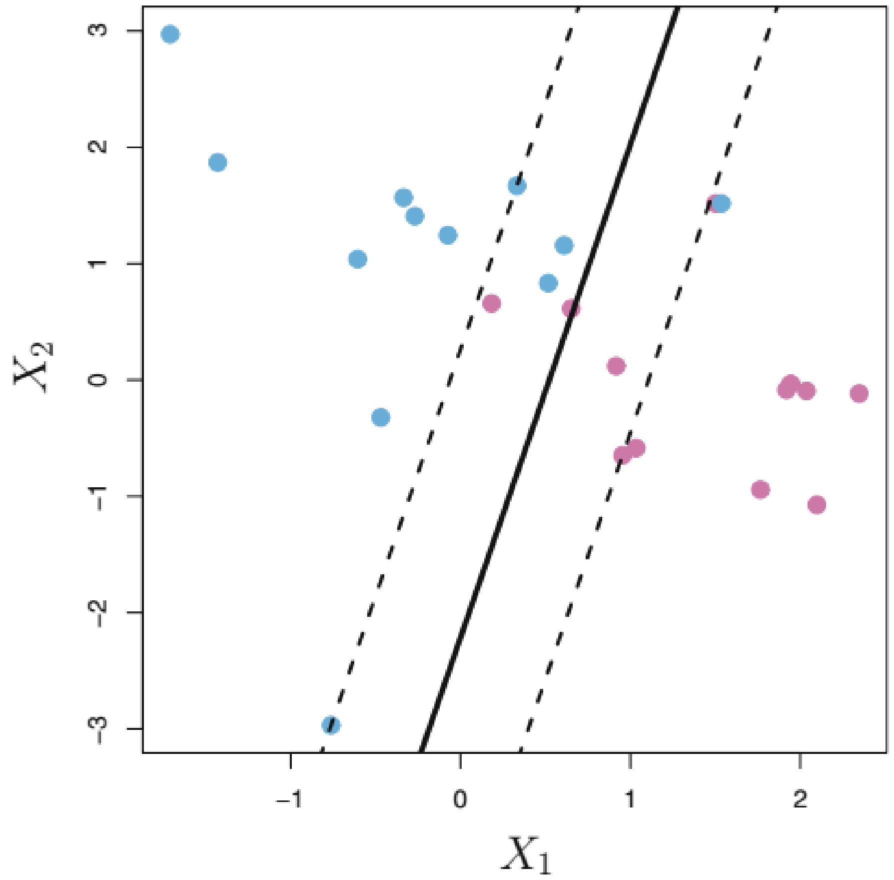
- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C4. Non-linearity

Linear separable problem

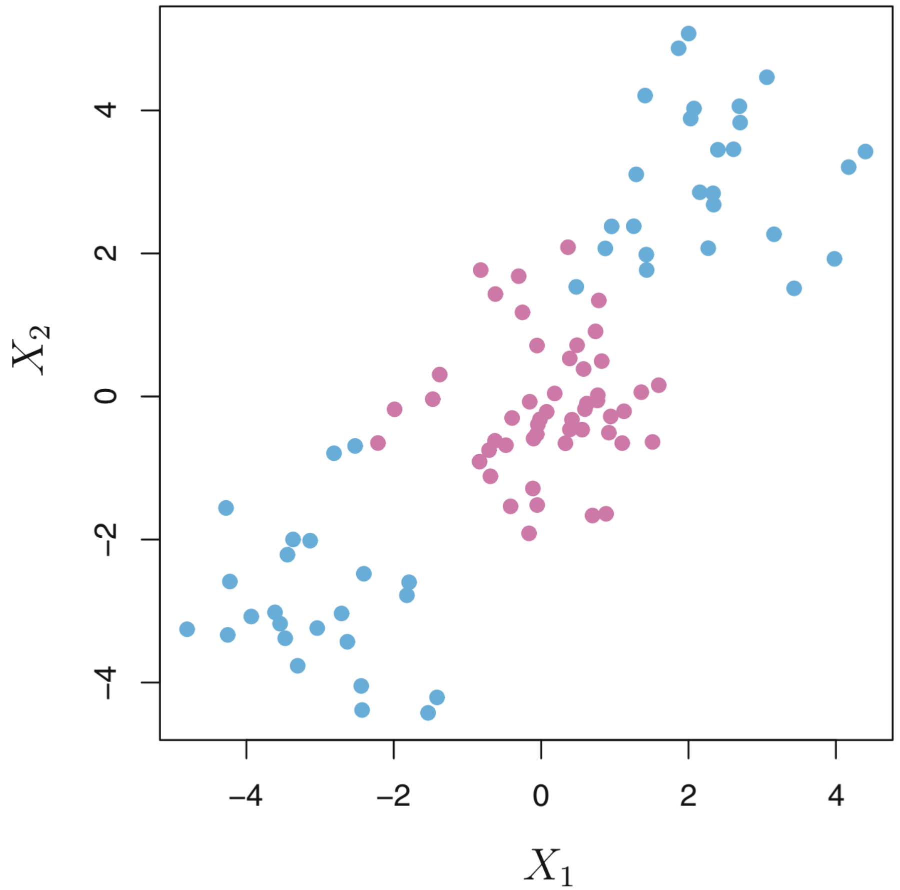
Separation plane : $w_1X_1 + w_2X_2 = 0$

e.g. Support Vector Classifier (soft SVM), logistic, LDA, perceptron …



Not linear-separable

How to build the model?

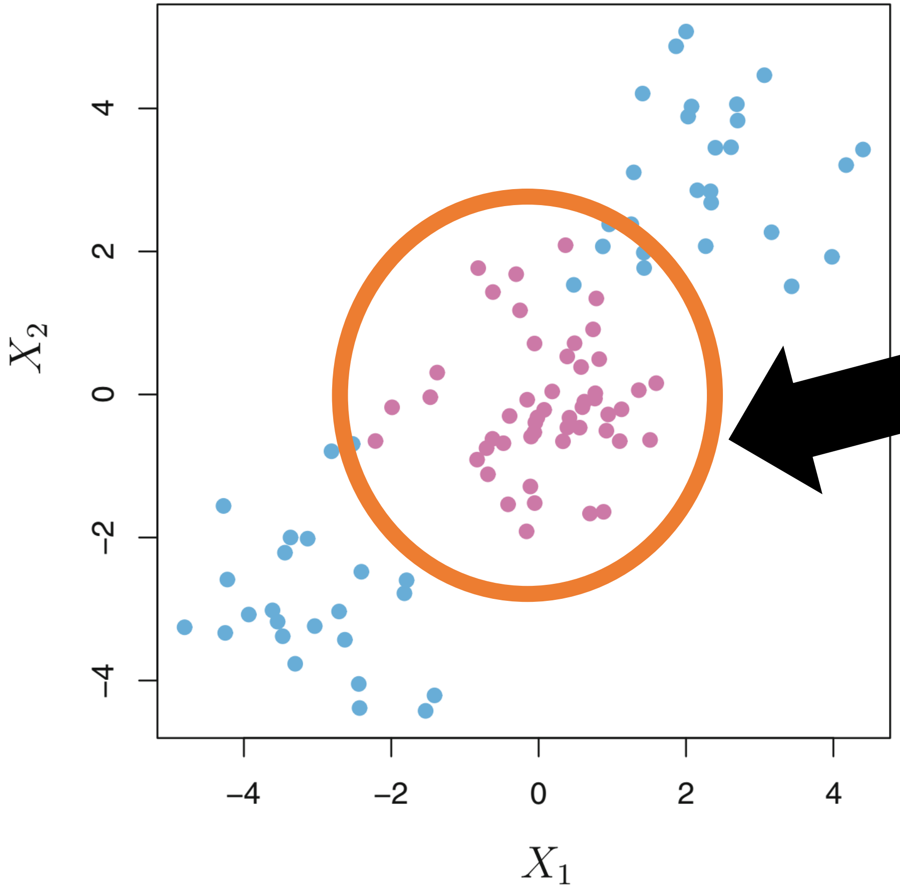


C4. Non-linearity

Not linear-separable

How to build the model?

- Nonlinear transformations
- Kernel based-methods
- Neural Networks



Original data set:

$$\{(X_1, X_2)_i\}_{i=1}^N$$

Model:

$$y = \mathrm{sign}(w_1 X_1 + w_2 X_2)$$

Parameters to be found:

$$w_1, w_2$$

Optimization problem:

e.g. max-margin at 2-dim space

Data set with extended dimension:

$$\{(X_1, X_2, X_1^2, X_2^2, X_1 X_2)_i\}_{i=1}^N$$

Model

$$y = sign(w_1 X_1 + w_2 X_2 + w_{12} X_1 X_2 + w_{11} X_1^2 + w_{22} X_2^2)$$

Parameters to be found:

$$w_1, w_2, w_{12}, w_{11}, w_{22}$$

Optimization problem:

e.g. max-margin at 5-dim space

- Want a non-linear decision boundary
- Keep the advantages of linear methods, e.g., optimization, interpretation

Decision boundary

$$w_1 X_1 + w_2 X_2 + w_{12} X_1 X_2 + w_{11} X_1^2 + w_{22} X_2^2 = 0$$

Disadvantages

- Explode at high-dim
- Possible solutions

Kernel (use symmetry), NNs (use non-linear activation)

Advantages of direct non-linear transform

- Simple for low-dim problem
- Interpretable as linear model
- Same way in regression

$$y = wx + b \rightarrow y = wx + w'x^2 + b$$

C4. Goodness of Fitting and Inference

Data set $\{(x_i, [y_i]\}_{i=1}^N$

- **Data mining/cleaning** (missing values, outliers, high leverages …)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, …)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity

Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(x_i, [y_i]\}_{i=1}^N$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, …)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, …)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, …)

Objective function

- **Error function** (RSS; Error rate; within-class distance, …)
- **Likelihood function**
- **Bayes posterior function**

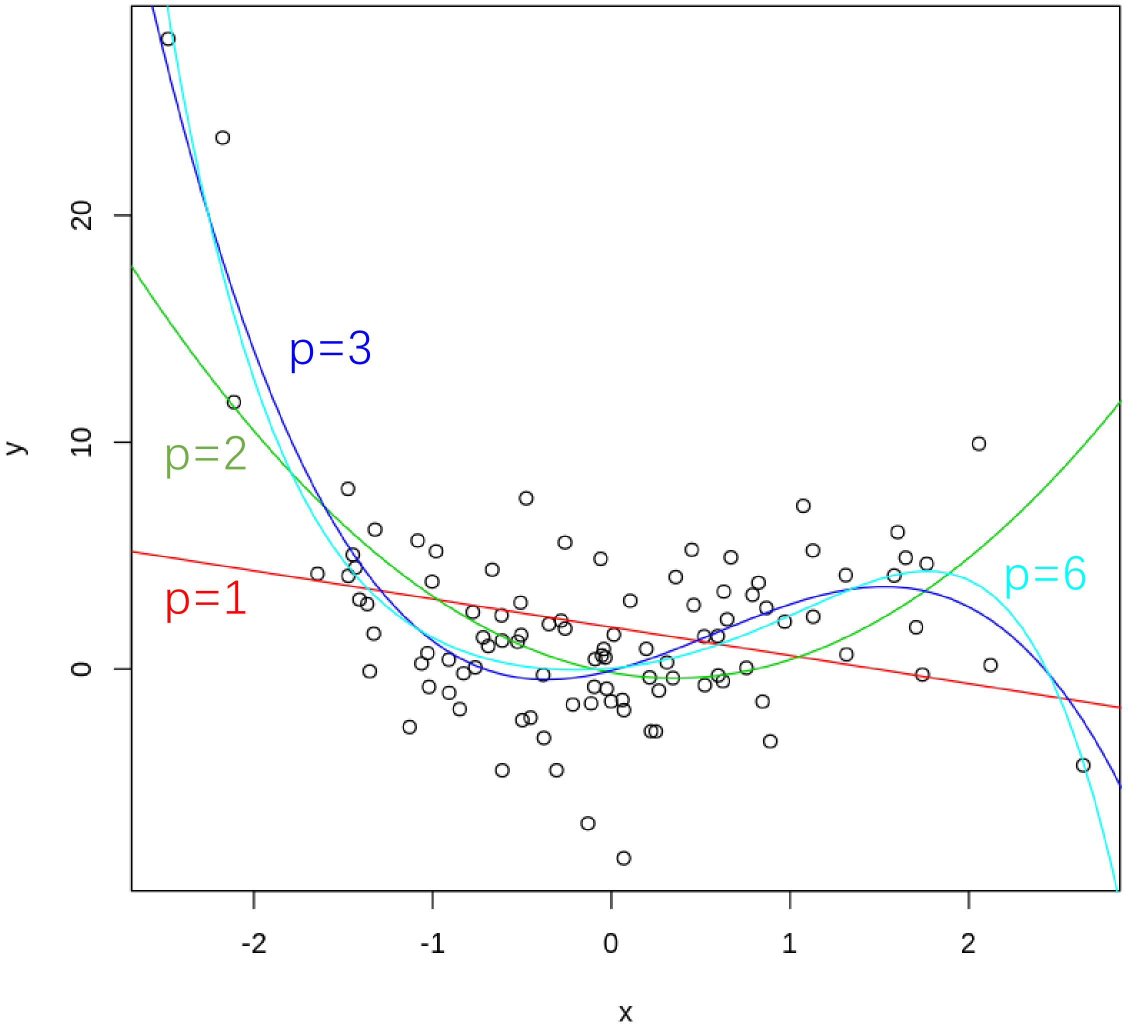
Optimization

- **Analytical** (LS for linear regr,)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

C4. Goodness of Fitting and Inference



Polynomial curve fitting revisit:

$$y = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_p x^p$$

Objective: RSS

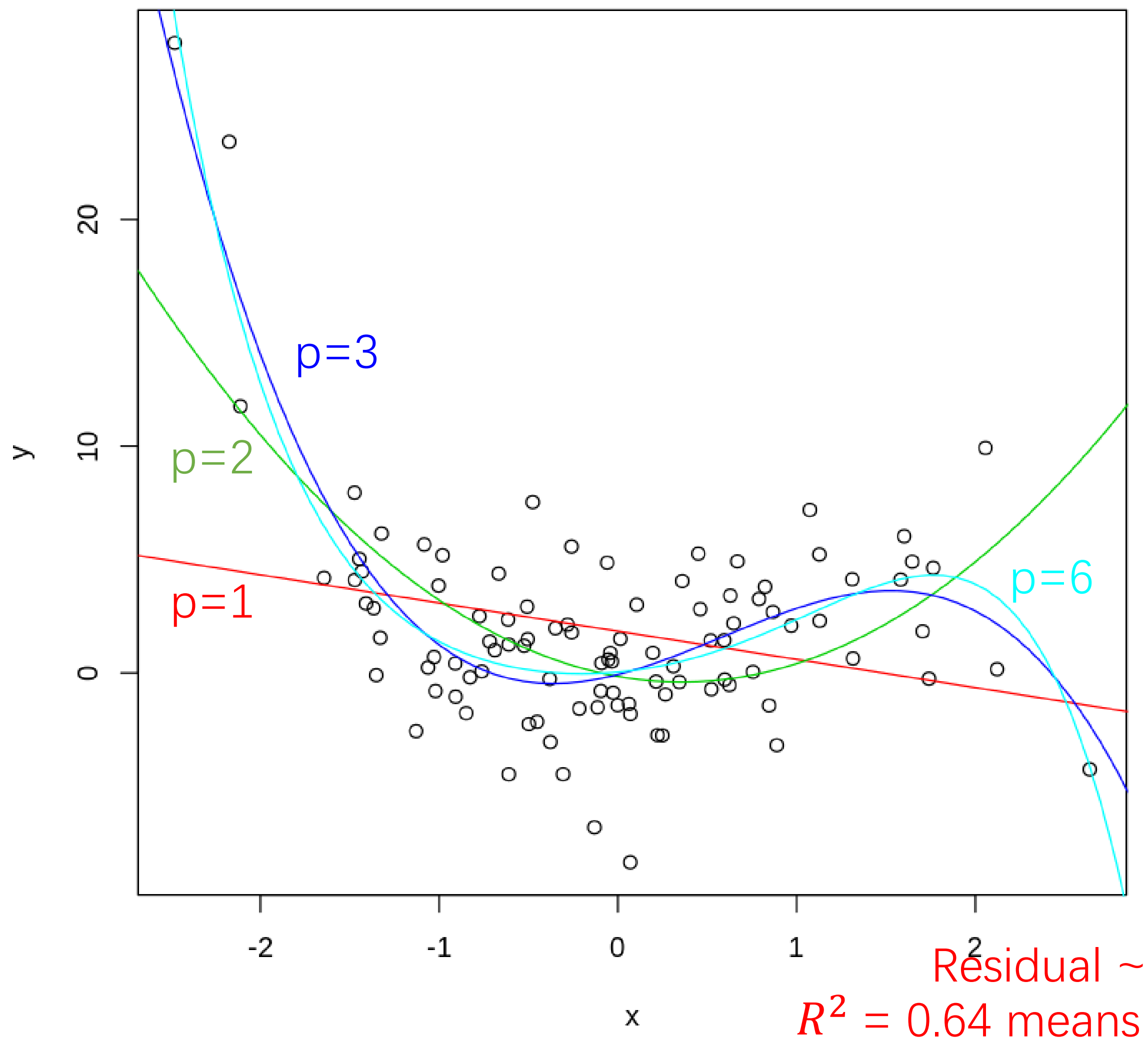
Optimization: Analytical

Question: In polynomial fitting, how to choose a best value p?

In R language:

```
lm.fit = lm(y~poly(x, 6))
xpred = seq(-4, 4, length=100)
ypred = predict(lm.fit, data.frame(x=xpred))
```

C4. Goodness of Fitting and Inference



Residual ~ 2.9

R^2 = 0.64 means 64% scatter of data is traced by this model

In R language:

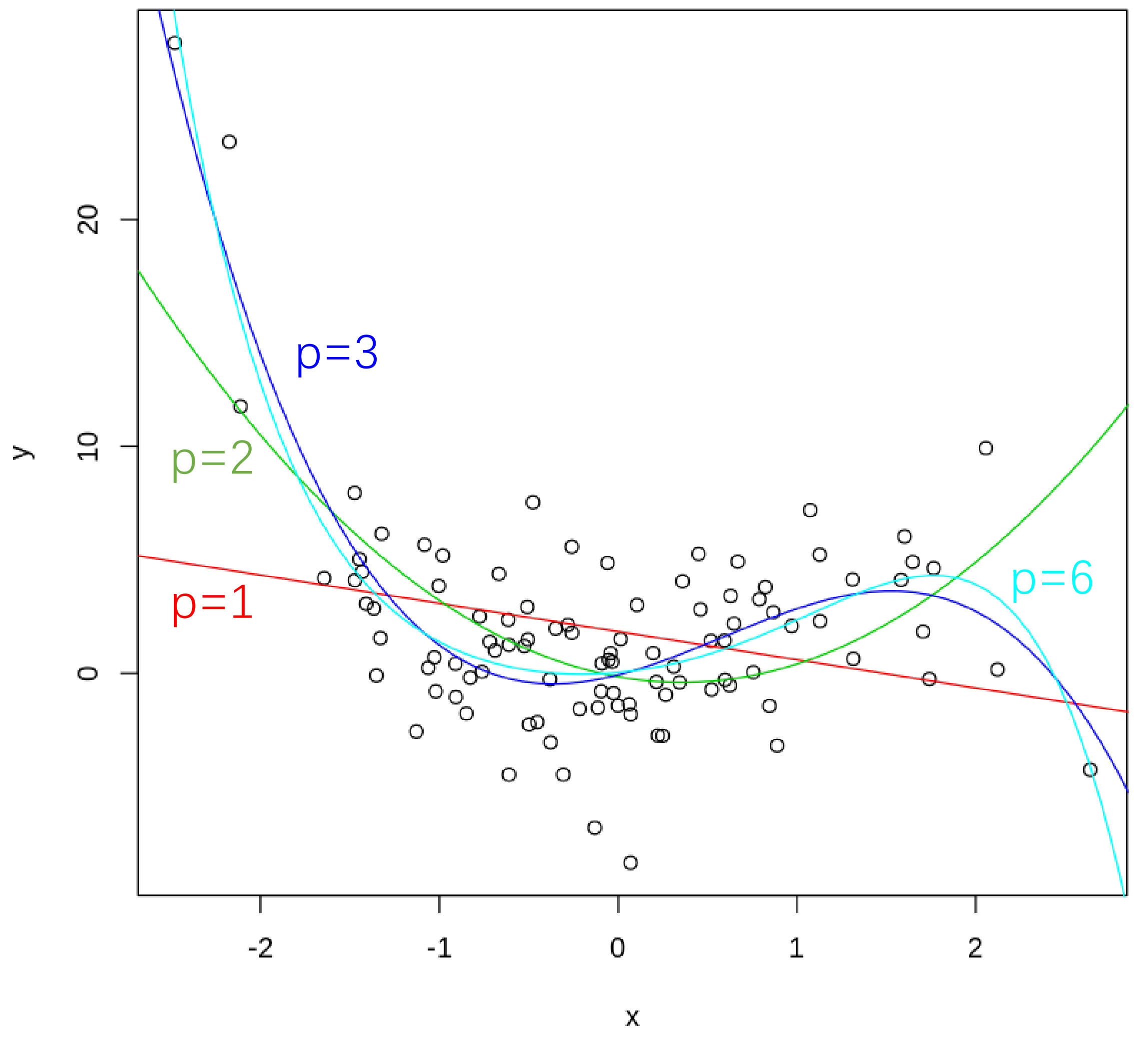
```
lm.fit = lm(y~poly(x, 6))
xpred = seq(-4, 4, length=100)
ypred = predict(lm.fit, data.frame(x=xpred))

summary(lm.fit)
```

```
Call:
lm(formula = y ~ poly(x, 6))

Residuals:
    Min      1Q Median      3Q     Max
-8.445 -1.621 -0.132  1.861  7.421

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.9200     0.2906   6.606 2.43e-09 ***
poly(x, 6)1   -12.4795     2.9064  -4.294 4.31e-05 ***
poly(x, 6)2    26.5001     2.9064   9.118 1.50e-14 ***
poly(x, 6)3   -23.5571     2.9064  -8.105 2.04e-12 ***
poly(x, 6)4     0.6426     2.9064   0.221   0.8255
poly(x, 6)5    -5.9518     2.9064  -2.048   0.0434 *
poly(x, 6)6     0.1292     2.9064   0.044   0.9646
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.906 on 93 degrees of freedom
Multiple R-squared:  0.6484,    Adjusted R-squared:  0.6257
F-statistic: 28.58 on 6 and 93 DF,  p-value: < 2.2e-16
```

The p-value of t-stats

The p-value of F-stats

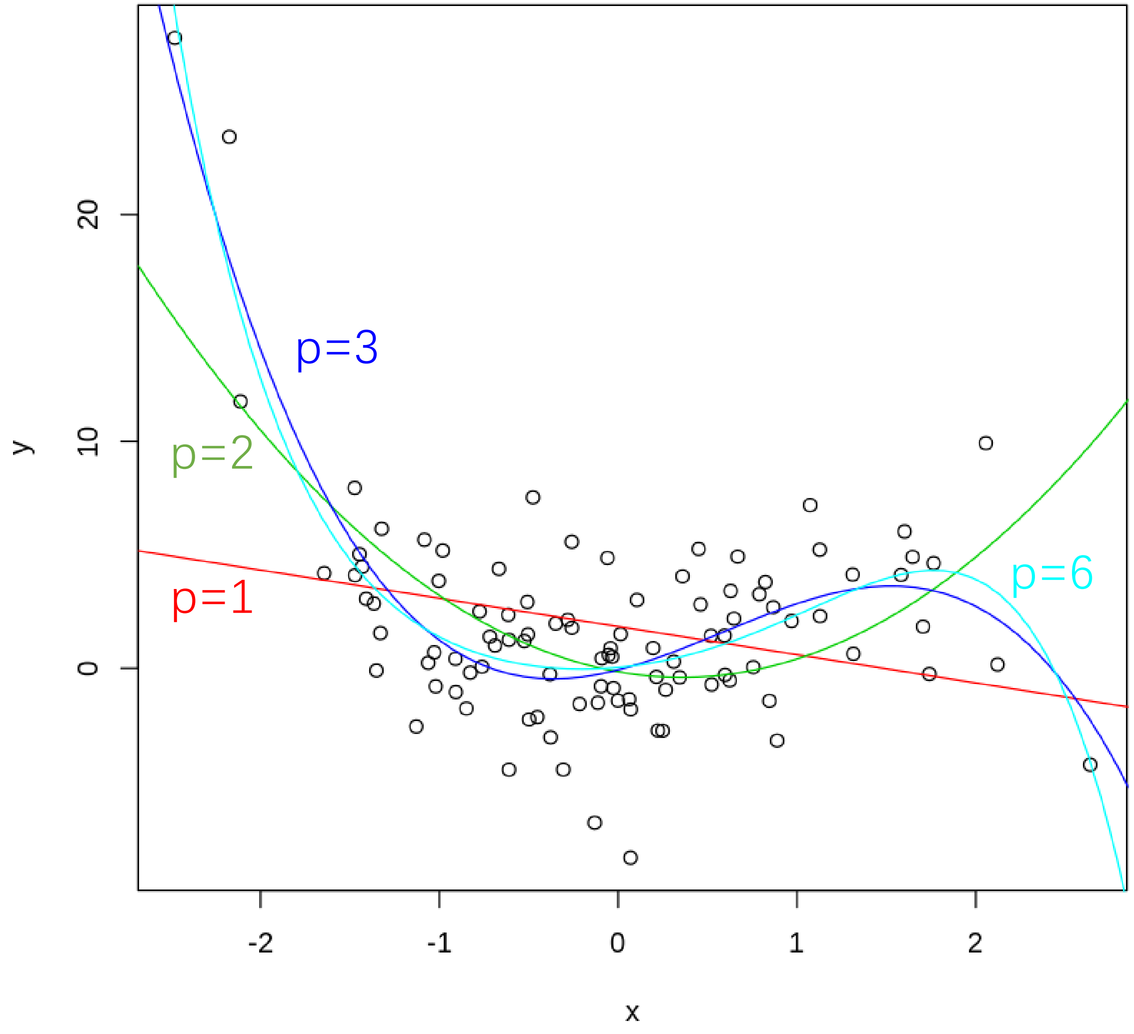
C4. Goodness of Fitting and Inference




```
1  lm.fit = lm(y~poly(x, 1))
2  summary(lm.fit)
```

```
Call:
lm(formula = y ~ poly(x, 1))

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1035  -2.7246  -0.6057   1.9765  22.8629 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.9200     0.4606   4.169 6.61e-05 ***
poly(x, 1)  -12.4795     4.6055  -2.710  0.00795 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.606 on 98 degrees of freedom
Multiple R-squared:  0.0697,	Adjusted R-squared:  0.06021 
F-statistic: 7.342 on 1 and 98 DF,  p-value: 0.007951
```

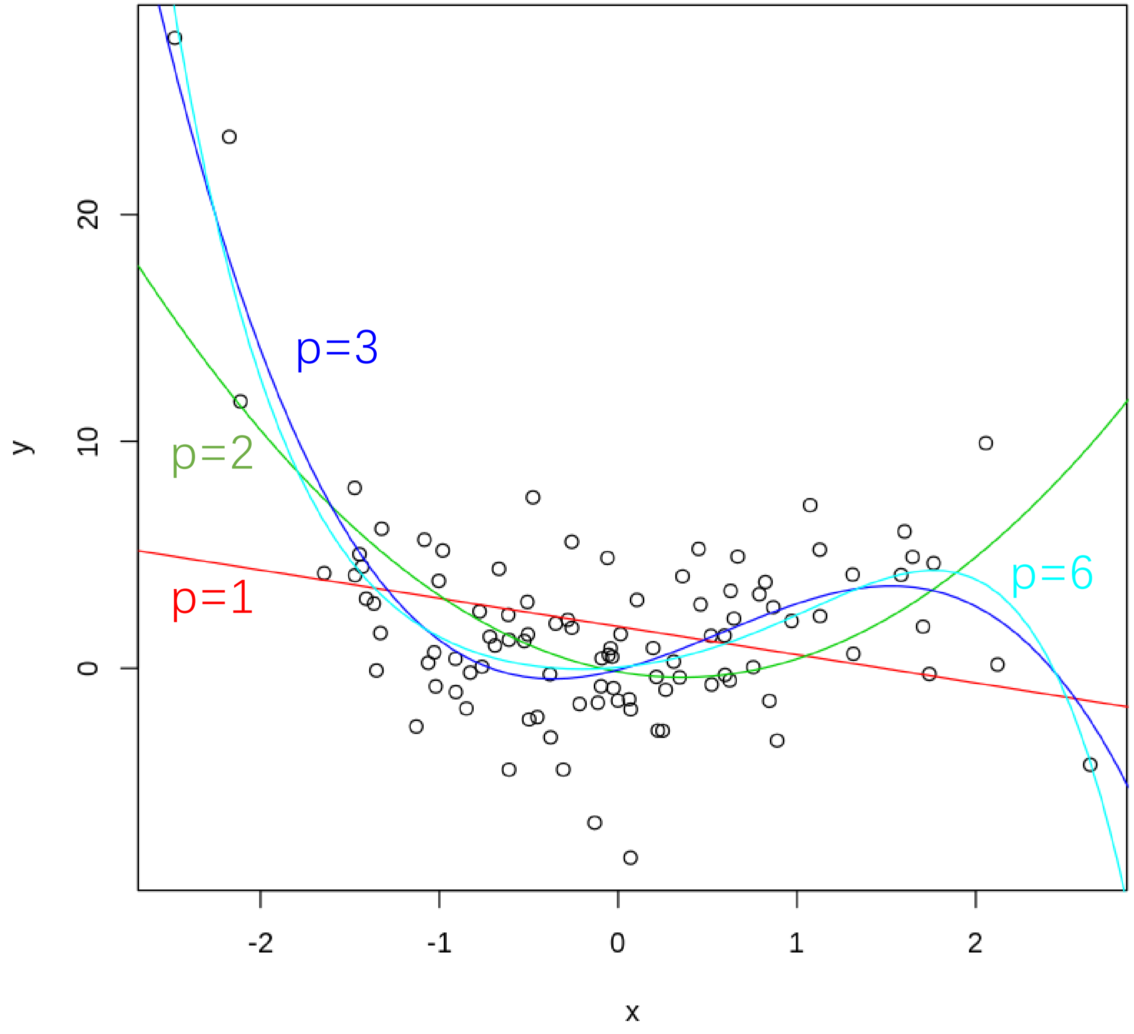
C4. Goodness of Fitting and Inference




```
lm.fit = lm(y~poly(x, 2))
summary(lm.fit)
```

```
Call:
lm(formula = y ~ poly(x, 2))

Residuals:
     Min       1Q   Median       3Q      Max 
-14.0943  -2.3387   0.0405   1.8061  12.4047 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.9200     0.3767   5.097 1.70e-06 ***
poly(x, 2)1   -12.4795     3.7669  -3.313   0.0013 ** 
poly(x, 2)2    26.5001     3.7669   7.035 2.83e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.767 on 97 degrees of freedom
Multiple R-squared:  0.384,	Adjusted R-squared:  0.3713 
F-statistic: 30.23 on 2 and 97 DF,  p-value: 6.236e-11
```

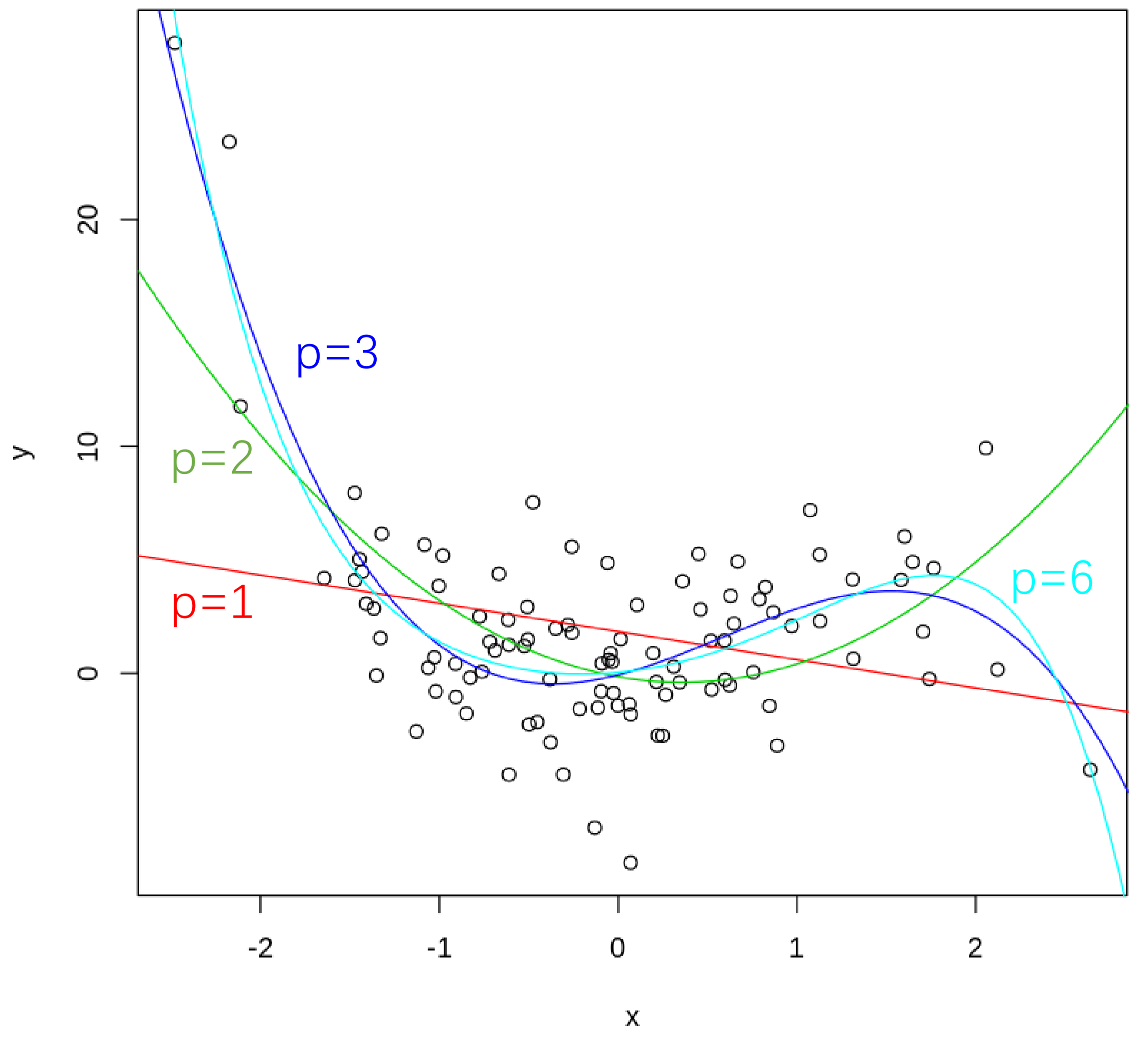
C4. Goodness of Fitting and Inference




```r
lm.fit = lm(y~poly(x, 3))
summary(lm.fit)
```

```
Call:
lm(formula = y ~ poly(x, 3))

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4277 -1.8608 -0.1296  1.7306  7.9533 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.9200     0.2925   6.564 2.67e-09 ***
poly(x, 3)1   -12.4795     2.9252  -4.266 4.66e-05 ***
poly(x, 3)2    26.5001     2.9252   9.059 1.56e-14 ***
poly(x, 3)3   -23.5571     2.9252  -8.053 2.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.925 on 96 degrees of freedom
Multiple R-squared:  0.6324,	Adjusted R-squared:  0.6209 
F-statistic: 55.04 on 3 and 96 DF,  p-value: < 2.2e-16
```

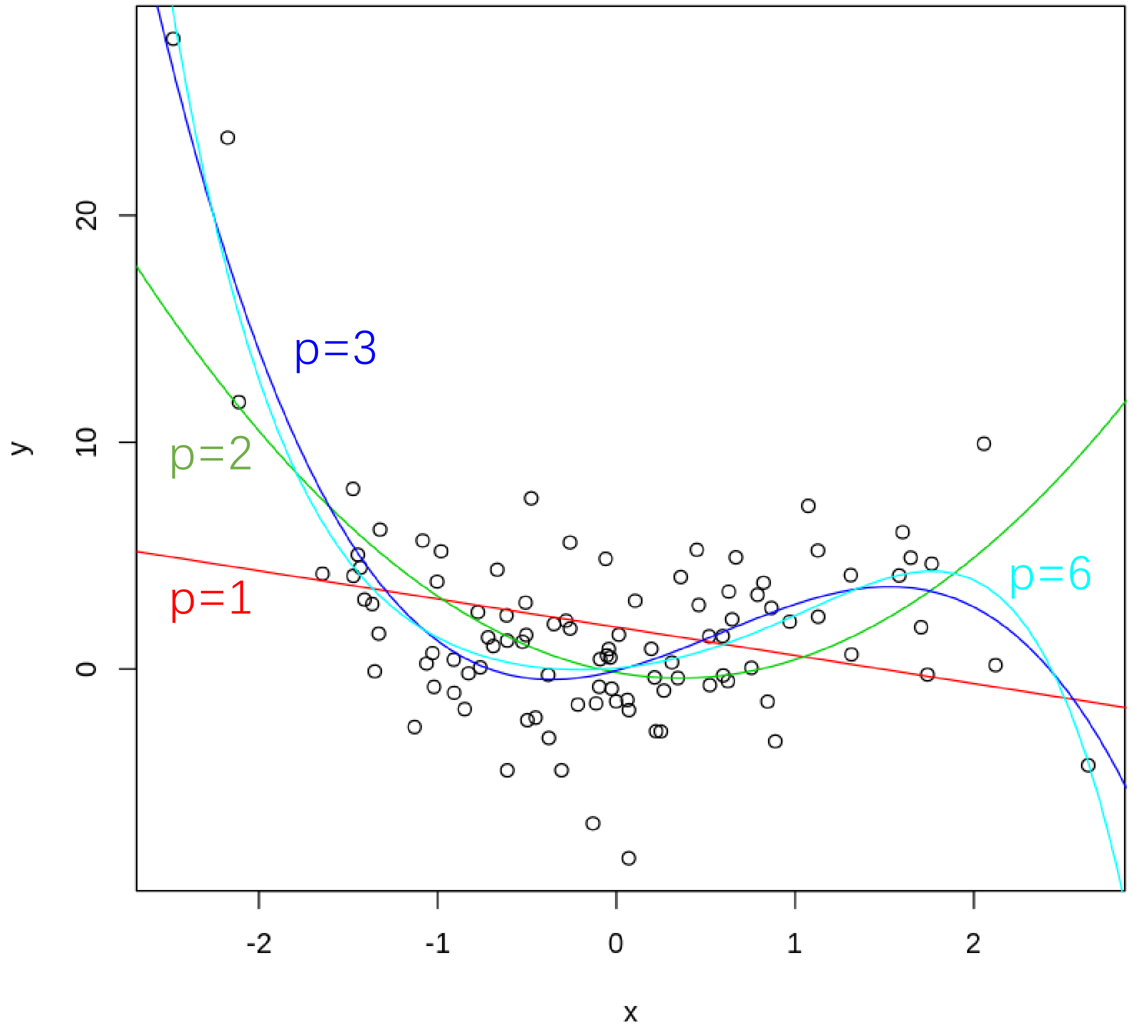
C4. Goodness of Fitting and Inference




```
lm.fit = lm(y~poly(x, 4))
summary(lm.fit)
```

```
Call:
lm(formula = y ~ poly(x, 4))

Residuals:
   Min     1Q Median     3Q    Max 
-8.476 -1.887 -0.131  1.719  7.931 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.9200     0.2940   6.531 3.21e-09 ***
poly(x, 4)1  -12.4795     2.9398  -4.245 5.09e-05 ***
poly(x, 4)2   26.5001     2.9398   9.014 2.11e-14 ***
poly(x, 4)3  -23.5571     2.9398  -8.013 2.83e-12 ***
poly(x, 4)4    0.6426     2.9398   0.219    0.827    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 95 degrees of freedom
Multiple R-squared:  0.6325,	Adjusted R-squared:  0.6171 
F-statistic: 40.88 on 4 and 95 DF,  p-value: < 2.2e-16
```

C4. Goodness of Fitting and Inference

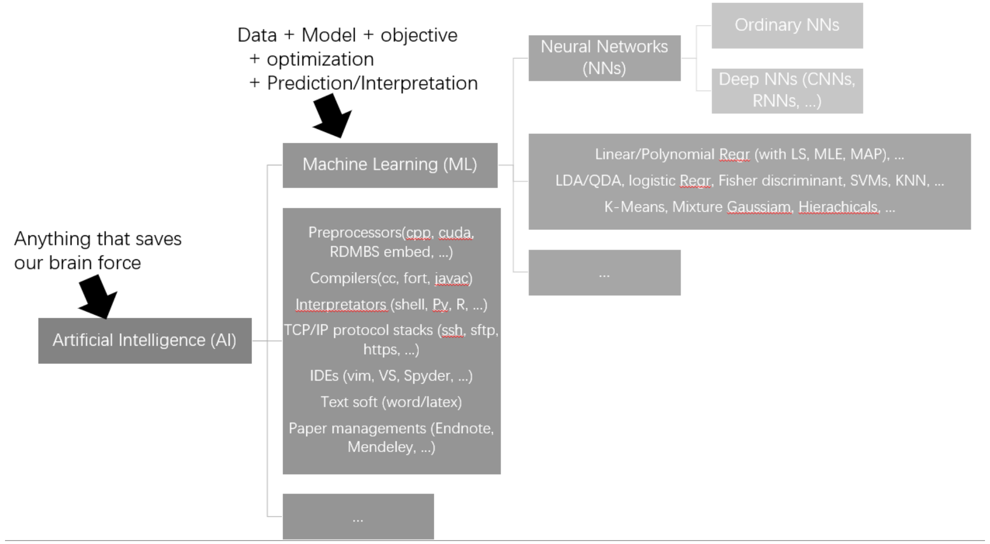


Summary on polynomial curve fitting:

- Goodness of fitting can be represented in many ways: $F, t, R^2, \mathrm{MSE}, \textbf{residual sd}$
- Inference, e.g., hypothesis test for parameters, can be handled with F, t
- By increasing the order p, and observing the p value, we can choose a confident model

What we have talked

Definition of ML



Framework of ML

Data set $\{(x_i, [y_i])\}_{i=1}^N$

- **Data mining/cleaning** (missing values, outliers, high leverages ...)
- **Type of question**
  - Supervised
    - Regression
    - Classification
  - Unsupervised problem
    - Density estimation
    - Clustering
    - Dimension reduction
- **Variable coding** (one to one, one to the rest, 1-of-K, ...)

Feature extraction

Nonlinearity
- Nonlinear transformations
- Kernel based-methods
- Neural Networks

Co-linearity
Dimension reduction

Model

Purpose: find $y = f(x)$ that is compatible with observations $\{(x_i, [y_i])\}_{i=1}^N$

- **Core problems**
  - The bias variance trade-off
  - The prediction-interpretation trade-off
- **Parametric vs. non-parametric**
- **Type of model**
  - Generative model (LDA, QDA, ...)
  - Probabilistic discriminative (Linear regr with LS; logistic regr, probit clas., soft-max clas.; Gaussian [Mixture])
  - Discriminant (Least Square Classifier, Fisher discriminant, perceptrons, trees, SVMs; K-Means)
- **Degree of freedom** (curse of dimensionality)
  - Analytical based methods (p-value)
  - Analytical approximation (AIC, BIC, ...)
  - Sampling partition (cross validation)
  - Model Regularization (Bayes prior, MLE regularization)
- **Model Combination**
  - Model ensemble (bagging, random forest, boosting)
  - Model mixing (Gaussian mixture, Bernoulli mixture, ...)

Objective function

- **Error function** (RSS; Error rate; within-class distance, ...)
- **Likelihood function**
- **Bayes posterior function**

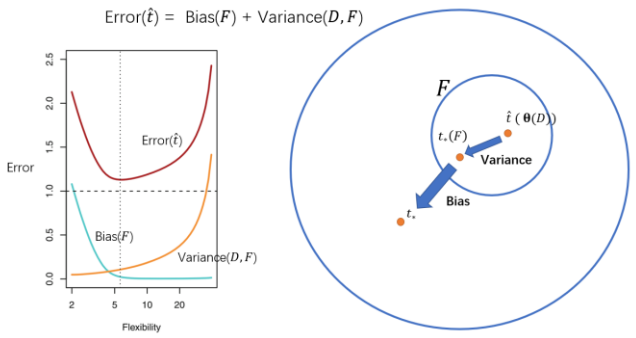
Optimization

- **Analytical** (LS for linear regr.)
- **Numerical** (maximum-margin for SVMs, EMs for model with latent variables, stochastic/batch gradient descent, Newtonian, error back-propagation for NNs)

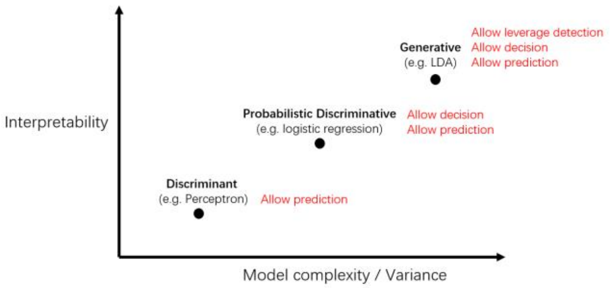
Goodness, prediction/decision and interpretation/inference

- R^2, MSE, error rate, confusion table and ROC curve
- F, t, χ^2 stats (with p-value)

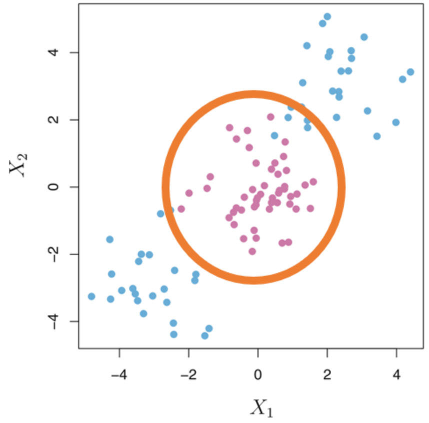
Bias-variance trade-off



Prediction-interpretability trade-off



Non-linearity



Goodness of fitting and model selection