Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry

patrick@mcsharry.net

www.mcsharry.net

Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Week 9

Course outline

Week	Description
1	Statistical learning
2	Linear models
3	Nonlinear models
4	Supervised learning
5	Unsupervised learning
6	Ensemble approaches

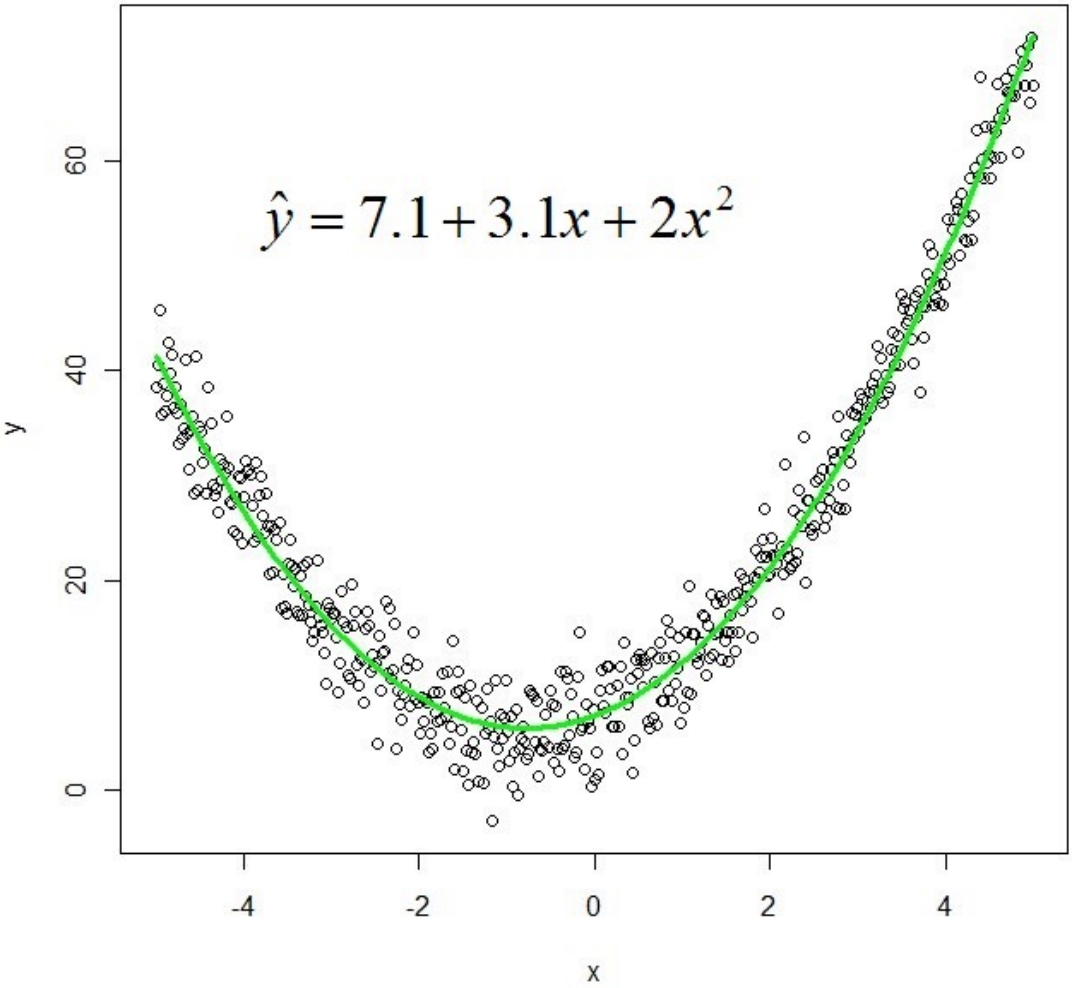
Applied Machine Learning

WEEK 9A

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Modelling nonlinear relationships	10
2	Discussion	What causes nonlinearity?	10
3	Case study	Heart rate dynamics	10
4	Analysis	Feature selection	20
5	Demo	Features and ApEn	20
6	Q&A	Questions and feedback	10

Nonlinear modelling



Poll

- What does non-linear mean?
- Input a word or two that caputures what non-linear means to you.

Slido.com
#36809

Nonlinearity

- **Mathematics**: Denoting or involving an equation whose terms are not of the first degree.
- **Engineering**: Involving a lack of linearity between two related quantities such as input and output: a non-linear network.
- **Literature**: Not sequential or straightforward: James Joyce’s stream-of-consciousness, non-linear narrative.

Feedback loops

- Feedback occurs when outputs of a system are "fed back" as inputs as part of a chain of cause-and-effect that forms a circuit or loop.
- Negative feedback: cruise control in a car to achieve a target speed such as the speed limit.
- Positive feedback: ice-albedo positive feedback loop whereby melting snow exposes more dark ground (of lower albedo), which in turn absorbs heat and causes more snow to melt.

Regimes

- A regime is a stable attractor of a system and represents characteristic behavior of that system.
- Climate, depression and recession may all be considered as regimes within their respective systems (atmosphere, mood, economy).
- A regime may be maintained by mutually reinforced processes or feedbacks.

Regime shift

- Regime change occurs when a smooth change in an internal process (feedback) or a single disturbance (external shocks) triggers a completely different system behavior.
- Such regime switching is one particular type of nonlinearity.
- Being able to detect regimes can be useful for classifying states that correspond to health and illness or order and disorder.

Poll

- Modelling the influence of temperature on electricity demand in a city with heating devices and air conditioning requires a polynomial of degree:

a) One

b) Two

c) Three

d) Four

Slido.com
#36809

Structural break or regime shift

- External influences – noise or shocks
- Dynamics change due to internal influences
- Regime switching
- Examples:
  - growth versus recession
  - mania versus depression
- Natural variability in systems makes it difficult to detect regime changes.
- Global oscillations in weather/climate

Surrogate data

- Surrogate data testing is a statistical proof by contradiction technique and similar to parametric bootstrapping used to detect non-linearity in a time series.
- The technique basically involves specifying a null hypothesis H_0 describing a linear process and then generating several surrogate data sets according to H_0 using Monte Carlo methods.
- A discriminating statistic is then calculated for the original time series and all the surrogate set.
- If the value of the statistic is significantly different for the original series than for the surrogate set, the null hypothesis is rejected and non-linearity assumed.

Surrogate algorithms

- There are two approaches to generating surrogates:
- 1. Parametric: surrogate data series are generate using a model that has been fit to the original data.
- 2. Non-parametric: surrogate data series are generated directly from the original data.
- The latter have the advantage of not requiring any parameter estimation.

Surrogate methods

- **Random shuffle**: surrogates generated by randomly shuffling the original data; obtaining same distribution and destroying linear correlations.
- **Random phases**: surrogate data are generated by the inverse Fourier Transform of the amplitudes of Fourier Transform of the original data with new (uniformly random) phases. This approach preserves the linear correlations in the data.

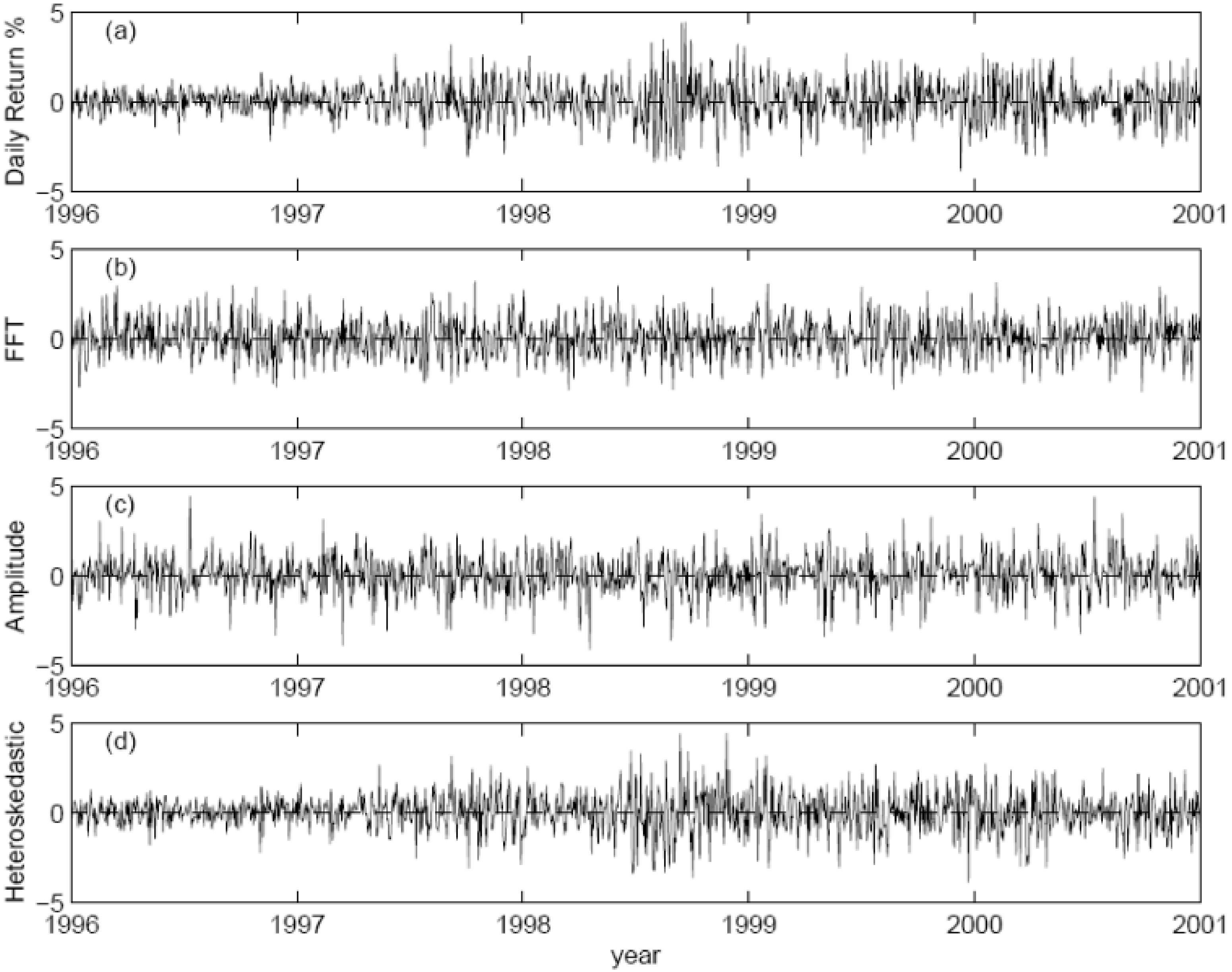
Surrogate methods

- Amplitude Adjusted Fourier Transform (AAFT) combines random shuffle and random phases, in that it preserves the distribution and the linear correlations.
- An iterative version of the AAFT is available to polish the surrogates to obtain a better a better match to the distribution and the linear correlations (autocorrelation function).

Surrogates

- Surrogate data provides a means of null hypothesis testing
- IID surrogates: independent identically distributed obtained by shuffling the original time series (preserves the unconditional distribution)
- FFT surrogates: obtained by Fourier decomposition of the original time series
- FFT surrogates preserve the linear correlations (but not the nonlinear correlations)
- Polished surrogates attempt to preserve both the unconditional distribution and the linear correlations
- Parametric model-based surrogates rely on AR and ARMA
- Null hypothesis should be *interesting* and *relevant* to the application
- Surrogates should look like the original!
- Use monte-carlo approach to sample distribution of test statistic for surrogate
- Test significance of value obtained from real time series

FTSE surrogates



Regularity and unpredictability

- Consider two time series:

  A: {1,2,1,2,1,2,1,2,1,2,1,2}

  B: {1,1,2,1,2,1,2,1,2,1,2,2}
- Traditional statistics, such as the mean and variance, cannot distinguish between A and B.
- However, series A is perfectly regular in that knowing that we observed a 1 allows us to predict the following 2 with certainty.
- On the other hand, series B is random and therefore unpredictable.
- How can we measure such regularity?

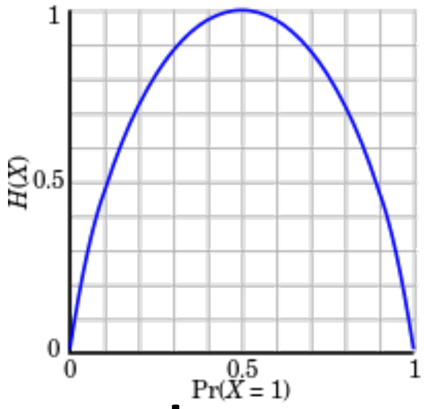
Information

- Information is measured by:

  I(p) = log(1/p) = - log(p).

- This definition satisfies three properties:
- I(p) ≥ 0: information is a non-negative quantity
- I(1) = 0: events that always occur do not communicate information.
- I(p_1 p_2) = I(p_1) + I(p_2): information due to independent events is additive.

Entropy



- Entropy is a measure of disorder or expected surprise.
- Flipping a fair coin has equal probability of heads and tails (p = q = 0.5) leading to maximum uncertainty and maximum entropy.
- An unfair coin (p ≠ q ≠ 0.5) has smaller entropy.
- The extreme case of a double-headed coin that never comes up tails, or a double-tailed coin that never results in a head leads to minimal entropy.

Entropy

- Entropy measures the average uncertainty in the value of the discrete-valued probability density.
- The recurrence probability density entropy (RPDE) for a probability of recurrences times, p(k), with k=1,…,K is given by:

  $H = -\Sigma_k\, p(k) \ln p(k)$.
- For a perfectly periodic signal with p(k)=1 for a single value of k and zero otherwise: $H_{per} = 0$.
- For a random signal with uniform density, $H_{iid} = \ln K$.
- Normalized entropy $H_{norm} = H/H_{iid}$ lies between 0 and 1.

Measuring Entropy

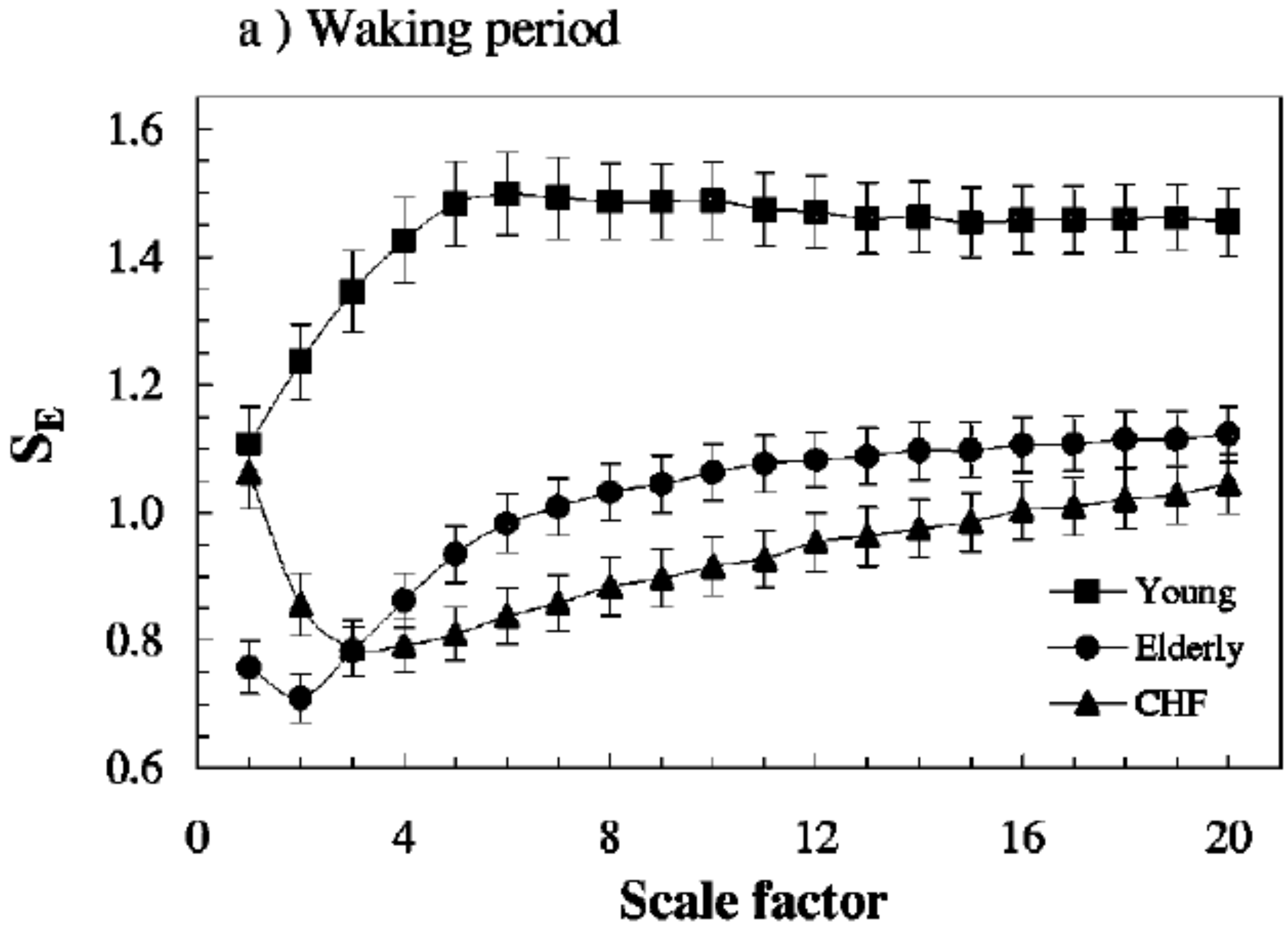
- Approximate entropy (ApEn) is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data.
- Standard formula for measuring entropy requires access to large amounts of data and are sensitive to noise.
- ApEn is a modification that allows estimation of entropy using empirical observations.
- Sample Entropy (SampEn) is independent of the amount of data.

Regularity in heart rate

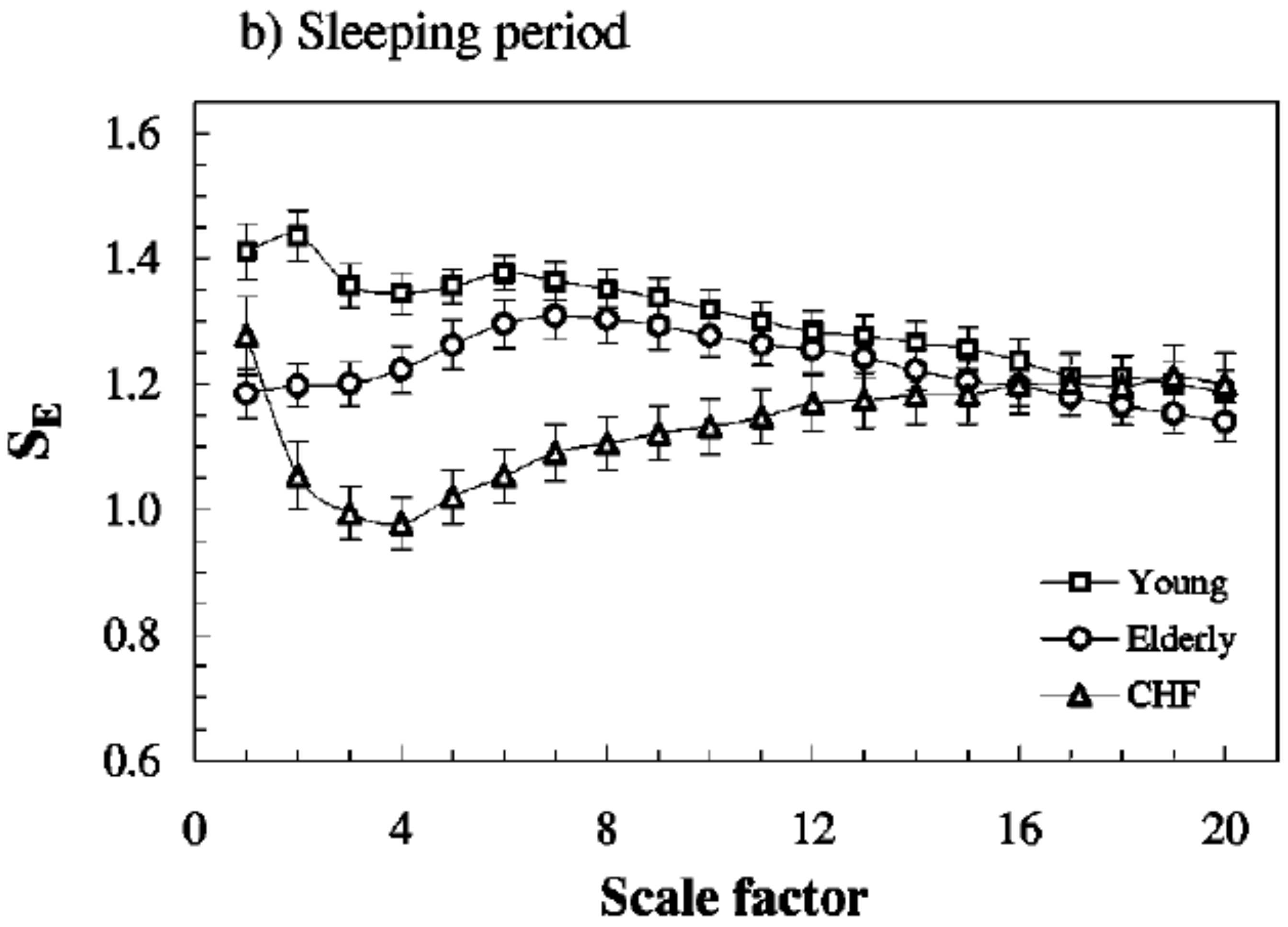
- Approximate Entropy was used to measure regularity in heart rate time series.
- Heart rate approximate entropy decreased with age and was higher in women than men ($p < 0.05$).
- Authors concluded that women are more complex than men!

Ryan et al. (1994). Gender- and age-related differences in heart rate dynamics: are women more complex than men? J Am Coll Cardiol 24(7):1700-7.

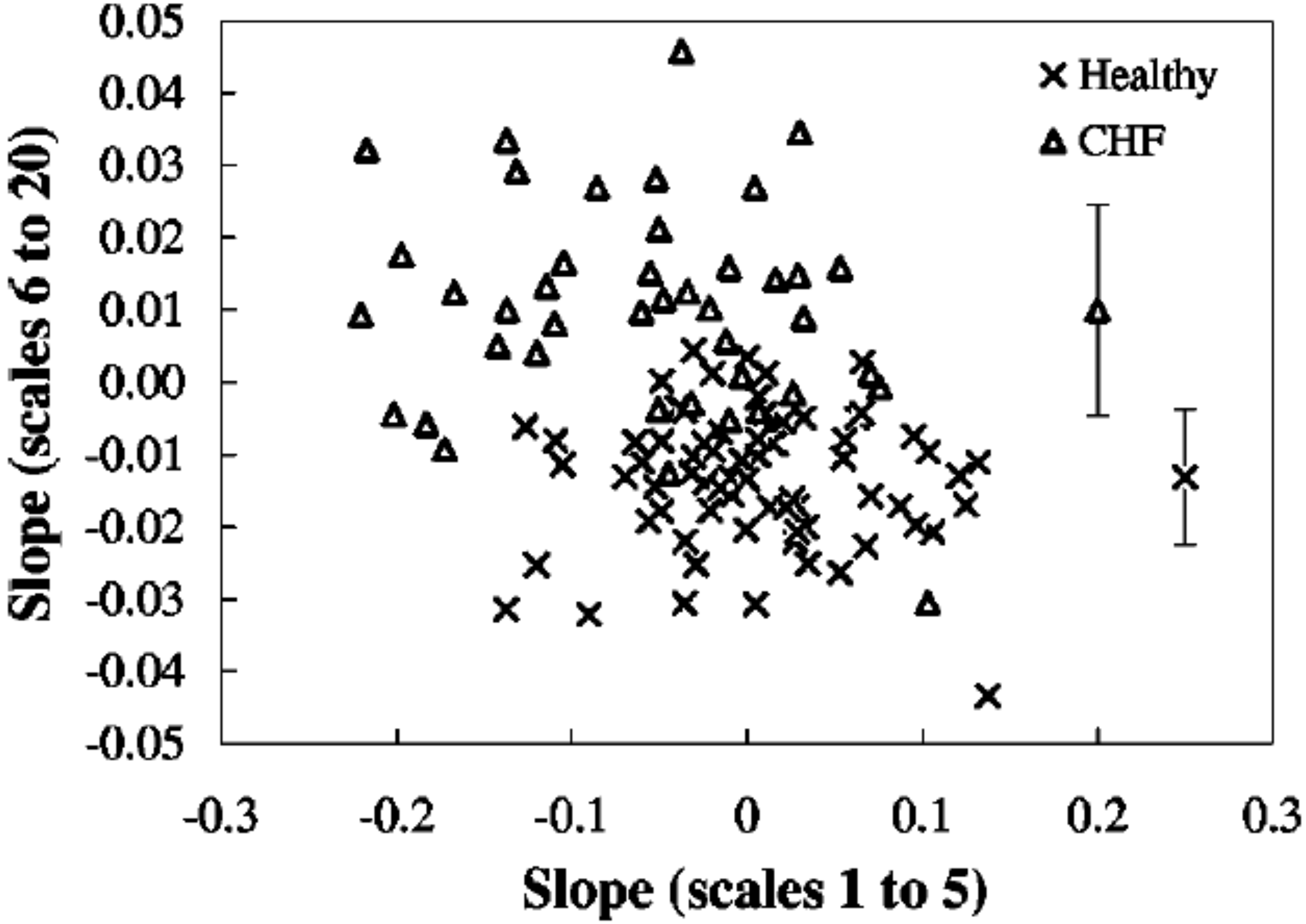
Entropy of heart rate



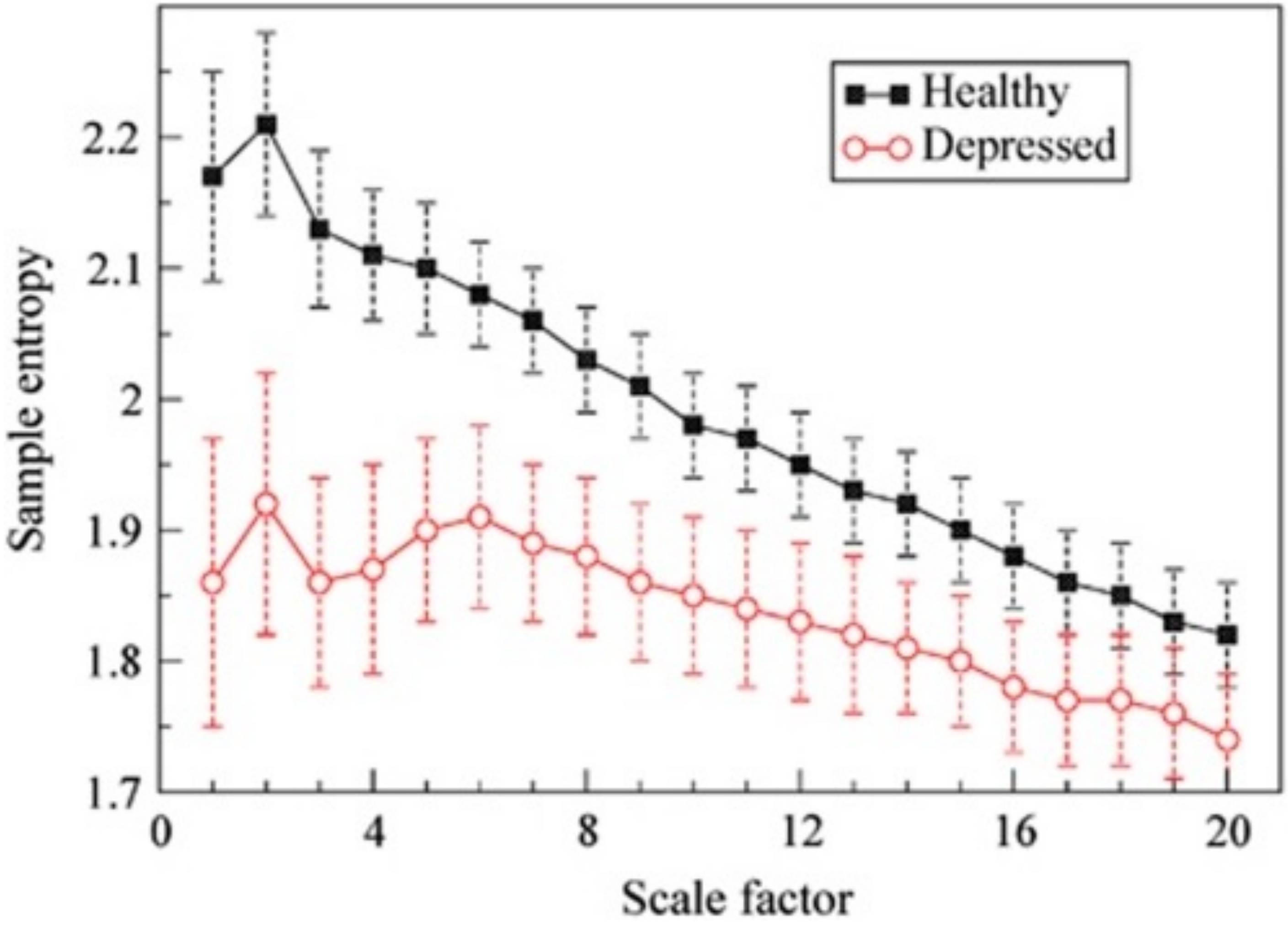
Entropy of heart rate



Healthy versus Congestive Heart Failure



Entropy and mood

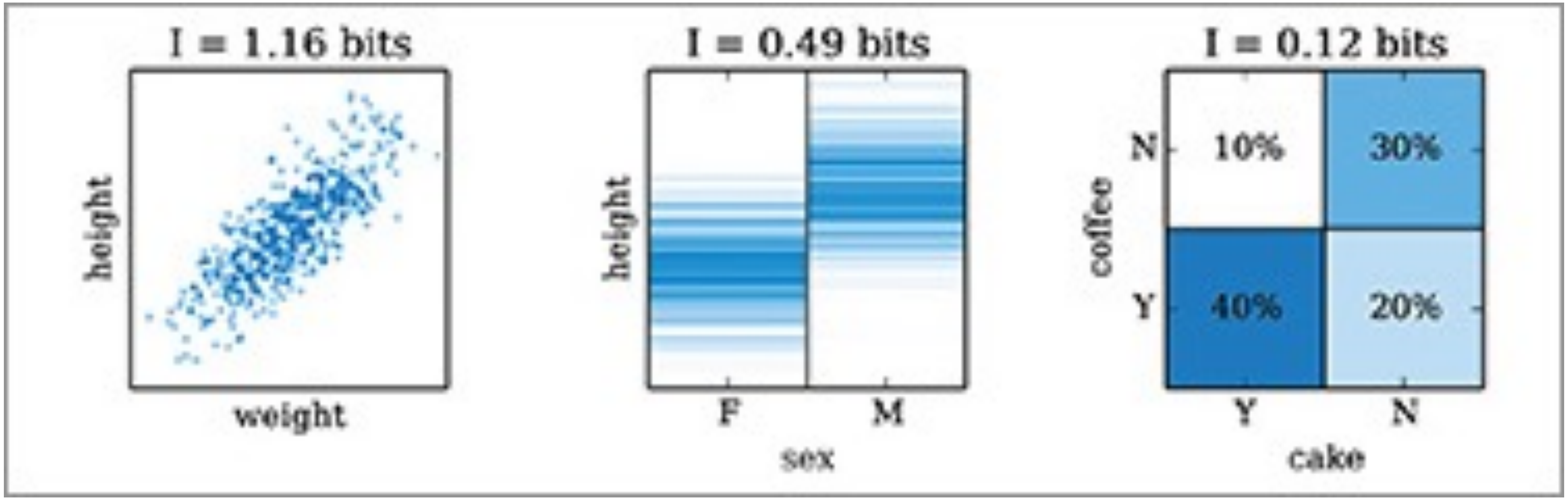


Leistedt et al. (2011). Decreased neuroautonomic complexity in men during an acute major depressive episode: analysis of heart rate dynamics. Translational Psychiatry 1, e27.

Mutual information

- Shannon (1948) showed that the concept of “information” can be formalized by a mathematical quantity.
- Mutual information quantifies the amount of information that one variable provides about another variable.
- It is typically measured in units called “bits”.
- A value of zero indicates no dependence.
- Larger values imply the existence of a relationship (which might be linear or nonlinear).

Mutual information



Kinney & Atwal (2014). Equitability, mutual information and the maximal information coefficient, PNAS 111(9):3354-3359.

Information theory

- Information:

  $I(x) = -\log p_x(x)$

- Entropy:

  $H(x) = -\int p_x(x)\log[p_x(x)]dx$

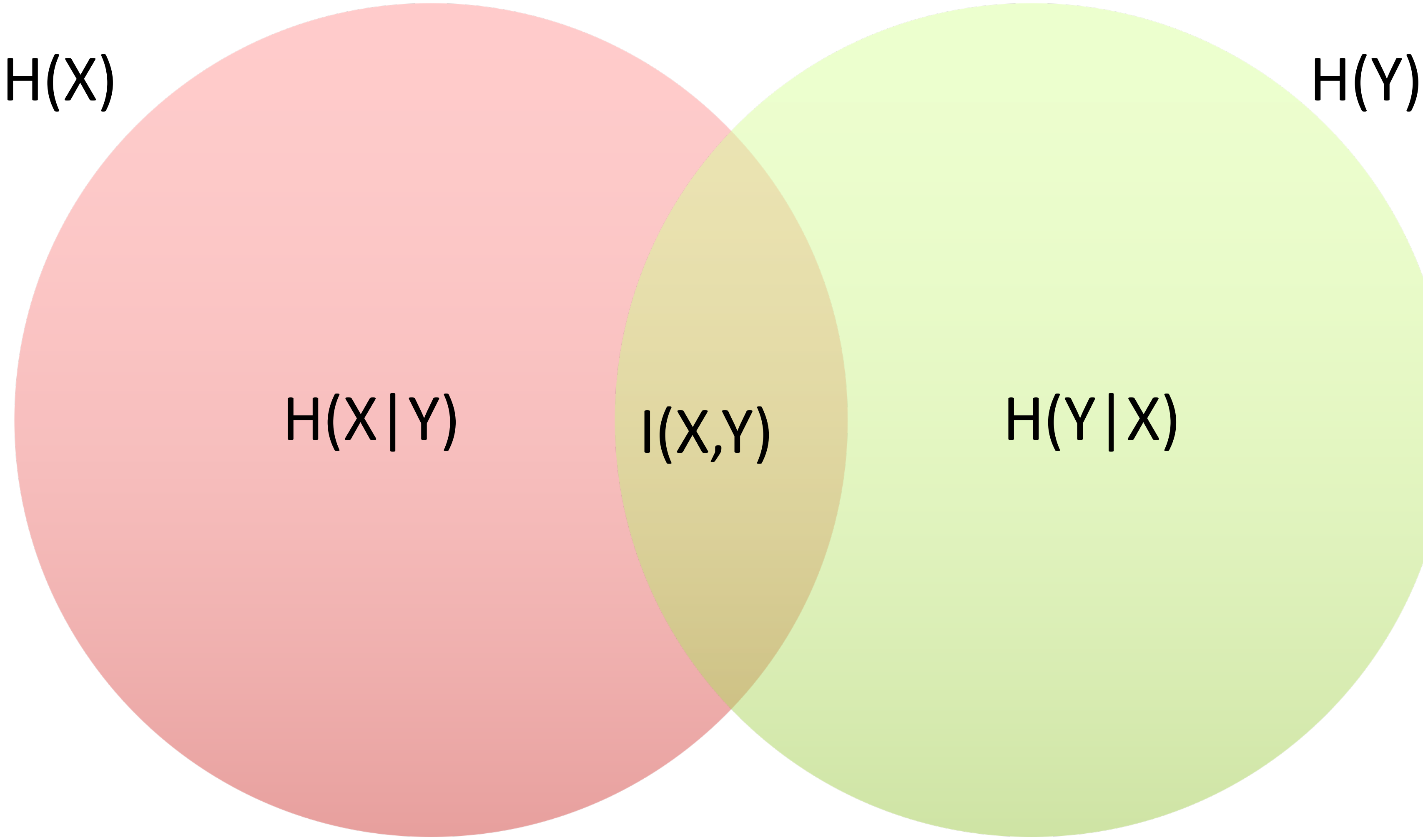
- Joint entropy:

  $H(x,y) = -\int\int p_{x,y}(x,y)\log[p_{x,y}(x,y)]dxdy$

- Mutual information:

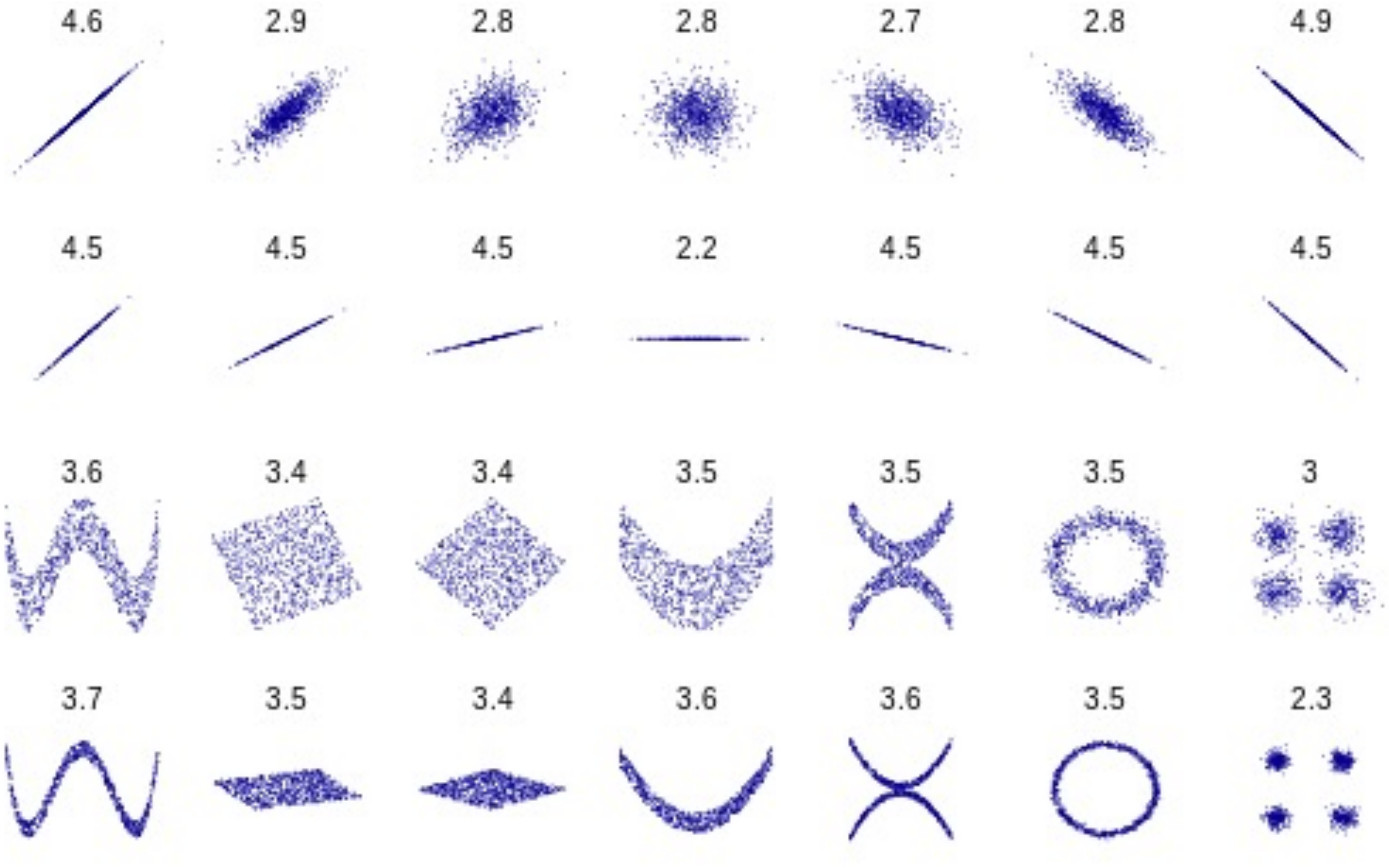
  $$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= \int\int p_{x,y}(x,y)\log[p_{x,y}(x,y)/(p_x(x)p_y(x))]dxdy \end{aligned}$$

Mutual Information



$$H(X,Y) = H(X|Y) + H(Y|X) + I(X,Y)$$

Mutual Information examples



Source: https://commons.wikimedia.org

Quiz

- Nonlinearity is relevant in the following stages of machine learning:

a) Feature engineering

b) Feature selection

c) Model construction

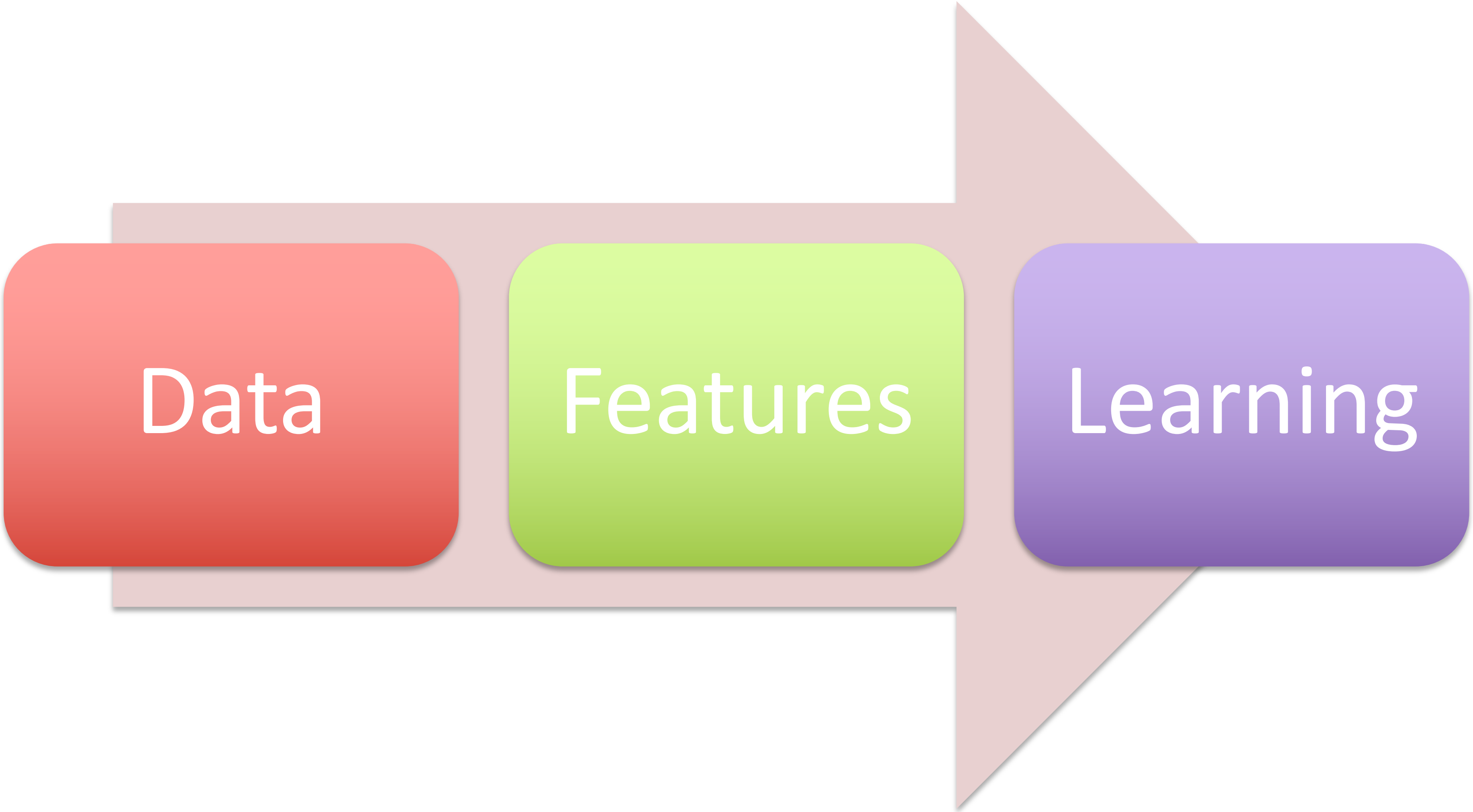
d) All of the above

Slido.com
#36809

Features

- Features refer to the inputs of a machine learning algorithm.
- Features are simply another name for an explanatory variable or predictive variable in a statistical model.
- Features are extracted in the sense that some pre-processing may be necessary to calculate the values of a given feature.

Learning process



Data

Features

Learning

Data standardization

- In many applications, we collect a number of different explanatory variables, x_1, x_2, x_m.
- It is likely that these variables will have different means, standard deviations and ranges.
- Some techniques are sensitive to the magnitude of the variables and it is useful to normalize the variables.

Z-score

- The z-score or standard score is the (signed) number of standard deviations an observation is above the mean.
- The z-score is defined as

  $z = (x – \mu)/\sigma$

- where μ is the mean and σ is the standard deviation.
- If $x \sim N(\mu, \sigma^2)$, then the $z \sim N(0,1)$.

Feature selection

- Feature selection refers to techniques for identifying which set of features, X, contain predictive information about the dependent variable y.
- Some machine learning techniques are extremely sensitive to the features being provided
- “Garbage in, garbage out”

Feature selection - correlation

- The simplest example of feature selection is to calculate the correlation coefficient between each feature and the dependent variable.
- The ranking of the absolute correlations is one approach for selecting features.
- Only those features which are significant at the 95% level are then selected for submission to the machine learning algorithm

Lasso feature selection

- The LASSO techniques introduced earlier can be used to select features.
- These features can then be used as inputs into a nonlinear machine learning technique.
- It is common practice to separate the feature selection and model construction into two stages in this way.

Minimum-redundancy-maximum-relevance (mRMR)

- Peng et al. (2005) introduced mRMR as a feature selection method that can use either mutual information, correlation, or distance/similarity scores to select features.
- mRMR relies on measuring relevancy and redundancy.
- mRMR penalises a feature's relevancy by its redundancy in the presence of the other selected features.

Relevance

- Consider a feature set S contain features x_i.
- The relevance of a feature set *S* for the class *y* is defined by the average value of all mutual information values between the individual features x_i and the class *y:*
-

$$D(S,y) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, y)$$

Redundancy

- The redundancy of all features in the set *S* is the average value of all mutual information values between the feature x_i and the feature x_j:

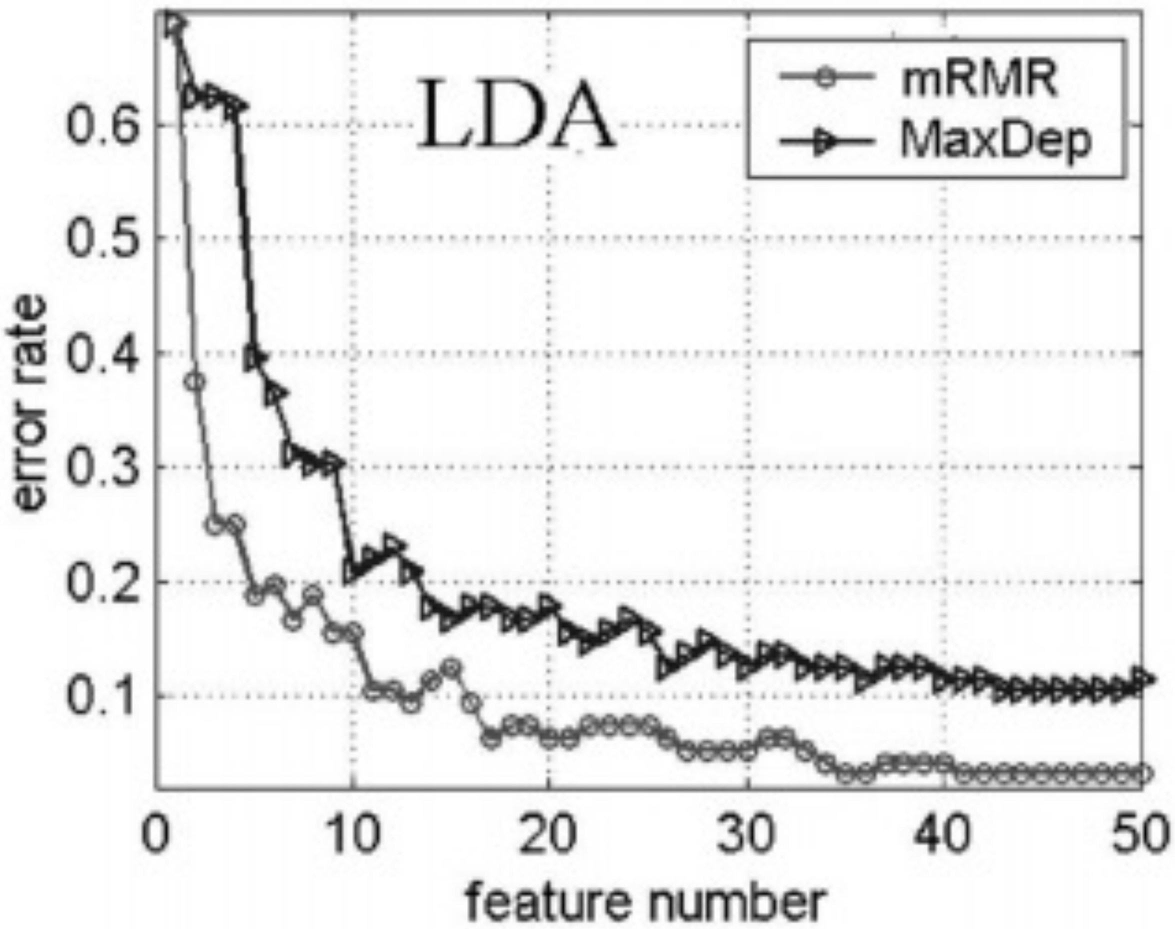
$$R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

mRMR

- The mRMR criterion is then a simple linear combination of the relevance and redundancy and is defined as follows:

$$mRMR = \max_{S}\left[D(S,y) - R(S)\right]$$

Feature Graph



For different numbers of features, the model accuracy is optimal for the features selected by MRMR.

Q&A

Applied Machine Learning

WEEK 9B

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Monitoring	10
2	Discussion	Nonlinear structures	10
3	Case study	Growth charts	10
4	Analysis	Nonlinear regression	20
5	Demo	Techniques for nonlinear modelling	20
6	Q&A	Questions and feedback	10

Monitoring

- We are often faced with the task of monitoring a given variable over time.
- In order to identify faults, anomalies or issues that needs a response, it is important to have a benchmark for comparison.
- This is a commonly used approach in medicine but can be applied to other disciplines.

Growth

- How would you model growth of a child in order to monitor development?
- This can be useful for relating age to weight when estimating drug dosages.
- It can also be used for tracking human development in terms of nutrition.

Poll

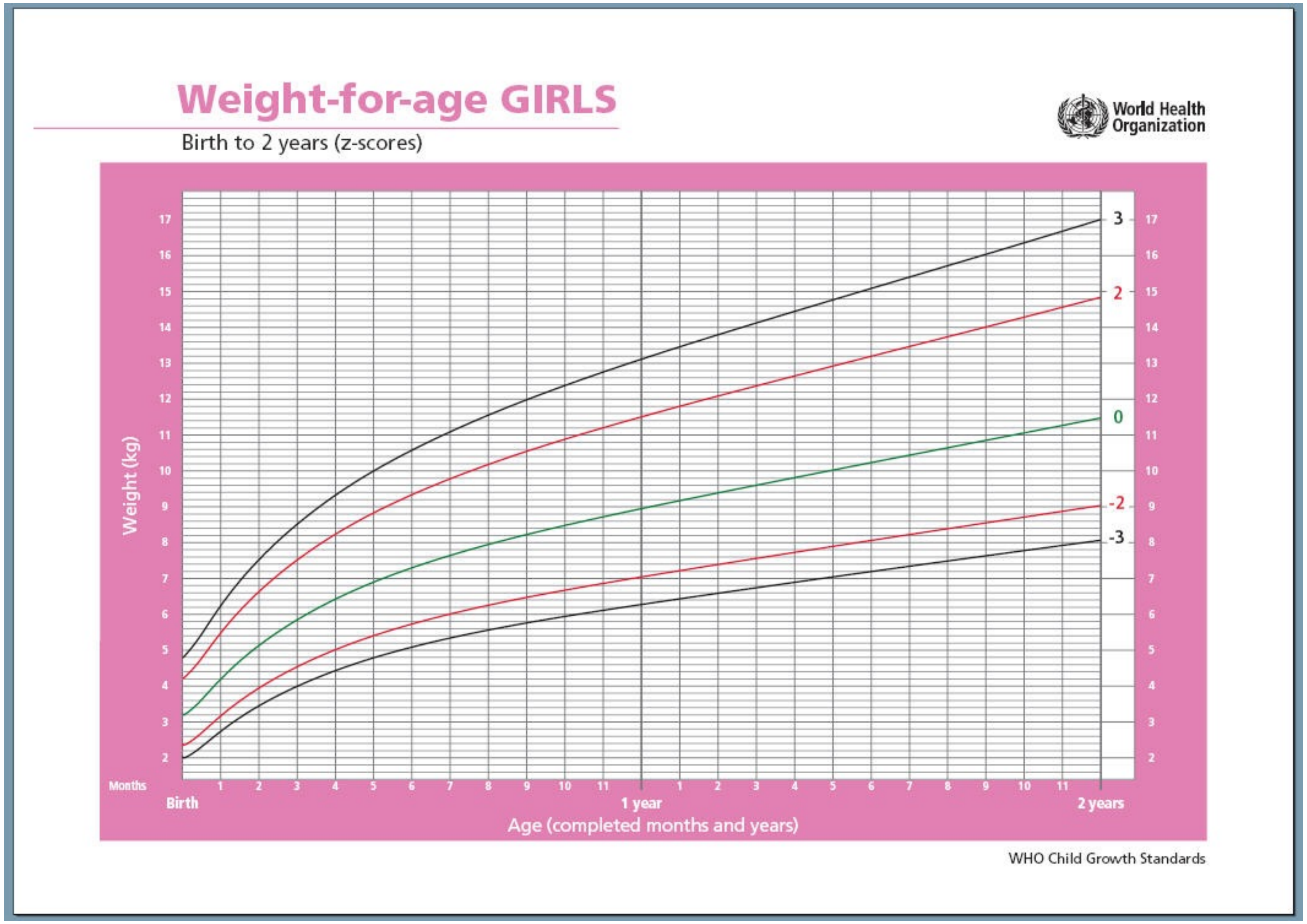
- Is the growth of a child from birth to two years of age a linear process?

- **Slido.com**
- **#50845**

Linear growth

- There are many studies about the most appropriate mathematical model for describing the relationship between weight and age in children.
- A simple linear relationship, known as the Advanced Paediatric Life Support (APLS) formula, is used to estimate the weight of children in paediatric emergencies in order to administer correct drug dosages.
- APLS estimates weight in Kg using age in years as: weight = [age + 4] × 2.
- The simplicity of this linear formula has the advantage of being easy to calculate in emergency situations

WHO growth chart



Poll

- What type of nonlinear model might work for describing growth of a baby?

- **Slido.com**
- **#50845**

Leffler formula

- However, as weight growth curves display evidence of nonlinearity, especially for young babies, more complex models have been developed.
- The Leffler formula provides distinct relationships for babies below and above the age of one year:

$$m = 0.5a_m + 4$$

$$m = 2a_y + 10$$

Nonlinear formula

- More complex models include the model of Theron and the non-parametric approach used by the WHO.
- Theron's formula is given by:

$$m = \exp(0.175571a_y + 2.197099)$$

- This is known as a log-linear relationship:

$$\log(m) = b + ca$$

Nonlinear structures

- There are many examples of nonlinear relationships:

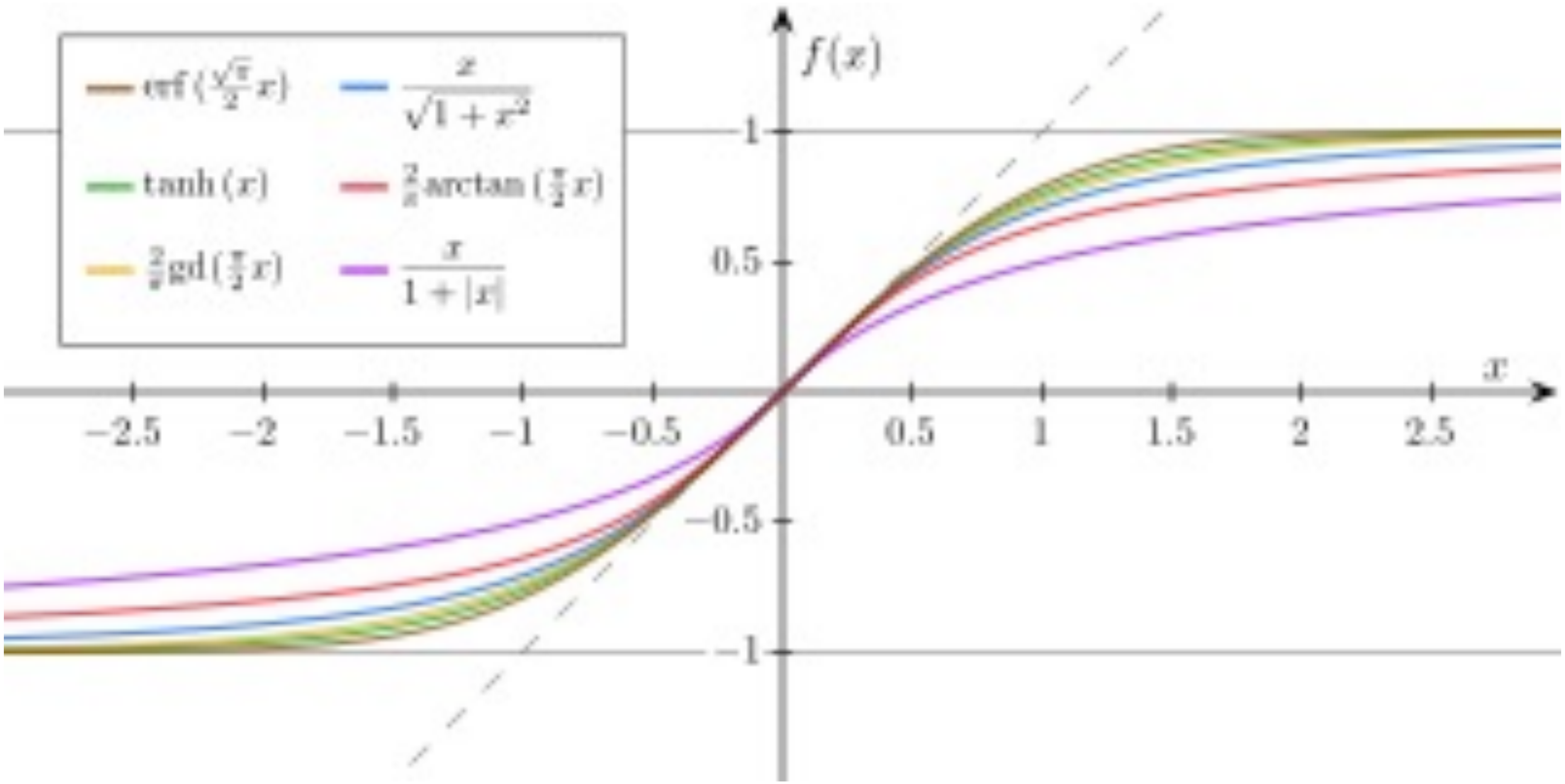
  $\log(y) = a + bx$

  $\log(y) = a + b\log(x)$

  $y = a + bx + cx^2$

- The relationship might follow a sigmoid structure.
- Essentially there are many mathematical formulae that have the appropriate structure.

Sigmoids



$\mathrm{erf}\left(\frac{\sqrt{\pi}}{2}x\right)$
$\frac{x}{\sqrt{1+x^2}}$
$\tanh(x)$
$\frac{2}{\pi}\arctan\left(\frac{\pi}{2}x\right)$
$\frac{2}{\pi}\mathrm{gd}\left(\frac{\pi}{2}x\right)$
$\frac{x}{1+|x|}$
$f(x)$
x
1
0.5
−0.5
−1
−2.5
−2
−1.5
−1
−0.5
0.5
1
1.5
2
2.5

Computational approaches

- For least squares estimation of pseudo linear models, it is possible to use singular value decomposition to estimate parameters
- This provides a unique solution which is known to coincide with a global optimum
- Estimating parameters of nonlinear systems often requires more complicated optimization techniques

Nelder–Mead optimisation

- The derivatives of the parameter space will generally be unknown
- Nelder–Mead (also known as the downhill simplex) method is a commonly used nonlinear optimization technique and is available as fminsearch in Matlab
- This approach is heuristic and often converges to local optima

Problems with local minima

- Suppose that the objective is to minimize a cost function f(**a**) given parameter vector **a**
- Due to the issues of local minima, it is necessary to consider alternative optimization techniques
- One such approach is to make N random samples of the parameter space inside specific boundaries and calculate $f_i = f(\mathbf{a}_i)$ for i=1,…,N

Shotgun approach

- It is then possible to take a fraction of the lowest values of f, say f_j for $\mathbf{a}_j$ for j=1,…,M
- Then by treating each of these parameter vectors as an initial condition, we use fminsearch to converge to the local minima $\mathbf{a}_j$* in the neighborhood of $\mathbf{a}_j$
- Finally we identify the parameter estimate as that **a** which gives the minimum

Poll

- Name some nonlinear models that you have heard about.

- **Slido.com**
- **#50845**

Examples of nonlinear models

- Global polynomial models
- Local polynomial models
- Threshold autoregressive (TAR)
- Smooth Transition autoregressive (STAR)
- Markov-Switching AR
- Exponential Smoothing

The pseudo-inverse

- The model parameters ***a*** are determined by solving the linear system of equations ***b = Ha***, where the design matrix is ***H*** and the dependent variable is represented by $b_i = s_{i+1}$
- Obtain parameters ***a*** which minimise

$$\chi^2 = \|\mathbf{b} - \mathbf{H}\mathbf{a}\|^2$$

- Both $||\boldsymbol{b} = \boldsymbol{Ha}||^2$ and $||\boldsymbol{a}||$ are minimised by choosing $\boldsymbol{a} = \boldsymbol{H}^{\dagger}\boldsymbol{b}$, where $\boldsymbol{H}^{\dagger}$ is the Moore-Penrose pseudo-inverse of ***H***

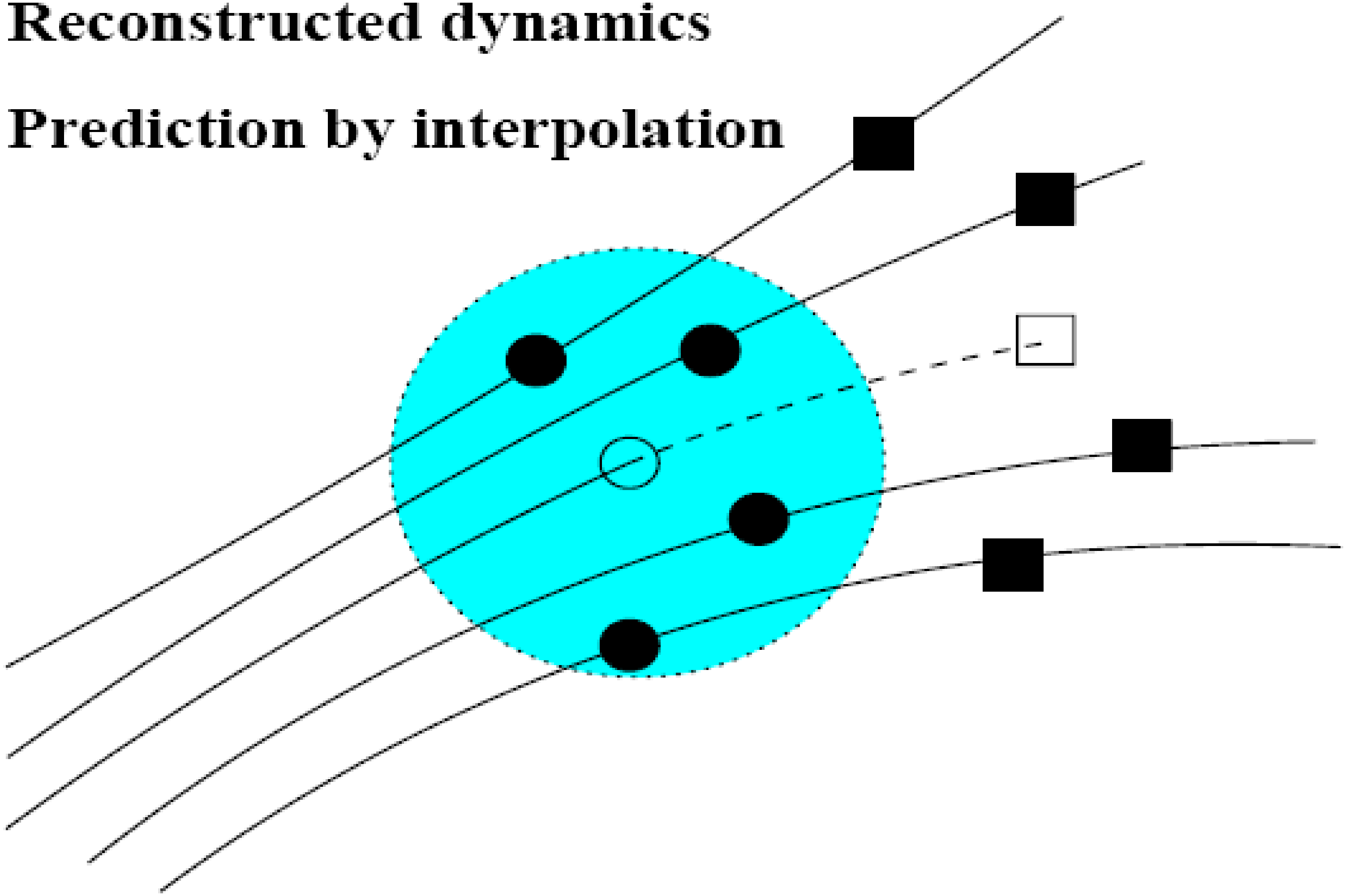
Global polynomial model

- Global polynomial model structure:

$$F(\mathbf{s}) = a_0 + \sum_{j=1}^{m} a_j s_{i-(j-1)\tau} + \sum_{j=1}^{m}\sum_{l=1}^{m} a_{jm+l} s_{i-(j-1)\tau} s_{i-(l-1)\tau} + \cdots$$

- Advantages:
  - Easy to understand and employ
  - Estimate parameters using SVD
  - Known to converge (via the Weierstrass approximation theorem)
- Disadvantages:
  - Number of parameters required for a polynomial of order *n* in an *m*-dimensional space is *(m + n)!/m!n!*
  - Risk of over-fitting increases with polynomial order *n*
- Unfeasible to fit high order polynomials in high dimensions

Extrapolation & Interpolation



Reconstructed dynamics

Prediction by interpolation

Local analogue & local average

- To make a prediction from ***s**$_i$*:
- First define a local neighbourhood *B(**s**$_i$)* about ***s**$_i$*:
  - select *k* nearest neighbours from the learning data, or
  - select all nearest neighbours within distance *r* of ***s**$_i$*
- Local analogue: use the future of the nearest neighbour as the prediction,

$$\hat{s}_{i+1} = s_{j+1} \text{ where } j \text{ minimises } \| \mathbf{s}_i - \mathbf{s}_j \|$$

- Local average: use the average of the future of all neighbours found in the neighbourhood

$$\hat{s}_{i+1} = \frac{1}{|B(\mathbf{s}_i)|} \sum_{j \in B(\mathbf{s}_i)} s_{j+1}$$

where |*B(**s**$_i$)*| denotes the number of state vectors in *B(**s**$_i$)*

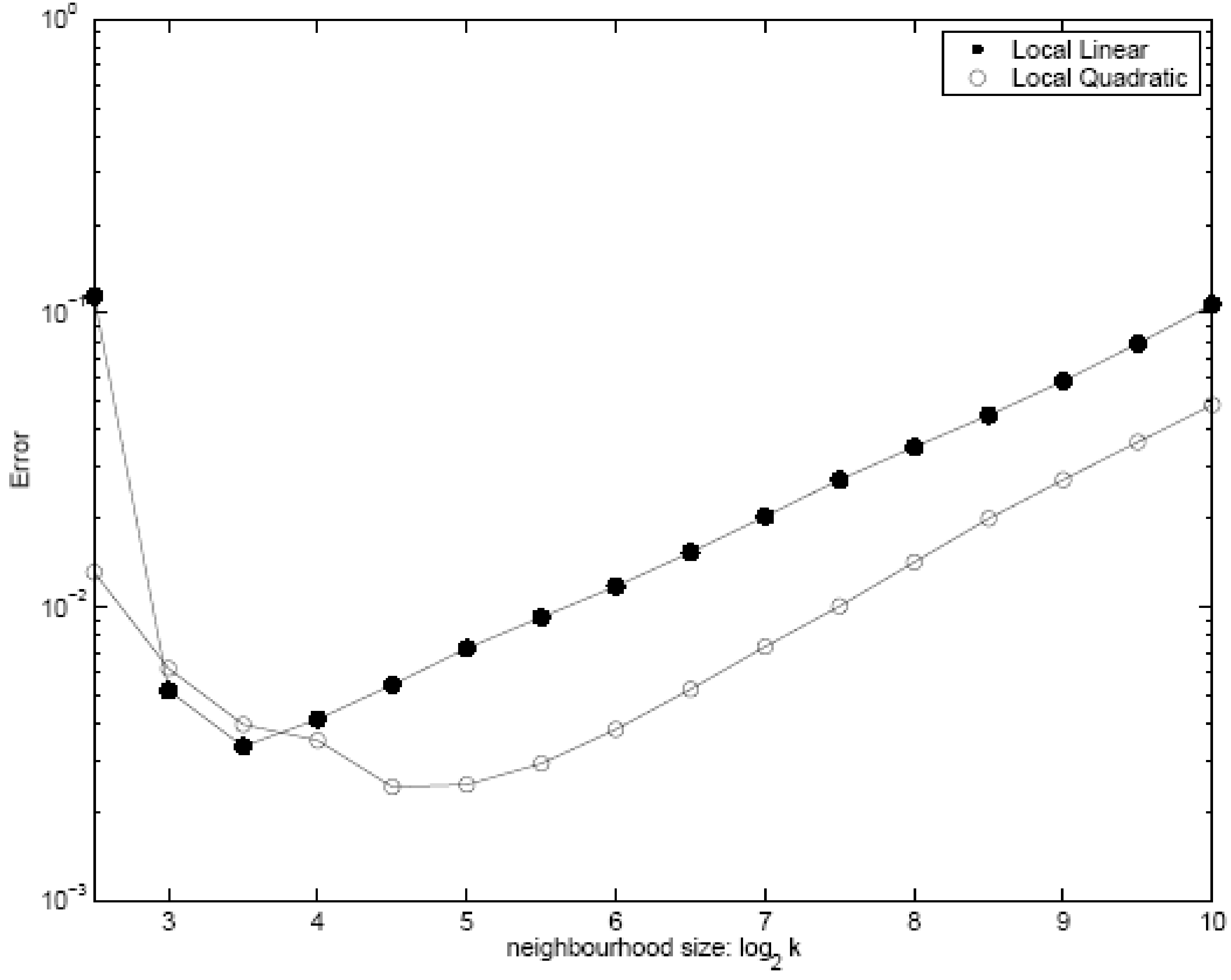
Local average and local analogue

- Note that as the number of neighbours tends to the total number of points in the training data set, the local average forecast becomes the unconditional mean
- Use cross-validation to estimate the optimal neighbourhood size by evaluating the model on out-of-sample data as a function of *k* or *r*
- Neighbourhood size balances resolution and noise with neglected nonlinear terms
- Dynamically adaptive and can be used with data streams (easily updated)
- Random Analogue Prediction (RAP): Choose near neighbour at random and use its future as a prediction (stochastic prediction)

Local polynomial modelling

- First define a local neighbourhood $B(\boldsymbol{s}_i)$ about $\boldsymbol{s}_i$:
  - select *k* nearest neighbours from the learning data, or
  - select all nearest neighbours within distance *r* of $\boldsymbol{s}_i$
- Fit a linear model using only these nearest neighbours in $B(\boldsymbol{s}_i)$
- $$F(\mathbf{s}) = a_0(\mathbf{s}_i) + \sum_{j=1}^{m} a_j(\mathbf{s}_i) s_{i-(j-1)\tau} + \sum_{j=1}^{m}\sum_{l=1}^{m} a_{jm+l}(\mathbf{s}_i) s_{i-(j-1)\tau} s_{i-(l-1)\tau} + \cdots$$
- Use Singular Value Decomposition (SVD) to solve the matrix equation
- Useful property: as *k* tends to the number of points in training set, the local polynomial model becomes a global polynomial model
- Use cross-validation to estimate the optimal neighbourhood size by evaluating the model on out-of-sample data as a function of *k* or *r*
- Neighbourhood size balances resolution and noise with neglected nonlinear terms
- Dynamically adaptive and can be used with data streams (easily updated)

Detecting low-dimensional dynamics



Regime switching models

- Existence of distinct regimes of activity
- Business cycles (expansion and contraction)
- Assume each regime can be described using a traditional linear model, AR(p)
- Regime may depend on external factors
- Regime can also be specified using a lag of the observed variable

Threshold autoregression

- Threshold autoregression (TAR) was proposed by Tong (1978)
- The TAR model structure is given by:

$$y_t = \phi_0^{(i)} + \sum_{j=1}^{p_i} \phi_j^{(i)} y_{t-j} + \varepsilon_t^{(i)} \text{ if } y_{t-d} \in R_i, i = 1,2,\ldots,k$$

- The *k* partitions, *R_i*, correspond to each regime and form a non-overlapping partition of the real line
- *d* is the delay parameter (or threshold lag),
- $p^{(i)}$ is the AR order in the ith regime
- $\varepsilon^{(i)}_t$ is a sequence of i.i.d. normal random variables such that $\left\langle \varepsilon^{(i)} \varepsilon^{(j)} \right\rangle = \sigma^{(i)} \delta_{ij}$

TAR applications

- Tong and Lim (1980) showed that the TAR model is capable of producing the asymmetric, periodic behaviour exhibited in the annual sunspot data and the Canadian Lynx data.
- Tyssedal and Tjostheim (1988) applied the TAR model to daily closing prices of IBM between 1959 and 1962
- Tong (1990) analysed the Hang Seng Index from 1984 to 1987
- Pope and Yadav (1990) employed a TAR model to characterise mispricing of the FTSE 100 index futures
- Cao and Tsay (1993) investigated monthly volatility
- Tiao and Tsay (1994) employed a TAR model to investigate the cyclical properties of US GNP
- Gao and Wang (1999) analysed the non-linear dynamics of the S&P 500 index

SETAR models

- A self-exciting threshold autoregressive (SETAR) model is a piecewise linear model
- SETAR provides local linear approximations using distinct AR models while switching between different regimes based on the value of the delay variable y_{t-d}
- In a financial setting, the return generating mechanism depends on the value of the price in previous period
- The key characteristics of a SETAR model include time irreversibility, asymmetric limit cycles and jumps
- A major advantage of this model is that the parameters can be estimated using least squares

SETAR(2,p,p) model

- The simplest SETAR models provides evolution of a process governed by switching between two distinct linear auto-regressions:

$$y_t = (1 - I(y_{t-d} > r))(\alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \sigma_1 \varepsilon_t)$$
$$+ I(y_{t-d} > r)(\beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \sigma_2 \varepsilon_t)$$

- $I(y_{t\text{-}d}>r) = 1$ if $y_{t\text{-}d}>r$ and zero otherwise and $\varepsilon_t \sim N(0,1)$
- Two regimes: AR(p) with coefficients $(\alpha_0,\alpha_1,\ldots,\alpha_p,\sigma_1)$ and $(\beta_0,\beta_1,\ldots,\beta_p,\sigma_2)$ and regime parameters r and d

SETAR estimation

- Split the sample into two regimes and perform OLS is each separately
- d takes on integer values 0,1,2,…
- r is allowed to take on each value y_{t-d}
- If p is known select pair (r,d) that minimises the overall sum of the squares in each regime
- If p is unknown, use Akaike (AIC) and minimise:

$$N_L \ln \hat{\sigma}_L^2 + N_U \ln \hat{\sigma}_U^2 + 2(p+1) + 2(p+1)$$

SETAR forecasting

- No analytical expression for multi-step forecasts
- Normal forecast error; DeGooijer & DeBruin (1997)
- Monte-Carlo simulations; Clements & Smith, 1997
- SETAR(2,p,p)

$$y_t = \begin{cases} \alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \sigma_1 \varepsilon_t, & y_{t-d} \le r \\ \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \sigma_2 \varepsilon_t, & y_{t-d} > r \end{cases}$$

SETAR Monte Carlo

- One step:

$$\widehat{y}_{t+1}^{MCj} = (1 - I_t(r))(\alpha_0 + \alpha y_{t-1}) + I_t(r)(\beta_0 + \beta_1 y_{t-1})$$

- Multiple steps:

$$\widehat{y}_{t+k}^{MCj} = (1 - I_{t+k-1}(r))(\alpha_0 + \alpha \hat{y}_{t+k-1}^{MCj}) + I_{t+k-1}(r)(\beta_0 + \beta_1 \widehat{y}_{t+k-1}^{MCj}) + \zeta_{k,j}$$

$$I_{t+k-1}(r) = I(\widehat{y}_{t+k-1}^{MCj} > r)$$

- Averaging gives the MC forecast:

$$\widehat{y}_{t+k}^{MC} = \frac{1}{N} \sum_{j=1}^{N} \widehat{y}_{t+k}^{MCj}$$

Q&A