TTK4260
Multivariate data analysis and machine learning
by Jan-Øivind Lima

Notes from exam preparation

Contents

§1. Introductions and motivations

The overarching workflow for data analysis is seen below:

Example (The main piece of the data analysis workflow).
- Consider the best approach given the data and objective
- Split the data into training/validation/test
- Decide on preprocessing of the training data; save this procedure for validation and test sets
- Proposing some alternative models
- Estimating the model parameters
- Choosing the best model structure
- Model interpretation given background knowledge (theory, literature, own experience)
- Identify possible outliers and decide whether to take them out or not
- Validating the chosen model(s)
- Estimating the performance of the model, including parameter uncertainties

For modeling there are several strategies to employ, here are a few of them.

- Mechanistic models / physicas based models: based on theory, first principles and/or domain knowledge (deduction)
- Data-driven models: Models based on carefully designed experiemntal (numerical of physics) data. They can be black box, grey box or white box (induction)
- Meta Modelling / Hybrid Modelling: A combination of the above two.

Some desired characteristics in the modeling approach are:
- Generalizability / Robustness
- Trustworthiness / Transparent / Explainable
- Computational efficiency yet accurate
- Dynamically adapting and evolving

§1.1. Dimensionality of a model

Most systems are observed iwth more sensors/variables than the actual underlying dimensionality(rank). Methods based on latent varibales will summarize the information in terms of

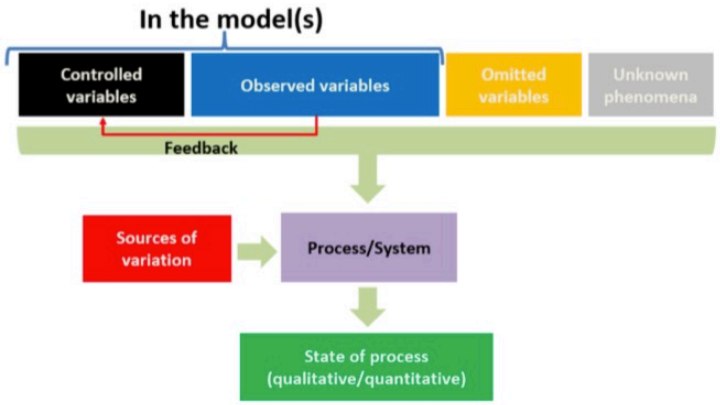


Figure 1: Variables in a system

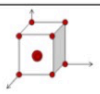
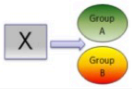
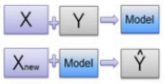
Tool	Symbol	Purpose
Design of Experiments (DoE) Full and Fractional Factorial Designs, Optimisation Designs		Gain maximum information from a minimum of experimental effort
Exploratory Data Analysis (EDA)		Find important sample groups or variable relationships in large data sets. Define classes.
Classification/Discrimination		Classify samples. Identify raw materials. Is the process under control (MSPC)?
Regression and Prediction		Predict quality for new samples and/or find out which variables that most influence a process

Figure 2: Tools used in this course

"super varibles" and reveal the individual variable's contribution to these. The rank of a system can mean different things:

- The **numerical** rank (This will for all real data be the smallest dimension of the data table)
- The **statistical** rank (found by some statistical criterion, e.g. the chosen validation scheme or maximum likelyhood)
- The **application** specific rank based on experience and interpretation

§1.2. Redundancy and selectivity

Most sensors are not measuring directly the inherent state or property of a system. Although individual sensors are not selectie, ,we can still use them for observin the system:

- **Qualitative**: Is the system under control?
- **Quantitative**: What is the water content in this rock on Mars?

§1.3. Typical pitfalls for valitation of classification or prediction models

- No independent test set
- Replicates in training and test sets respectively
- Use information in the model that is not present during prediction (for example using time as a variable)
- Selection bias (especially when tuning hyper-parameters)
- Extending sample sets with simulated samples to e.g. handle skewed distributions, then split in training/test
- Feature selection on all data, then split in training/test (see above)

- And what is not often mentioned: Random training/test when the samples must be stratified due to meta-information

§2. Principal Component Analysis (PCA) introduction

PCA is the "mother" of all exploratory multivariate methods, below are some examples demonstrating the need/utility of the exploratory data analysis.

Example.
- Pattern recognition
- Dimensionality reduction
- Clustering and classification
- Outlier detection
- Condition monitoring and predictive maintenance, e.g. of wind turbines
- Denoising
- Data imputation
- Speed up training of machine learning models (using latent variables)

§3. Least squares estimators

Indtroductory definitions:

- $y \in \mathcal{Y}$: random variable (r.v.) from a probability distribution P:

$$y \sim P(\theta) \qquad \theta \in \Theta$$

(Example: $y \sim \mathcal{N}(m, \sigma^2)$)

- $y_1, \ldots, y_N = \boldsymbol{y} \in \mathcal{Y}^N$: sampled data

Definition (statistic)

$$\phi(\boldsymbol{y}) : \mathcal{Y}^N \mapsto \mathbb{R}^M$$

is a statistic of $\boldsymbol{y}$ if
- *it is measurable*
- *it does not depend on θ*

Figure 3: Introductory definitions of least squares estimators

§3.1. Estimator

Estimator of θ

- $y \in \mathcal{Y}$: random variable (r.v.) from a probability distribution P:

$$y \sim P(\theta) \qquad \theta \in \Theta$$

- $y_1, \ldots, y_N = \boldsymbol{y} \in \mathcal{Y}^N$: sampled data

Definition (estimator (of θ))

any statistic of $\boldsymbol{y}$ with co-domain equal to Θ, i.e., any

$$\phi(\boldsymbol{y}) : \mathcal{Y}^N \mapsto \Theta$$

that

- *is measurable*
- *does not depend on θ*

Figure 4: Estimator

§3.2. Least sqares form geometrical interpretations

First the assumptions:

data generation model: $y_t = f(u_t; \theta) + v_t$

dataset: $\mathcal{D} = \{(u_t, y_t)\}_{t=1,\ldots,N}$

hypothesis space: $\theta \in \Theta$

Problem: *find a* $\widehat{\theta} \in \Theta$ *that "best explains"* $\mathcal{D}$

Figure 5: Assupptions for model.

This leads to the geometrical interpretation:

$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$ fixed $\quad \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{bmatrix}$ fixed $\quad \begin{bmatrix} f(u_1;\theta) \\ f(u_2;\theta) \\ f(u_3;\theta) \\ \vdots \end{bmatrix}$ manifold in θ

example with $\theta \in \mathbb{R}^2$:

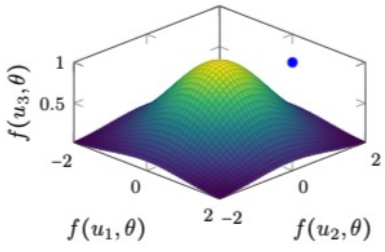


Figure 6: Geometrical interpretation of least squares

§4. Statistical performance indexes

There are three measures of central tendencies, they are **mode**, **median** and **mean**.

Example.
- **Mode**: most commonly observed value
- **Median**: midpoint
- **Mean**: average

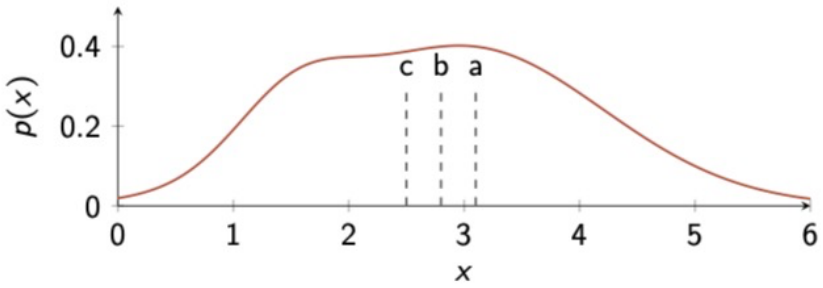


Figure 7: Central tendency

Measures of dispersion are **range**, **variance** and **standard deviation**.

Example.
- **Range**: maximum - minimum
- **Variance**: mean of the squared differences between the elements of a dataset and their mean
- **Standard deviation**: square root of the variance

Measures of association are **covariance** and **correlation**.

Example.
- **Covariance**: measure of the joint variability of two random variables
- **Correlation**: measure of the strength and direction of the linear relationship between two random variables

Some other measures are **skewness** and **kurtosis**:

Example.
- **Skewness**: how symmetric a probability distribution is
- **Kurtosis**: how tailed a probability distribution is

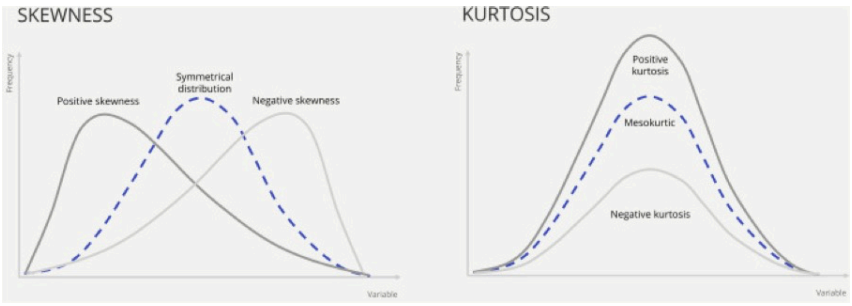


Figure 8: Skewness and kurtosis

Some units discussed in the lecture:

Example.
- regression metrics (MAE, MSE, RMSE, R2)
- classification metrics (accuracy, precision, recall, F1-score, sensitivity, specificity, ROC, AUC)
- Computer Vision metrics (PSNR, SSIM, IoU)
- timeseries related metrics (fit)

§5. Potential exam questions *(provided on BlackBoard)*

Q&A 5.1.
- *What do we mean with internal and external validity? May you make some practical examples?*
  - Internal validity: cause-effect accuracy. External validity: generalizability to other settings.

Q&A 5.2.
- *What are least squares estimators? What is the geometrical intuition behind them? And how can they be formulated mathematically?*
  - Minimizes squared errors; geometrically, orthogonal projection onto model space.

Q&A 5.3.
- *What are maximum likelihood estimators? What is the probabilistic intuition behind them? And how can they be formulated mathematically?*
  - Find parameters that maximize data likelihood under assumed distribution.

Q&A 5.4.
- *Why should one divide a dataset into training, test, and validation sets? And how should one select such sets from the original dataset?*
  - To avoid overfitting and assess generalization. Split randomly or with cross-validation.

Q&A 5.5.
- *What does cross validation mean? How should it be used? And why should one use a cross validation approach instead of using a training, test, and validation sets based approach?*
  - Splits data into folds; rotates training/testing for robust performance estimate.

Q&A 5.6.
- *What are maximum a posteriori estimators? What is the probabilistic intuition behind them? And how can they be formulated mathematically?*
  - Like MLE but includes a prior. Maximizes posterior probability.

Q&A 5.7.
- *What are the conditions about existence and uniqueness of the LS, ML and MAP estimates?*
  - Require full-rank data matrix; uniqueness if objective is strictly convex.

Q&A 5.8.
- *Derive the maximum likelihood estimator for the separable problem* $yi = \vartheta ui + vi$ *with* $vi \sim N(0, \sigma^2)$ *with* σ^2 *known,* ϑ *unknown and deterministic*
  - MLE: $\hat{\vartheta} = (\Sigma ui^2)^{-1}\Sigma uiyi$; derived by maximizing normal likelihood.

Q&A 5.9.
- *Derive and comment the bias-variance tradeoff*
  - High bias = underfit; high variance = overfit. Tradeoff affects generalization.

Q&A 5.10.
- *What do we expect to see when training and testing different estimators with different model order complexities? How should we account for the effects that we see on the statistical performance indexes?*
  - Training error ↓, test error has U-shape. Use validation/cross-validation to find sweet spot.

Q&A 5.11.
- *Which statistical performance indexes would you consider when dealing with a regression problem? And which peculiarities / usages do they have?*
  - MSE, RMSE, R^2; MSE penalizes outliers, R^2 shows variance explained.

Q&A 5.12.
- *Which statistical performance indexes would you consider when dealing with a classification problem? And which peculiarities / usages do they have?*
  - Accuracy, precision, recall, F1; choose based on class imbalance.

Q&A 5.13.

- *What does "design of experiments" mean? And "factorial design"? Which alternative factorial-design based alternatives do you know, and what are the tradeoffs among them?*
  - DOE: structured experiments. Factorial: test all factor combinations. Alternatives: fractional, reduces runs.

Q&A 5.14.
- *What does PCA mean from a geometrical point of view? How is it formulated mathematically, and how does it connect with SVD?*
  - Projects data onto directions of max variance. Uses SVD of data matrix.

Q&A 5.15.
- *What are the uses of PCA? And how can its results be interpreted?*
  - Dimensionality reduction, noise filtering. Scores show samples; loadings show variable influence.

Q&A 5.16.
- *What are the uses of the loadings plots and scores plots, in a PCA?*
  - Loadings: variable contribution. Scores: sample positions in PC space.

Q&A 5.17.
- *How can one decide how many components should be used when analysing some data through PCA? And how can one decide whether a sample is an outlier or not, through PCA?*
  - Use scree plot or variance explained. Outliers = distant scores/residuals.

Q&A 5.18.
- *How does the Ockham's razor principle connect with the model order selection problem? Which alternative strategies can be used to solve the model order selection problem?*
  - Simpler models preferred. Use AIC, BIC, cross-validation.

Q&A 5.19.
- *What does "rotated PCA" mean? How does this concept connect with PCA, from both geometrical and mathematical points of view?*
  - Rotates PCs to simplify interpretation. Maintains variance, changes axes.

Q&A 5.20.
- *Which type of problems does the ICA algorithm solve? Which assumptions does it require? And how does it work, from intuitive perspectives?*

- Separates mixed signals assuming non-Gaussianity and independence.

Q&A 5.21.
- *What does "total least squares" mean? How does this concept connect with least squares, from both geometrical and mathematical points of view?*
  - Accounts for errors in both X and y. Minimizes orthogonal distances.

Q&A 5.22.
- *What does ANOVA mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - Tests group means. Uses variance decomposition. Practical in experiments.

Q&A 5.23.
- *What does PLS mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - Projects predictors to explain both X and y. Good for multicollinearity.

Q&A 5.24.
- *What does MLR mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - Linear relation: $y = X\beta + \varepsilon$. Used for prediction and inference.

Q&A 5.25.
- *What does PCR mean? Which type of analyses does this approach serve? How is it formulated mathematically, and how can one use it in practice?*
  - PCA + regression on PCs. Reduces multicollinearity.

Q&A 5.26.
- *How do PLS, MLR, and PCR relate to each other? In which cases does one expect one of them to work better than the other ones, and viceversa?*
  - MLR needs full-rank X; PCR and PLS work with correlated X; PLS often better with small data.

Q&A 5.27.
- *What is the NIPALS algorithm? How does it work, from a graphical perspective? Which advantages does it bring over SVD, when used to compute a PCA?*
  - Iterative PCA for one component at a time. Handles missing data.

Q&A 5.28.
- *What does metamodelling mean? When would one want to use a metamodelling approach? What are the potential shortcomings of a metamodel?*
  - Approximate model of a complex system. Faster but may lack accuracy.

Q&A 5.29.
- *What do "stationarity" and "ergodicity" mean? Why are these two concepts important when dealing with statistical analyses of time series? And what would the lack of stationarity and ergodicity imply in practice?*
  - Stationarity: stats constant over time. Ergodicity: time average ≈ ensemble average. Needed for reliability.

Q&A 5.30.
- *Which LTI model structures do you know that are suitable to do control-oriented modelling of discrete time MISO systems?*
  - ARX, ARMAX, OE, BJ; choice depends on noise structure and goals.

Q&A 5.31.
- *What is the principle behind prediction error methods? When should the focus be on prediction errors, when identifying a dynamical model?*
  - Minimize error between predicted and actual outputs. Use for forecasting.

Q&A 5.32.
- *What are the implications of choosing an ARX, instead of an ARMAX, instead of an OE model structure when doing system identification? And what are the implications of choosing different model orders? How should one choose a specific structure and order?*
  - ARX is simple, ARMAX models noise, OE models dynamics. Use AIC/BIC.

Q&A 5.33.
- *What are Hammerstein Wiener models, and what are their usages?*
  - Block models with static nonlinearity + linear dynamics. Used in nonlinear systems.

Q&A 5.34.
- *What is a p value? How should it be computed? What is its usage for? And its drawbacks?*
  - Probability of data under H_0. Used for hypothesis testing. Can be misinterpreted.

Q&A 5.35.

- *What is a statistical test? How can it be interpreted from geometrical perspectives? And from mathematical perspectives?*
  - Decision rule to accept/reject H_0. Geometrically: distance from null region.

Q&A 5.36.
- *What are the statistical performance indexes associated to a statistical test? And which concepts may one use to say that a test is "better" than another one?*
  - Type I/II errors, power. UMP tests maximize power under constraints.

Q&A 5.37.
- *What are the differences between simple and composite hypotheses? How do the formulations of hypothesis testing algorithms change, depending on which type of hypothesis is considered?*
  - Simple: one distribution. Composite: a range. Testing becomes more complex.

Q&A 5.38.
- *What does the Linear Discriminant Analysis algorithm do? How? Which advantages and disadvantages does it have?*
  - Finds linear boundary to separate classes. Works well for Gaussian classes.

Q&A 5.39.
- *What does the Partial least squares discriminant analysis algorithm do? How? Which advantages and disadvantages does it have?*
  - PLS adapted for classification. Handles correlated variables.

Q&A 5.40.
- *What does the logistic regression algorithm do? How? Which advantages and disadvantages does it have?*
  - Models class probability using sigmoid. Easy to interpret.

Q&A 5.41.
- *What does the Support Vector Classification algorithm do? How? Which advantages and disadvantages does it have?*
  - Finds max-margin hyperplane. Robust to overfitting.

Q&A 5.42.
- *What is the kernel trick? Where may one use it, and why?*
  - Implicitly maps data to higher dimension. Used in SVM, PCA.

Q&A 5.43.
- *What does the K-means algorithm do? How? Which advantages and disadvantages does it have?*
  - Clusters by minimizing within-group variance. Simple but needs k.

Q&A 5.44.
- *What does the DBSCAN algorithm do? How? Which advantages and disadvantages does it have? What are its differences with the k-means algorithm?*
  - Density-based clustering. Finds arbitrary shapes, handles noise.

Q&A 5.45.
- *What do decision trees and random forests do? How? Which advantages and disadvantages do they have?*
  - Trees split by features; forests use many trees. Trees overfit, forests generalize.

Q&A 5.46.
- *What is the need for features selection? And what are the differences between feature engineering and selection?*
  - Reduces overfitting and complexity. Engineering creates features; selection chooses among them.

Q&A 5.47.
- *What are the pros and cons of using a wrapper method for feature selection, or an embedded method for the same sake?*
  - Wrapper: accurate but slow. Embedded: fast and model-integrated.

Q&A 5.48.
- *Why is categorical cross entropy loss used instead of mean squared error as a cost function in classification problems as against the regression problems?*
  - Cross-entropy handles probabilities and class labels better than MSE.

Q&A 5.49.
- *Which regularization technique is more effective in feature selection? And why?*
  - Lasso (L1) promotes sparsity by shrinking some weights to zero.