HIGH DIMENSIONAL STATISTICS

SUMMARY & EXAM

2024

THE QUESTION OF THE COURSE

- How can we deal with large data sets, especially when p is large compared to n?

- What are general concepts?

- What are particular methods?

GENERAL CONCEPTS

How can we deal with large data sets, especially when *p* is large compared to *n*?

Bias-variance tradeoff

Overfitting to noise
Training vs. test data

Sparsity

Reduce overfitting by shrinkage and variable selection.
Keep models interpretable.

Resampling

Cross-validation for model building
Bootstrap for assessing uncertainty

PARTICULAR METHODS

How can we deal with large data sets, especially when p is large compared to n?

Large-scale inference

Simple univariate approach via P-values, Q-values, posterior probabilities

Dimension reduction

Transform data to only a few most informative dimensions via PCA, SVD, t-SNE, UMAP

High-dimensional regression

Use variable selection and/or shrinkage via penalized regression models

LARGE-SCALE INFERENCE

- HDS1: What are P-values? What is the multiple testing problem? How does FWER try to solve it? How does Bonferroni method work?
- HDS2: What is FDR? How is it different from raw P-value thresholds or FWER correction? How does BH method work? How does BY method work?
  - HDS_error_control (slides)
- HDS3: How to estimate π_0? What is Q-value?
- HDS4: What does Bayes formula say about turning significance threshold procedure to talk about probability of a hypothesis? And what about conditioning on the observed data, not just significance threshold? What is local FDR?

HIGH DIMENSIONAL REGRESSION

- HDS5: Need for training and test data. Overfitting. Bias-variance tradeoff. AIC and BIC, definitions and what do they approximate? Stepwise selection (forward, backward, best subset). Cross-validation.
  - Very important concepts here! See also HDS5_slides.
- HDS6: Penalized regression objective functions (ridge, LASSO, elastic net). What are the differences btw these different penalties (shrinkage vs. variable selection, prediction vs. interpretation)? Role of CV in fitting penalized regression? Pre-processing data?
  - See also HDS6_slides.
- HDS7: How does bootstrap work and what is it used for?
- HDS8: Spike-and-slab prior, posterior distribution of coefficients (incl. number of non-zeros), Susie model and credible sets

DIMENSION REDUCTION

- HDS9: What is the goal of PCA? How does PCA achieve the goal? What are the different matrix decompositions that yield PCA? What are scores and loadings and how to compute them from different matrix decompositions?
- HDS9: SVD definition and interpretation through linear mapping and through low-rank approximations to a matrix. Connections to PCA and efficient ways to compute PCA.
  - See also HDS9_slides
- HDS10: How are t-SNE and UMAP different from PCA? (No technical details of t-SNE or UMAP will be asked in exam.)

EXAM

- Mon 4.3 10.15-13.00 remotely via Moodle
- The rules:
  - 1.You are free to use all material from the course notes and exercise solutions (and elsewhere from the web for that matter).
  - 2.You are not allowed to communicate with other course participants during the exam.
  - 3.You are not allowed to ask anyone for help; whether the other person is on this course or is not on this course makes no difference.
  - 4.You are not allowed to help anyone else during the exam.
  - 5. If you have some problem during the exam, send email to matti.pirinen@helsinki.fi.

EXAM

- Mon 4.3 10.15-13.00 remotely via Moodle
- Questions
  - Explaining concepts, interpreting figures
  - Running R-commands as has been done in exercises
    - external packages that may be needed: qvalue, glmnet, BoomSpikeSlab, Rtsne, umap.
    - no computationally heavy loops or fine-tuning of plots
  - Mathematics required: basic linear algebra operations and probability
    - no complicated proofs, no need for calculus (no derivatives, no integrals)

EXAM

- Mon 4.3 10.15-13.00 remotely via Moodle
- Return your solutions by 13.00 (strict) to Moodle the same way you have returned exercises
  - Preferably use R markdown and knit your answers to PDF or HTML format
    - Check that the file you will return is shown correctly by viewer/browser
    - Check that you return the correct file