MA8701 Advanced methods in statistical inference and learning

Part 5: Closing, L22: Final lecture

Mette Langaas

4/23/23

Outline

- Learning outcomes [10.15]
- Final reading list
- Part 1: Central theoretical aspects [10.20]
- Student-presentation on outlier detection with connections to missing data imputation [10.35]
- Part 2: Central theoretical aspects [10.50]
- Part 3: Central theoretical aspects [11.15]
- Part 4: Central theoretical aspects [11.30]
- Oral exam schedule [11.45]
- Course feedback [11.55]

Learning outcome

(The student should be able to)

1. Knowledge
   - Understand and explain the central theoretical aspects in statistical inference and learning.
   - Understand and explain how to use methods from statistical inference and learning to perform a sound data analysis.
   - Be able to evaluate strengths and weaknesses for the methods and choose between different methods in a given data analysis situation.

2. Skills

Be able to analyse a dataset using methods from statistical inference and learning in practice (using R or Python), and give a good presentation and discussion of the choices done and the results found.

3. Competence

- The students will be able to participate in scientific discussions, read research presented in statistical journals.
- They will be able to participate in applied projects, and analyse data using methods from statistical inference and learning.

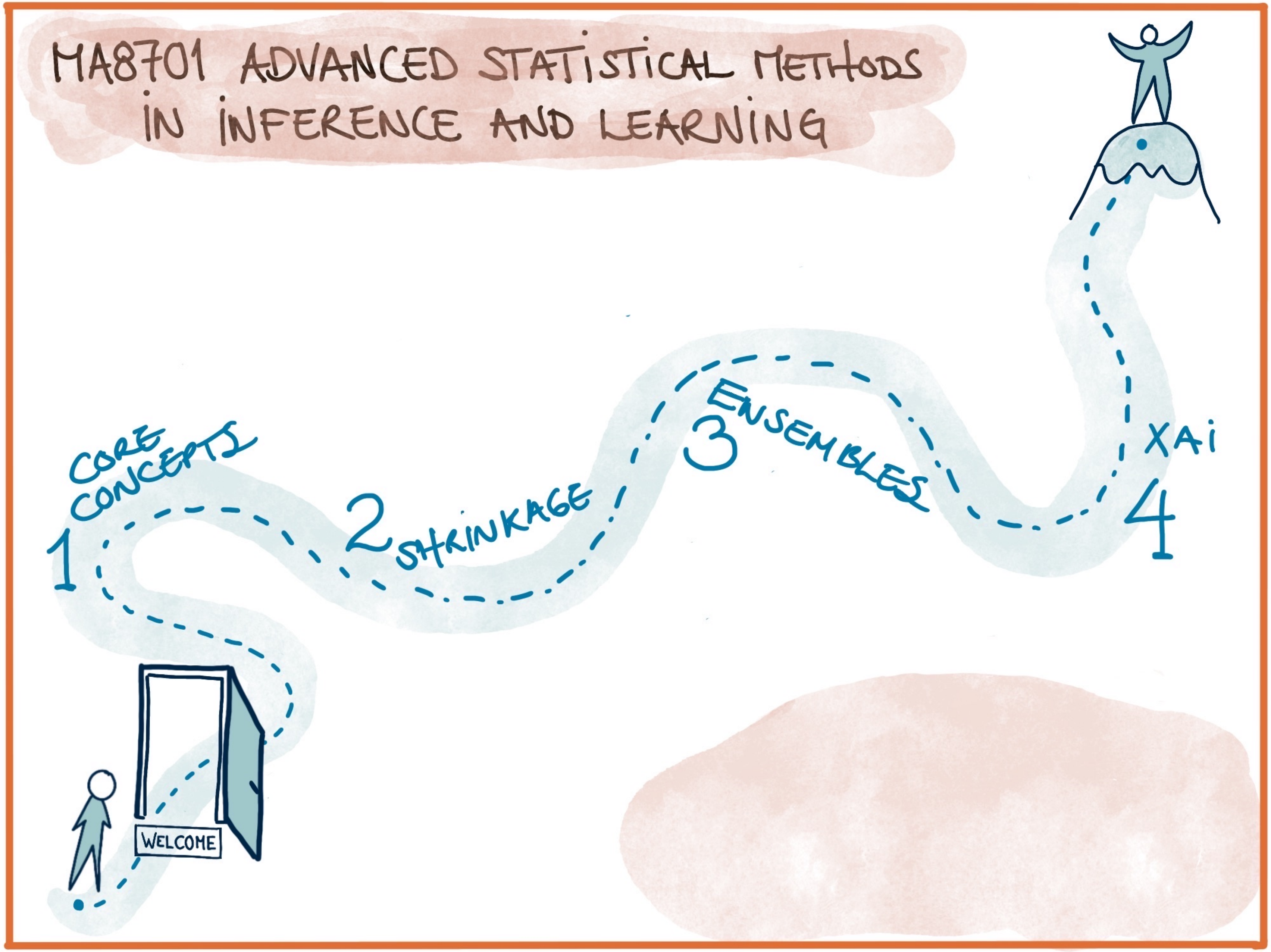
Compulsory activity 2023

- Data analysis project (analyse, write report, review)
- Article presentation (present and discuss)

Final reading list

- https://wiki.math.ntnu.no/ma8701/2023v/curriculum
- List of central theoretical aspects: https://wiki.math.ntnu.no/ma8701/2023v/curriculum#central_theoretical_concepts
- All slides/notes at the wiki https://wiki.math.ntnu.no/ma8701/2023v/handout and for Part 4 on Bb (under Handouts).

MA8701 ADVANCED STATISTICAL METHODS IN INFERENCE AND LEARNING



Part 1: Core concepts

- Starting point: decision theoretic framework
- Model selection and assessment (mind-map)
- Missing data (elements and schematic)

Starting point: decision theoretic framework

Covariates $\quad\quad\quad\quad$ response

$\mathbb{X} \in \mathbb{R}^p$ $\quad$ and $\quad$ Y, G

↑ also $\{0,1\}$

joint $p(x,y)$

conditional distribution $p(y|x)$

marginal $p(x)$, $p(y)$

Aim: to find function $f(X)$ for predicting Y (or $\hat{G}(\mathbb{X})$ for G)

one reason for LM!

mvN: $E(Y|X) = X\beta$ and $Var(Y|X)$ indep of $\mathbb{X}$

$L(Y, f(X)) = (Y - f(X))^2$

$f(x) = E(Y|X=x)$

Loss function $L(Y, f(X))$

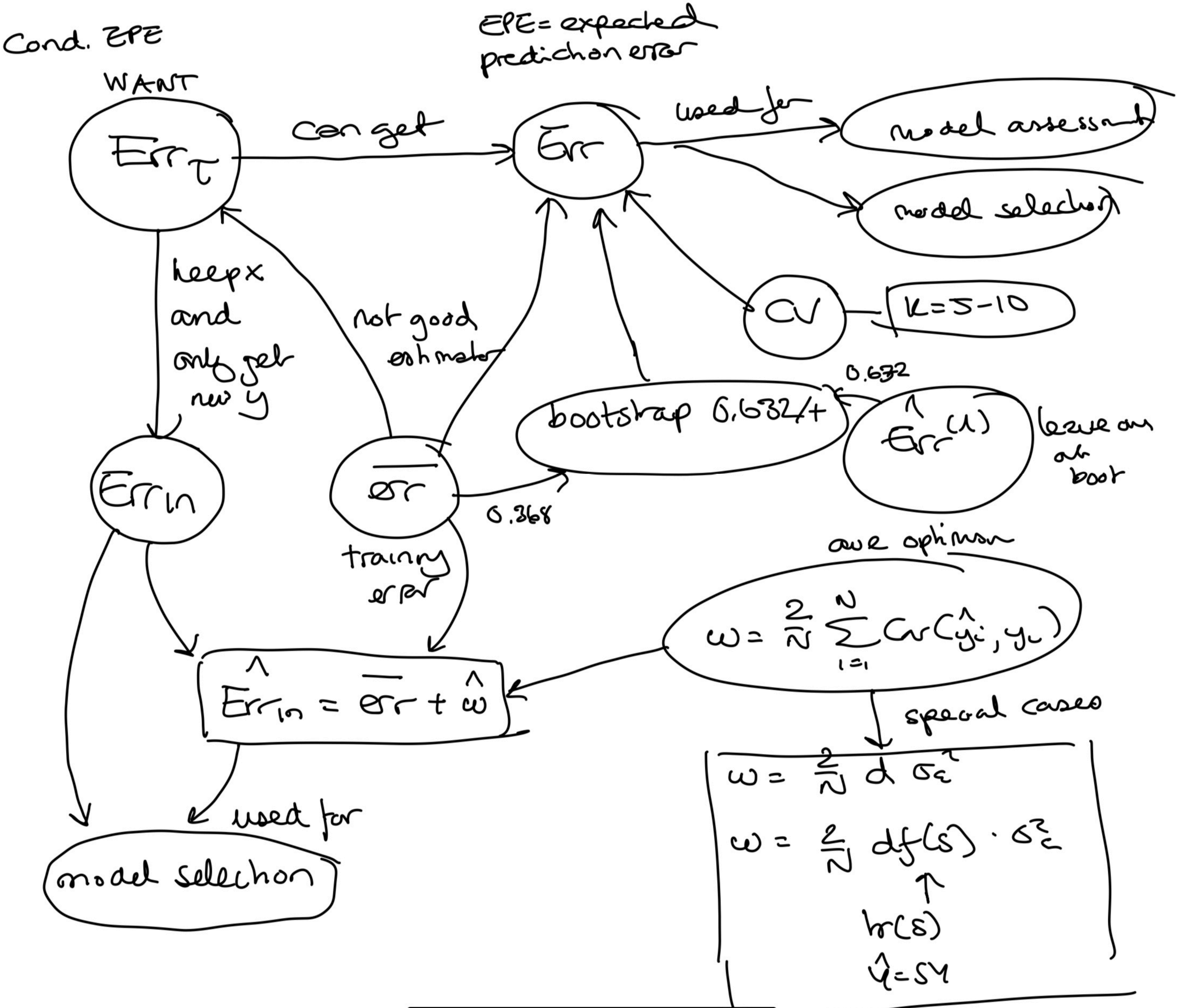
Choose f to minimize $\underset{Err}{EPE}(f) = E_{X,Y}[L(Y, f(\mathbb{X}))]$

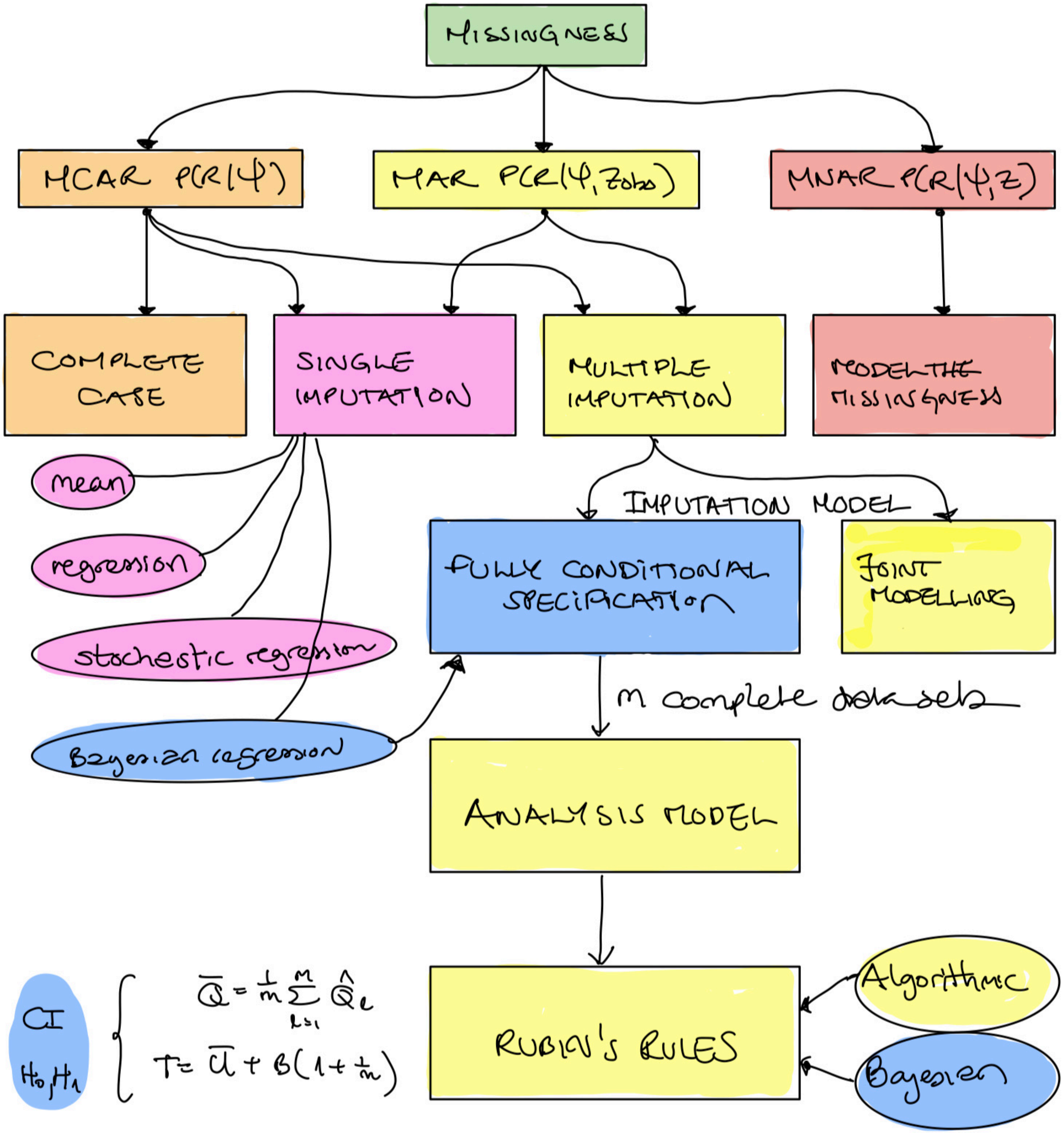
(Derive!)

$\hat{G}(x)$ = classify to the most probable class

$L(Y, \hat{G}(x)) = \begin{cases} 0 & \text{correct class} \\ 1 & \text{wrong class} \end{cases}$

Theoretical results for the optimal $f(x)$, $\hat{G}(x)$.


Cond. EPE
WANT
Err_{τ}
can get
EPE = expected prediction error
$\overline{\mathrm{Err}}$
used for
model assessment
model selection
keep x and only get new y
not good estimator
CV
K=5-10
0.632
bootstrap 0.632/+
$\hat{\mathrm{Err}}^{(1)}$
leave one out boot
Err_{in}
$\overline{\mathrm{err}}$
0.368
training error
ave optimism
$\omega = \frac{2}{N}\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$
$\hat{\mathrm{Err}}_{in} = \overline{\mathrm{err}} + \hat{\omega}$
special cases
$\omega = \frac{2}{N} d \, \sigma_{\varepsilon}^2$
$\omega = \frac{2}{N} df(S) \cdot \sigma_{\varepsilon}^2$
$tr(S)$
$\hat{y} = Sy$
used for
model selection



$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l$$

$$T = \bar{U} + B\left(1+\frac{1}{m}\right)$$

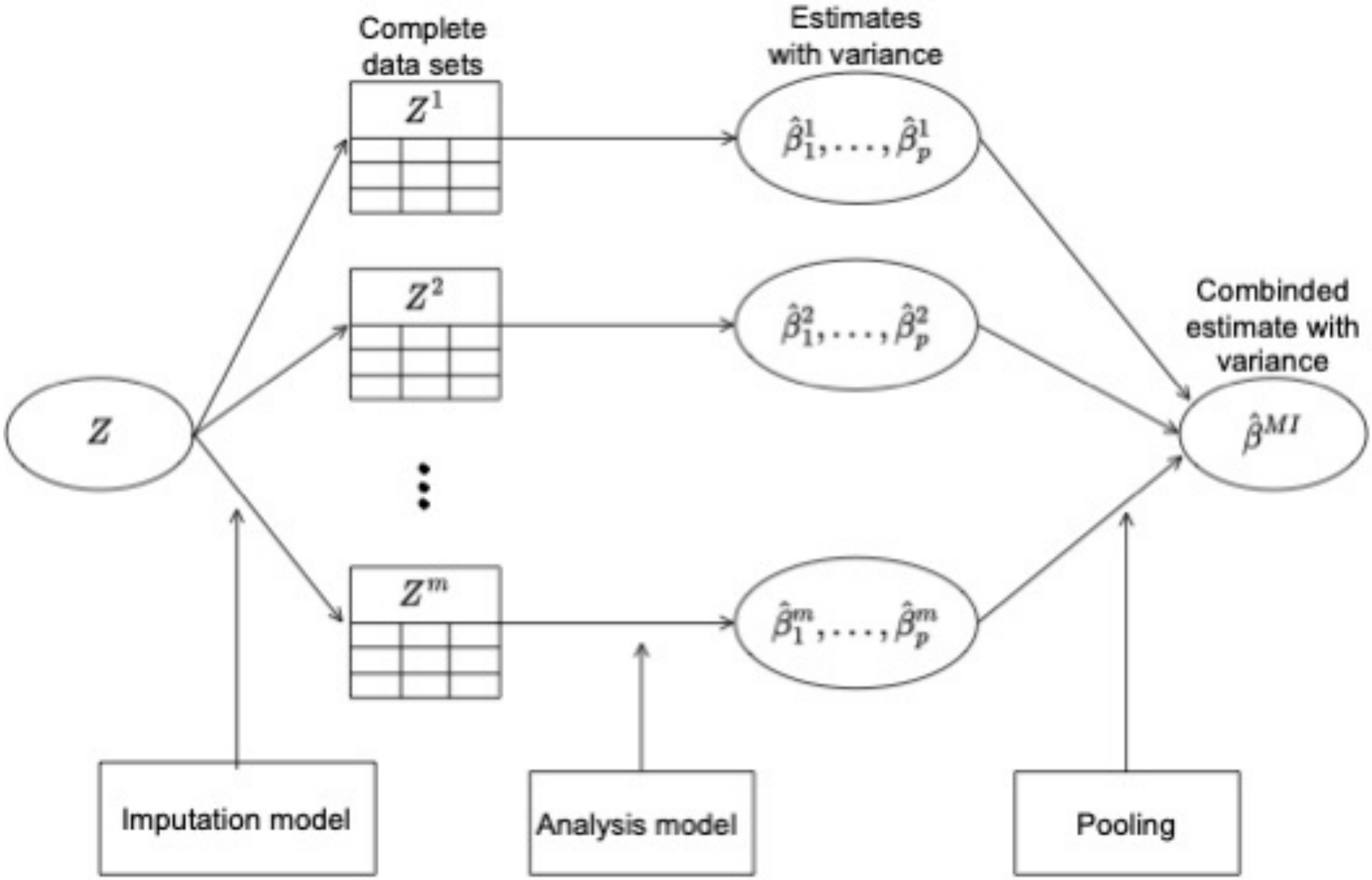

Complete data sets
Estimates with variance
Z^1
$\hat\beta_1^1, \dots, \hat\beta_p^1$
Z^2
$\hat\beta_1^2, \dots, \hat\beta_p^2$
Combinded estimate with variance
Z
$\hat\beta^{MI}$
Z^m
$\hat\beta_1^m, \dots, \hat\beta_p^m$
Imputation model
Analysis model
Pooling

Figure by Marte Bøe Ludvigsen

Prepared exam question 1) Missing data

We have discussed different methods for handling missing data, and the most complex of these is multiple imputation. Main elements of the multiple imputation method is the imputation model, the analysis model and use of Rubins rules for pooling of results.

a) If the imputation model is based on chained equations, what are the choices we need to make to run the model?

b) Let us say I have made m complete datasets in a), and then I perform logistic regression as my analysis model. What do I do with the m sets of regression estimates I have gotten?

c) What are challenges when using multiple imputation as part of a larger data analysis set-up?

Example exam questions

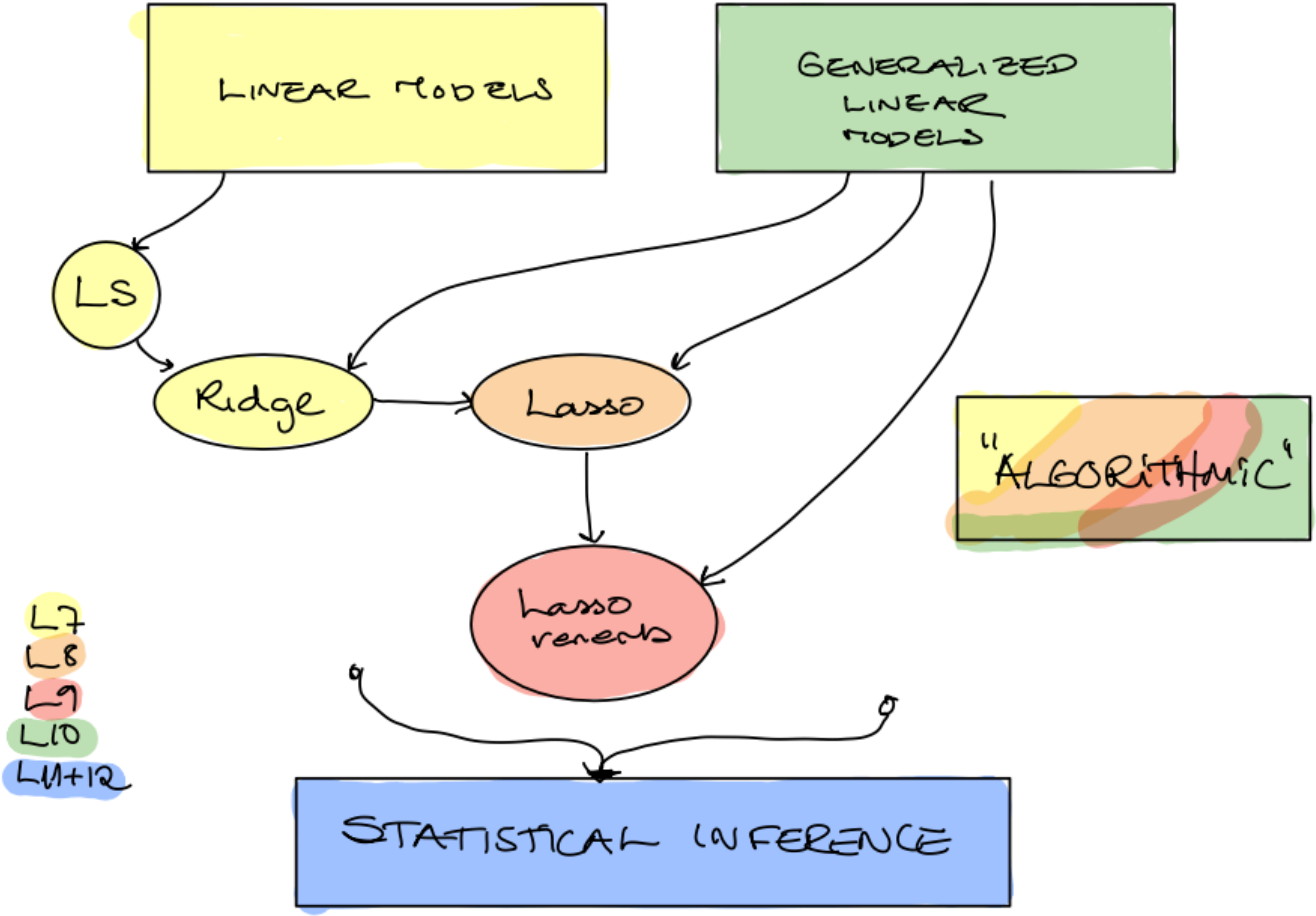
- Why do we “always” aim to estimate the conditional mean $\mathsf{E}(Y \mid X)$?
- There are a lot of Err variants in Part 1, what are core differences between the variants and how is that related to their use?
- What is Bayesian linear regression and where in the course has this played a role?
- I do not have much data (so not possible to use a train-validation-test split). How can I use my data if I both need to do model selection and model assessment?

and from the list of central topics “obvious” other questions can be found: https://wiki.math.ntnu.no/ma8701/2023v/curriculum#central_theoretical_concepts

Student-presentation on outliers and missing data imputation

Part 2: Shrinkage

(for this part we have valuable experience with from the Data analysis project!)


LINEAR MODELS
GENERALIZED LINEAR MODELS
LS
Ridge
Lasso
Lasso variants
"ALGORITHMIC"
L7
L8
L9
L10
L11+12
STATISTICAL INFERENCE

Topics

- From Gauss-Markov theorem (smallest variance among all linear unbiased estimators) to “better” biased estimators (here also theoretical excercises)
- Ridge regression - squared L2 penalty and closed form solution
- Lasso regression - the L1 penalty, only closed for one, two, and for orthogonal covariates.
- Lasso variants - solving different challenges and graphical presentation constraints
- From LM to GLM
- Parameter estimation with cyclic coordinate descent
- Statistical inference (see also article presentation CDF): bootstrapping, Bayesian interpretation, debiased lasso, multisample splitting and polyheder result.

Prepared exam question 2) Lasso regression

Lasso regression is the answer to an L1 penalty-problem, and the L1 penalty has been used in several models in this course.

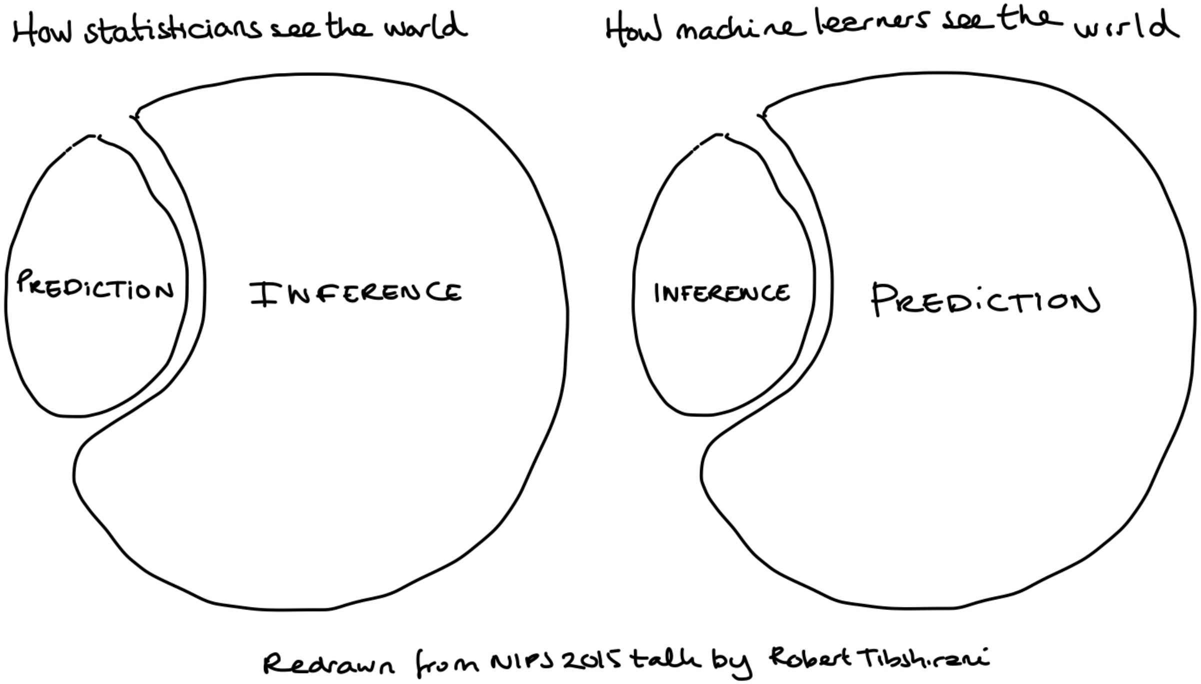
a) Write down the model and additional assumptions for a linear regression model with L1 penalty.

b) Explain how model parameters are estimated, and give properties of the parameter estimator.

c) Comment on changes needed when moving from the linear lasso regression to logistic lasso regression model.

Prepared exam question 3) Prediction vs. inference

In MA8701 one aim has been to move from a focus on prediction to statistical inference, that is moving from the right to the left figure.



How statisticians see the world
PREDICTION
INFERENCE
How machine learners see the world
INFERENCE
PREDICTION

Redrawn from NIPS 2015 talk by Robert Tibshirani

a) Choose one situation where we have done this, and elaborate.
b) What are challenges, and what is gained by using statistical inference?

(Hints for some situations: inference for the lasso, selective inference. Other situations exist.)

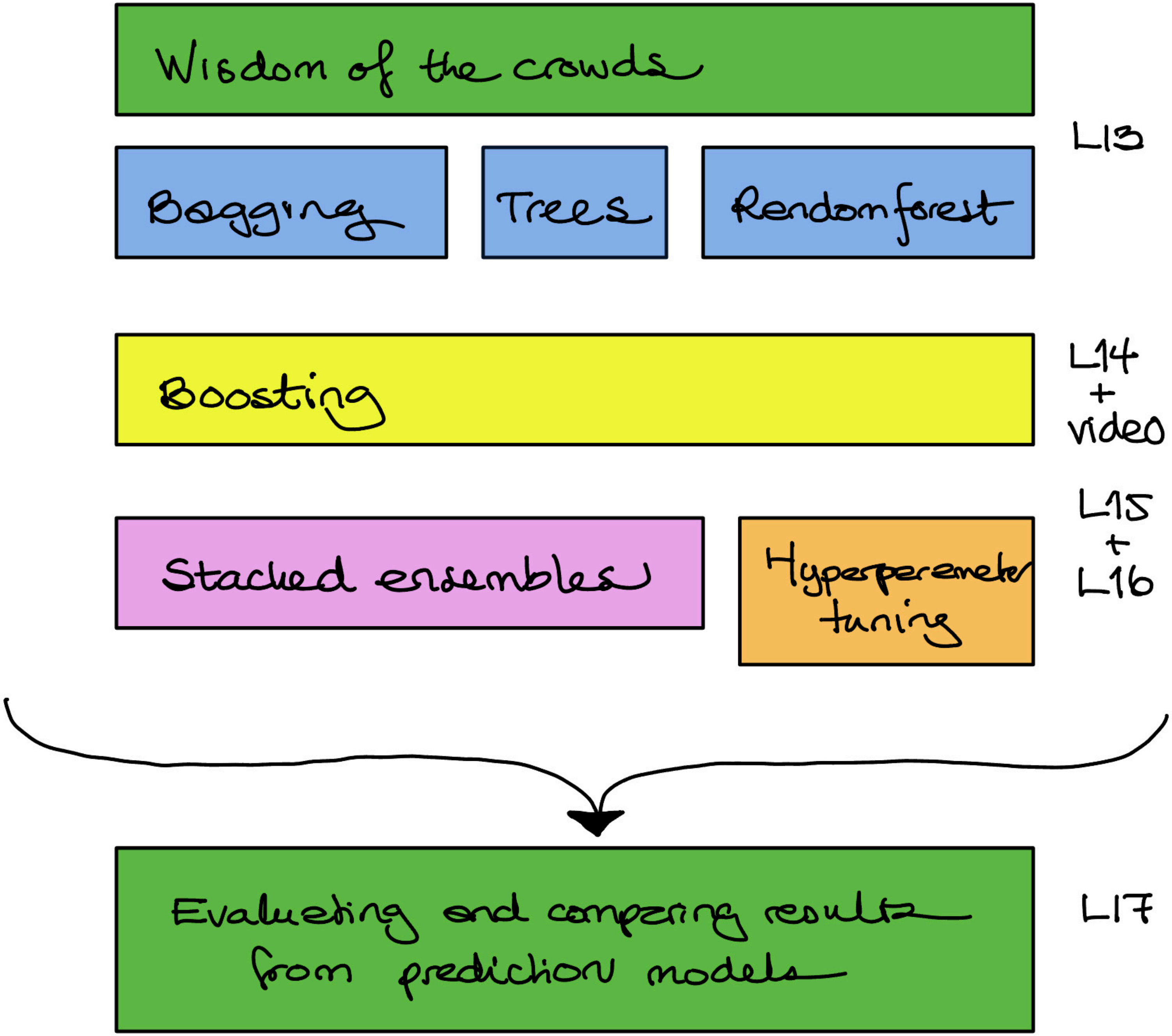
Example exam questions for Part 2

- How can we compare two estimators when $p > 1$? What is the role of positive definite matrices in the comparison?
- Explain the expression

$$\text{argmin}_{\beta}[\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2]$$

- What is it the expression used for?
- What is the bridge penalty?
- How can we construct a CI for our regression parameters when we have used the lasso for parameter estimation?

Part 3: Ensembles



Topics

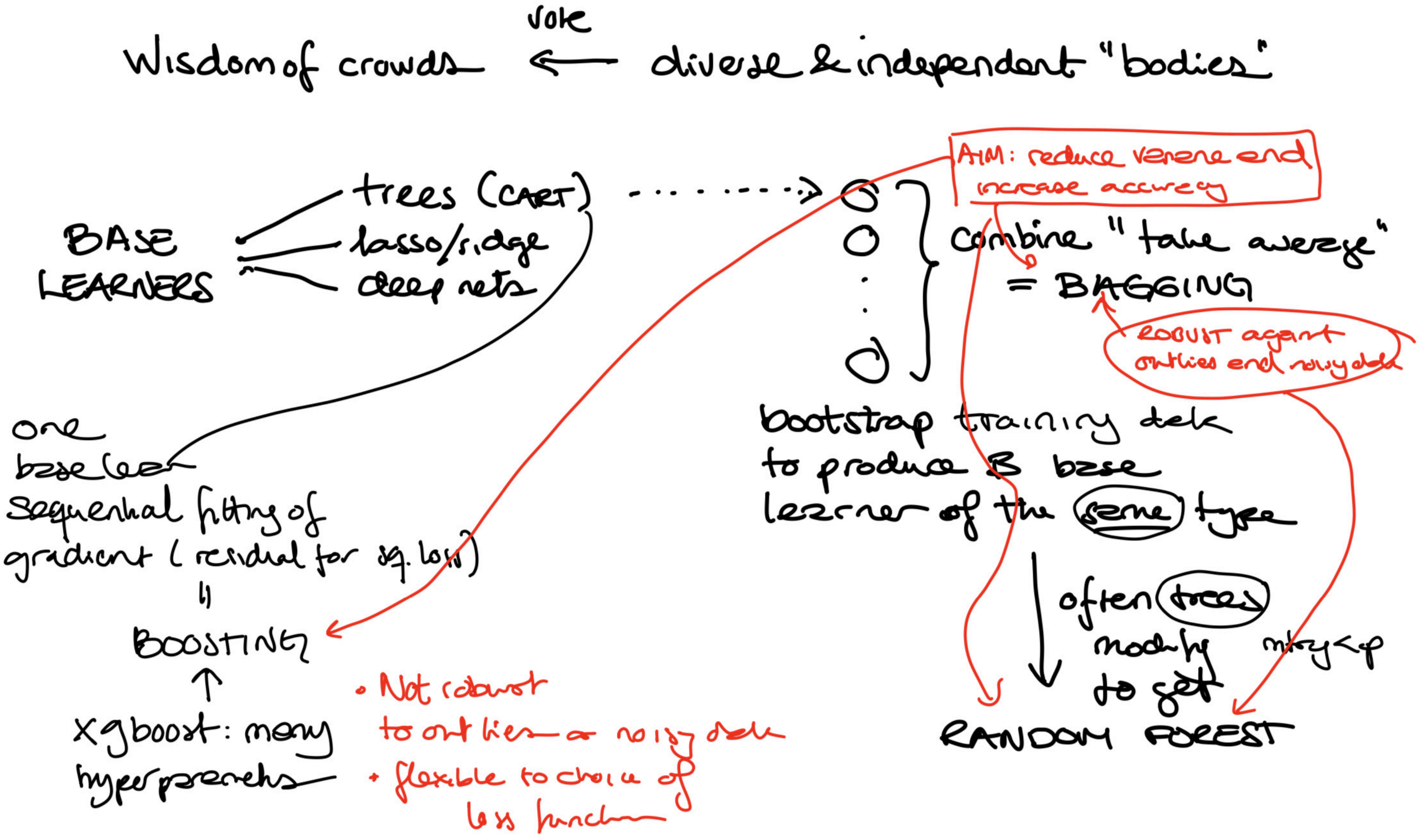
- Why do we want to form ensembles?
- Bootstrapping to form ensembles: using trees to do bagging and random forest.
- Out of bag estimation
- Ada.boost: the first boosting methods - principles.
- Stagewise vs stepwise methods.
- Depth of trees
- Gradient tree boosting and extreme gradient tree boosting.
- Stacked ensembles: training data (level 0), base learners, meta learner. Important role of v-fold CV for parameter estimation.
- Hyperparameter tuning: grid search and iterative search. Bayesian optimization
- Data rich situation: inference on test set, data poor: inference based on test folds in CV.

Prepared exam question 4) Ensembles

An ensemble can be constructed in different ways. Assume that our aim is regression (to minimize squared loss) and we use regression trees (with binary splits) as base learners.

a) What are the main differences between a random forest and a gradient boosting tree?

b) Which statistical principles (that we have learned about in this course) are used when moving from the gradient boosting tree to the xgboost? Hint: many are related to xgboost hyperparameters.



Wisdom of crowds ← vote diverse & independent "bodies"

BASE LEARNERS
- trees (CART)
- lasso/ridge
- deep nets

AIM: reduce verience end increase accurecy

combine "take averege" = BAGGING

ROBUST agenst outlies end noisy data

bootstrap training data to produce B base learner of the same type

often trees modifying to get RANDOM FOREST

mtry < p

one base learner sequential fitting of gradient (residual for sq. loss)

⇓

BOOSTING

↑

Xgboost: many hyperparemeters

- Not robust to outlies or noisy data
- flexible to choice of loss function

Stacking: simple + more complex base learners → best to have a diverse set

- asymptotic oracle properties for lin com in meta learner
- can be used to "avoid hyperparameter tuning"

- avoid model selection - no waste

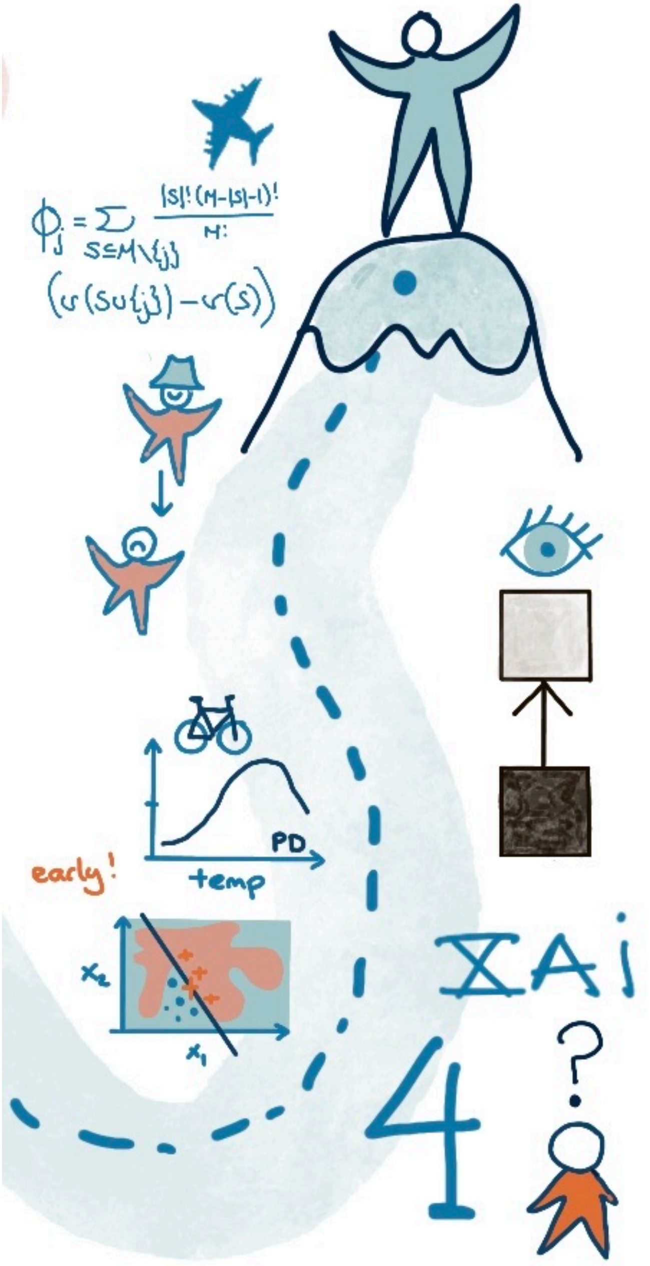
Example exam questions for Part 3

- What is the “Wisdom of the crowd”, and how have we used that in the course?
- Explain:

$$\hat{l}(y, \hat{y} + f_k(x)) = l(y, \hat{y}) + g(y, \hat{y}) f_k(x) + \frac{1}{2} h(y, \hat{y}) f_k^2(x)$$

- What are hyperparameters and how can these be “chosen”?
- What can we use Blaker´s confidence interval for? Why is this a good choice for CI?
- To test the hypothesis that method A is better than method B I want to compare the misclassification counts of the two methods on the same test set. How do I do that?
- What is the role of the ROC-AUC in statistical learning?

Part 4: Explainable AI


$\phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(M-|S|-1)!}{M!} (v(S \cup \{j\}) - v(S))$
PD
temp
early!
x_2
x_1
XAI
4
?

Topics

- How interpretable are the methods we have worked with?
- Global vs local, specific vs agnostic
- Linear model: Shapley regression
- Treebased methods: Importance plots
- ICE to PDP-plots, enhanced to ALE-plots.
- LIME. Role of ficticious data and local approximations.
- Counterfactuals. Which features to alter to obtain a different decision? Changes: few, small and likely. Three methods.

- Shapley values: understand and calculate by hand for 3-4 players.
- Properties: Efficiency, symmetry, dummy and linearity properties.
- Approximation methods for Shapley regression.
- Shaply for prediction: contribution function to be estimated via conditional distributions.
- Theoretical result for linear regression of independent covariates.
- Two challenges: computational complexity for sum over possible models and estimating the contribution function. Acknowledge what are challenge, but not go into details on solutions.

Prepared exam question 5) The Shapley values

The Shapley valuse are relevant in explainable AI.

a) What is the philosophy behind the Shapley value?

b) The Shapley regression is a global method (also referred to as the LMG-method). How does it relate to the Shapley value?

c) What are challenges if you want to use Shapley values for prediction explanation for a black box model?

Example exam questions for Part 4

- What are the interpretable methods we have learned about in this course and how interpretable are they (really)?
- What are local and global XAI methods and what are differences in usage of these?
- Explain “this plot”. And then a plot from this part.
- Explain the concept behind the PDP-plot by using the ICE-plot. Why is the ALE-plot needed?
- Explain one usage of a counterfactual explanation.

Oral exam schedule

May 10, 15, 22 - in room 822, 8th floor on Sentralbygg 2.
<s.ntnu.no/MA8701V2023examsignup>

- (00-03) Welcome and set-up
- (03-13) Student present "prepared question (with notes, no slides, write by hand on chalk board)" (no interruption).
- Put away notes.
- (13-20): If needed, follow-up question(s), else (preferably) either
  - if more theory needed/wanted: "explain formula/equation" or
  - if understanding needed/wanted: "explain figure/concept"
- (20-27): If needed/wanted: randomly select one (or more) short question from a topic that the student has not touched upon.
- (27-30): "Closing" with a more personal question on take-home messages from the course

Students get email in the end of the day (after all students that day), with pass/fail grade. Pass limit is B (around 70%).

How to prepare for the oral exam?

Group discussion:

- How do you plan to prepare?
- Will you arrange to meet and discuss difficult topics with fellow students?
- Do you want a few time slots where you may come to the lecturers office for questions, or is it ok to just contact lecturer when need?
- We have not used our https://mattelab2023v.math.ntnu.no/c/ma8701/107 - but it is available if you want!

Course feedback

Some observations about the course

Agree?

- Mainly a frequentist course, but some of the concepts and methods have a Bayesian version that might give insight into why and how the methods work, then Bayesian methods will be used.
- Focus is on regression and classification, and unsupervised learning is not planned to be part of the course.
- The required previous knowledge is listed because this is a PhD-course designed for statistics students. The background make the students go past an overview level of understanding of the course parts (move from algorithmic to deep understanding).

“Required” previous knowledge

Many of you did not have this required knowledge - how important was that?

- TMA4267 Linear statistical methods
- TMA4268 Statistical learning
- TMA4295 Statistical inference
- TMA4300 Computer intensive statistical methods
- TMA4315 Generalized linear models
- Good understanding and experience with R, or with Python, for statistical data analysis.
- Knowledge of markdown for writing reports and presentations (Rmarkdown/Quarto, Jupyther).
- Skills in group work - possibly using git or other collaborative tools.

Give feedback to the course evaluation

1) Final meeting with the reference group is on Thursday May 25 at 10.15: Contact Philip, Didrik or Jacob to give feedback - or/and
2) answer the IE studentevaluation (for all IMF courses)