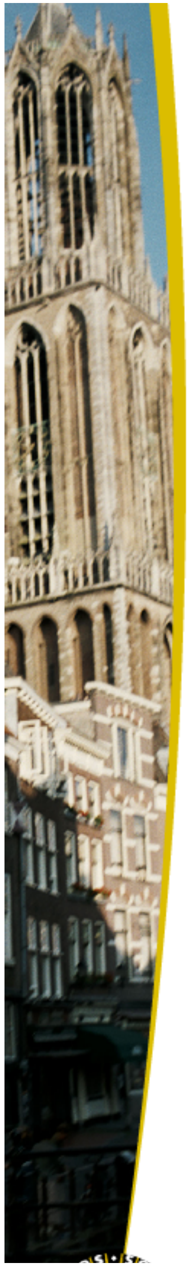
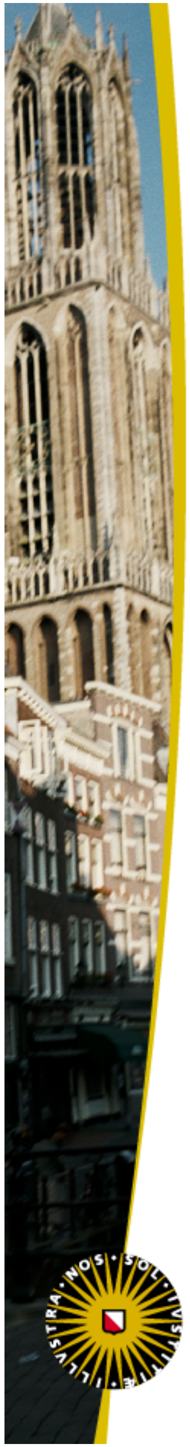


Resampling & Stability selection

Lützen Portengen, PhD
IRAS, Environmental Epidemiology Division
Utrecht University, the Netherlands

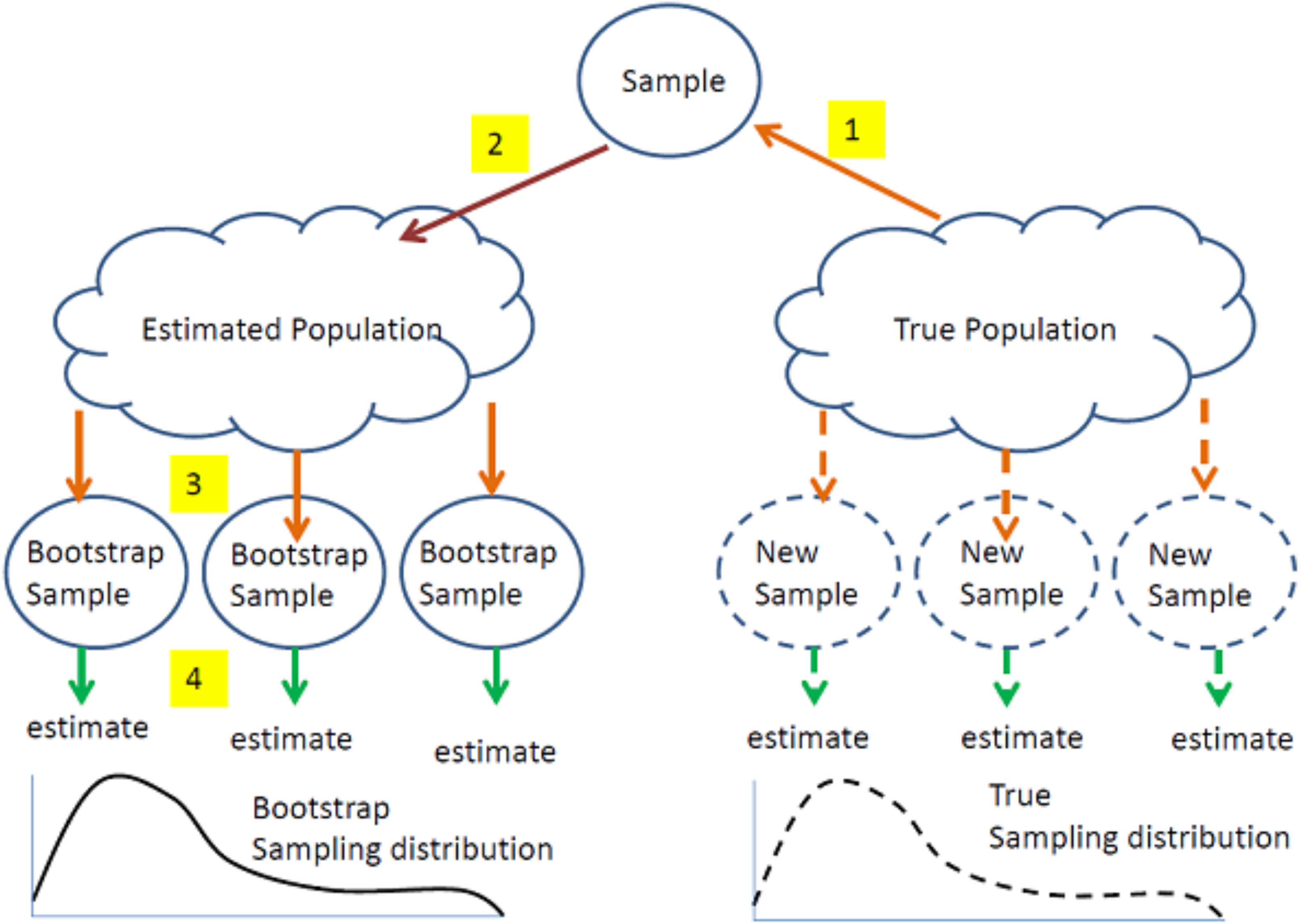
Universiteit Utrecht
INSTITUTE for RISK ASSESSMENT SCIENCES
ENVIRONMENTAL EPIDEMIOLOGY & VETERINARY PUBLIC HEALTH



Roadmap

- What is resampling?
- Approaches to resampling:
  - Subsampling
  - Bootstrapping
- Uses of resampling:
  - Model validation: how good is the model?
    - parameter tuning/model selection
  - Model uncertainty: parameters, structure
- Application to exposome-wide analyses
  - Stability selection

What is resampling?



Resampling methods

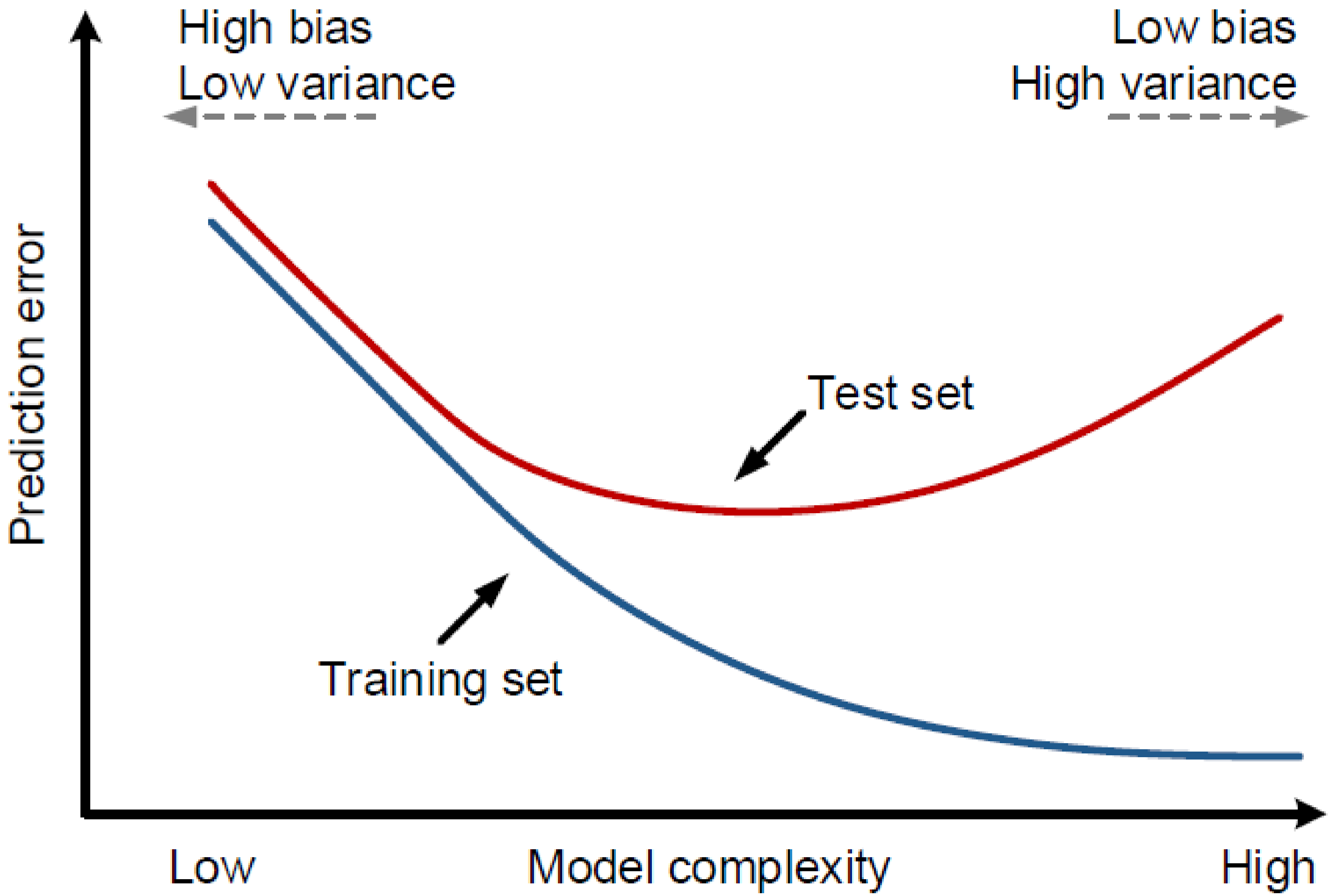
Subsampling

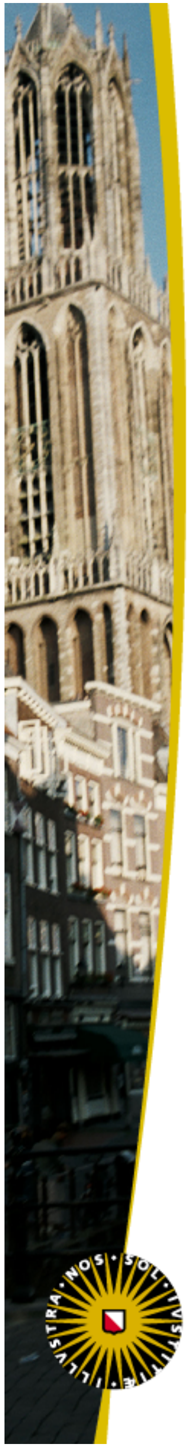
- Sample N* (<N) observations without replacement $\rightarrow$ new (sub)sample could have come from the original distribution. Mostly used to investigate model validity.

Bootstrapping

- Sample N* (=N) observations with replacement $\rightarrow$ new (sub)sample could have been a sample from (a discrete version of) the original distribution. Mostly used to investigate model uncertainty.

Use: prediction error





Cross-validation

K-fold cross-validation for est. prediction error:

- Repeated, systematic subsampling: divide sample in K non-overlapping folds
- K-1 folds as training sets to estimate model parameters
- other fold as test set to assess prediction error
- repeat K times and average prediction error

LOOCV → K = N

Optimal choice for #folds?

Expected test prediction error:

$$\frac{1}{N-N_*}\sum_{i=1}^{N-N_*}(Y_i - M(X_i)) = \frac{1}{N_*}\sum_{i=1}^{N_*}(Y_i - M(X_i)) + \frac{2}{N_*}\sum_{i=1}^{N_*} cov(Y_i, M(X_i))$$

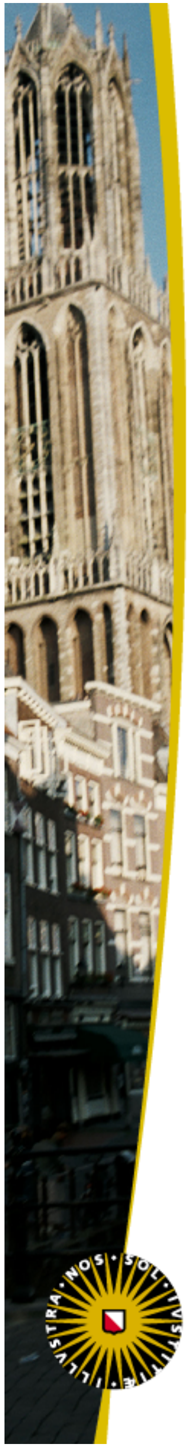
Equivalent to AIC (if error distribution correct).

Universiteit Utrecht

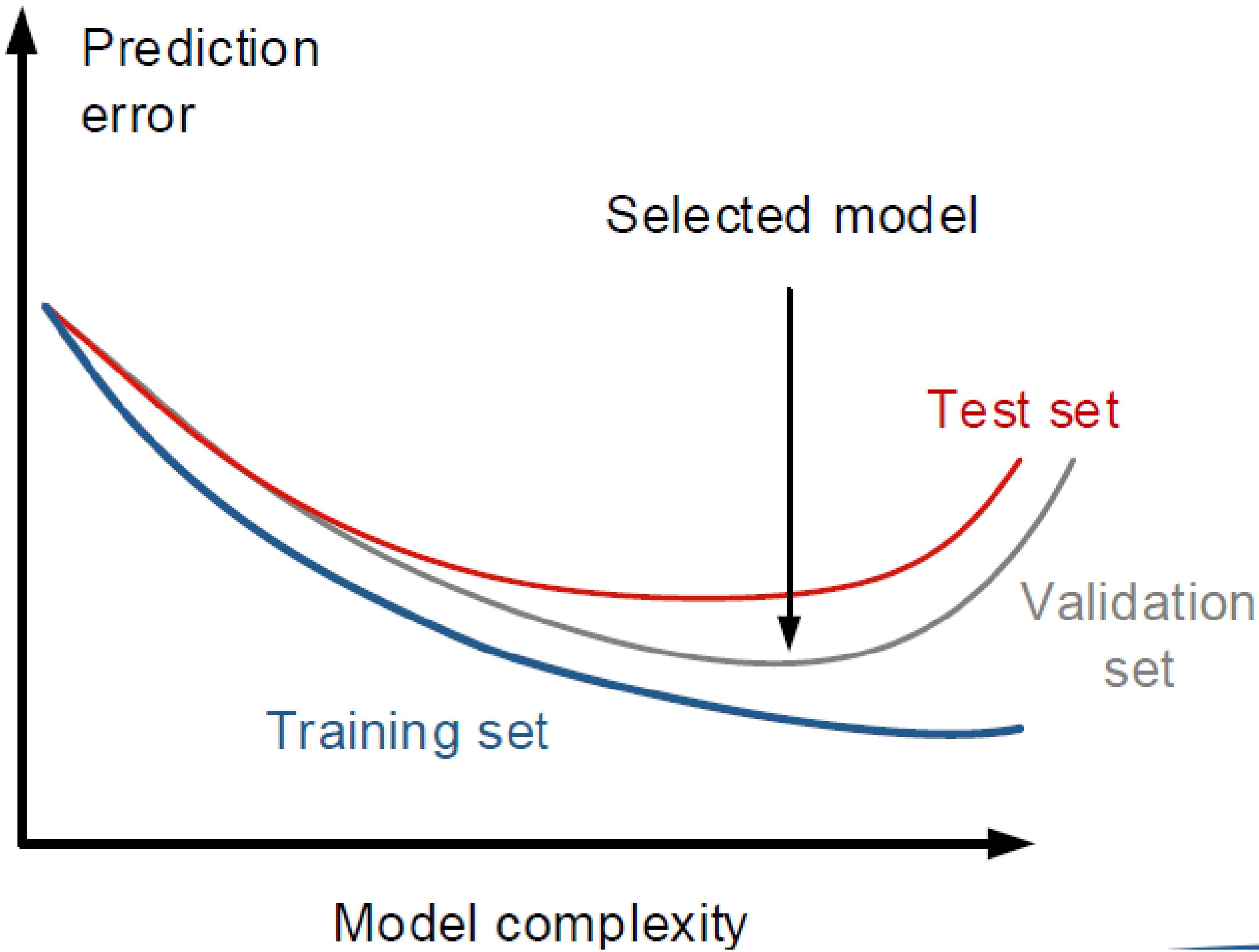
Stratified sampling

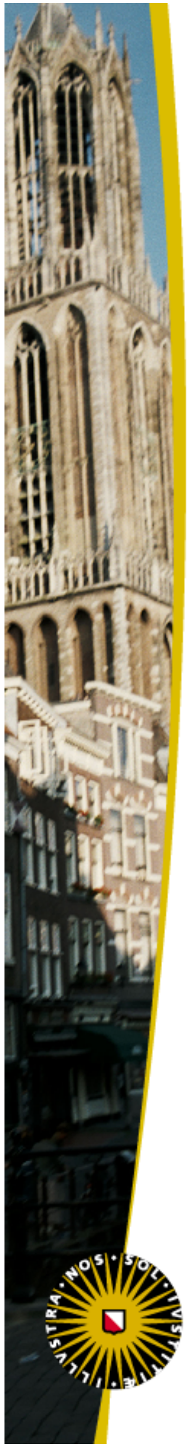
Stratified sampling

- unbalanced response classes
- unbalanced distribution of predictors
- clustered data, time series
- beware of twinning (presence of near identical cases in training and test sets)



Use: parameter tuning

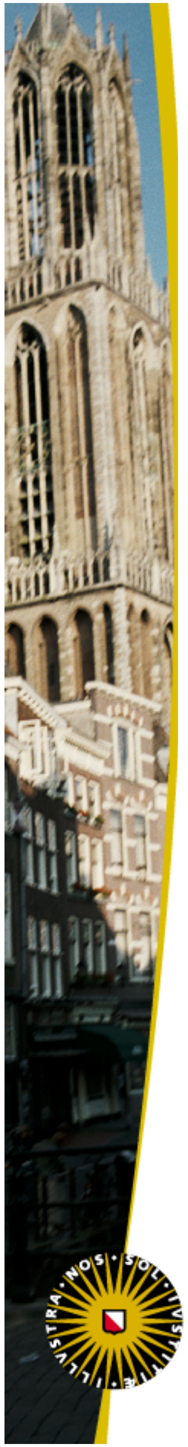


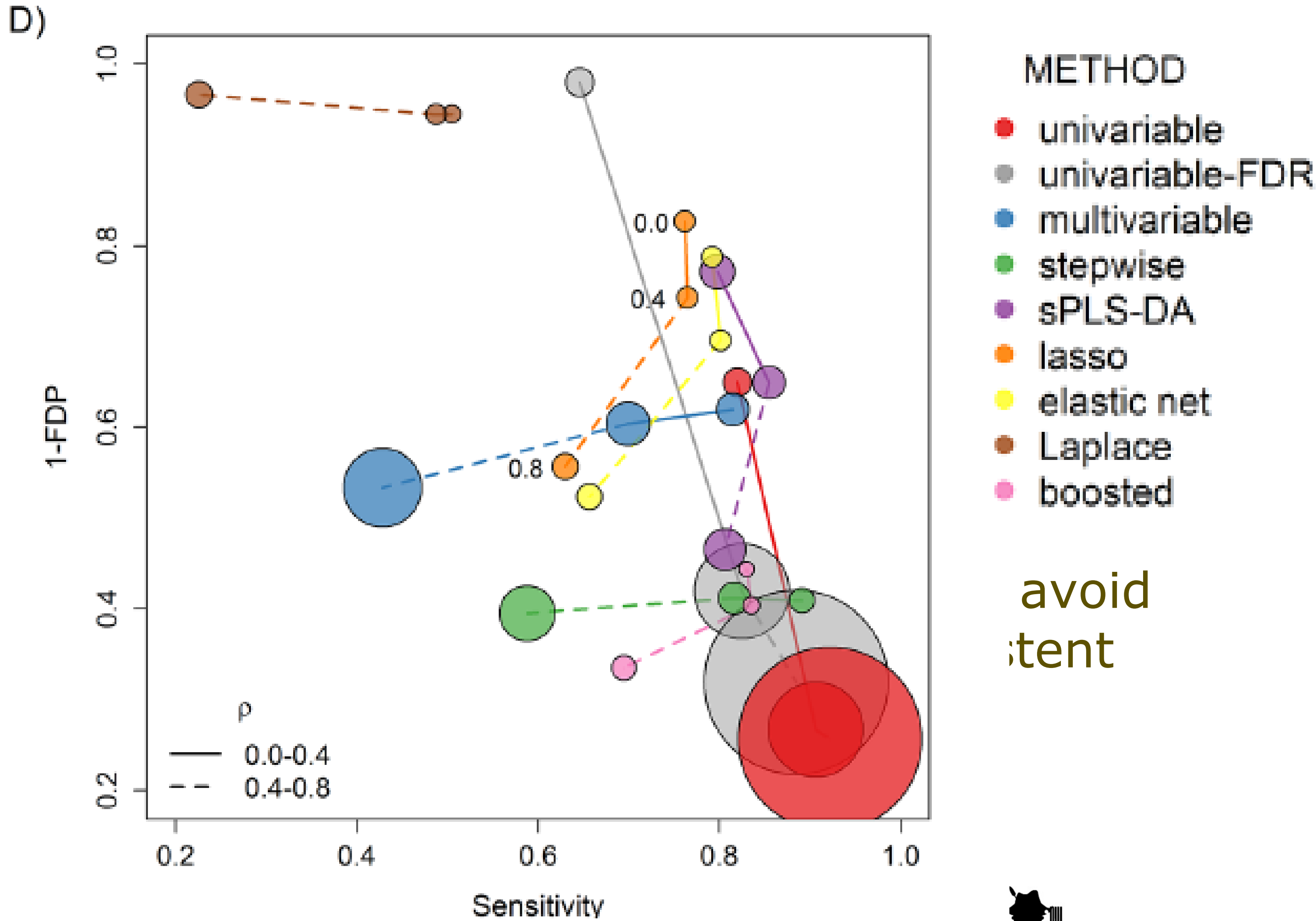


Use: model uncertainty

Model uncertainty:

- estimate variability of parameters (e.g. a slope coefficient) that results from sampling
- bootstrapping often used when analytical approaches (in the form of standard errors) are not directly available (e.g. SE for ratio of parameters)
- estimate variability (instability) of model structure (variable selection models)
- identify influential observations

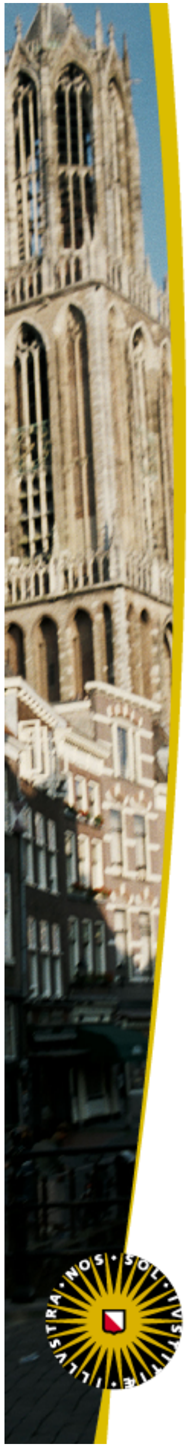




avoid
tent

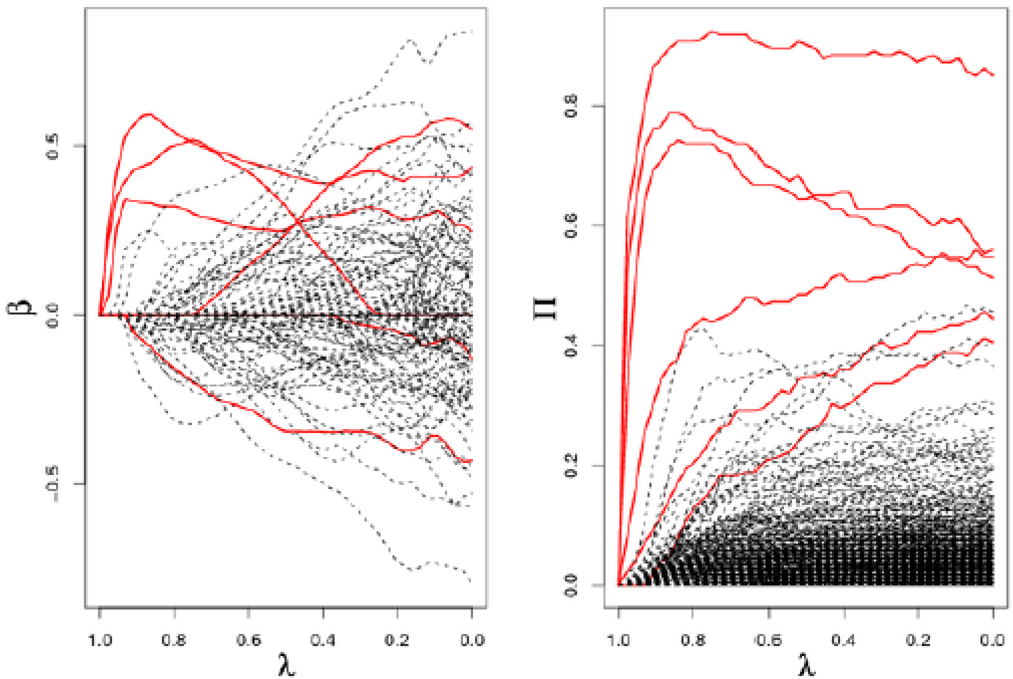
Universiteit Utrecht

StableBoy SELECTIONS



Stability selection 2/6

- setup grid of values for tuning parameter that affects selection of features (i.e. penalties for lasso, pvalues for stepwise selection)
- repeated subsampling (1:1): estimate P($\beta_i \neq$ 0) across grid



Universiteit Utrecht

Stability selection

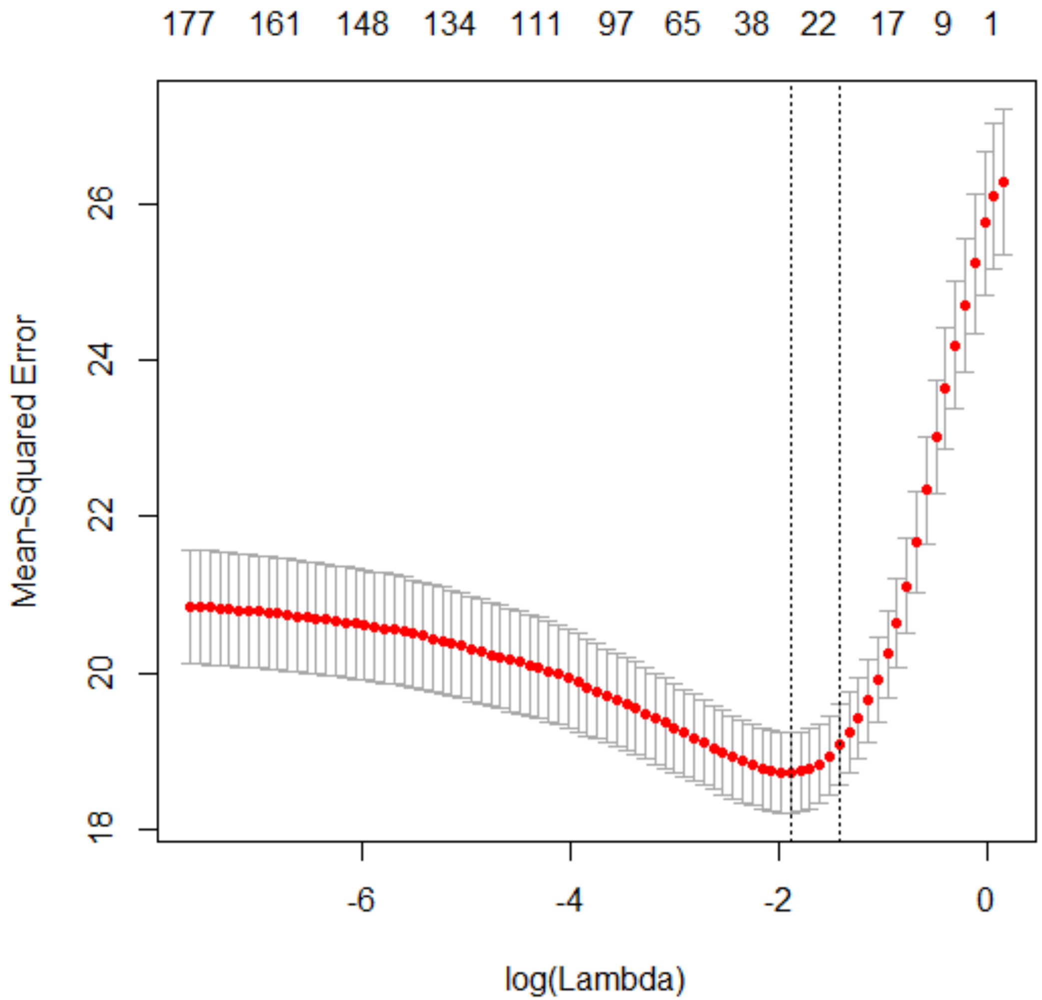
Meinshausen (2010) provides an upper bound on the number of falsely selected variables (V):

$$E(V) \le \frac{1}{2\pi_{thr}-1}\frac{q_{\Lambda}^2}{p}$$

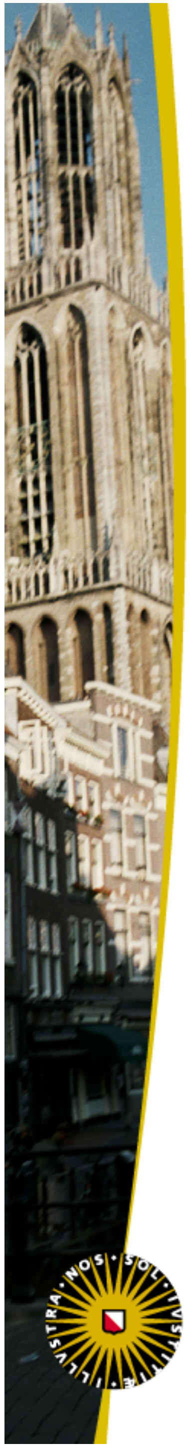
for a chosen cutoff (ϖ_{thr}) and *E(V)* we can choose the reqularisation region Λ so that the maximum number of non-zero coefficients q_{Λ} equals the calculated value. We then select all variables that have max($P\lambda(\beta i \neq 0)$) > ϖ_{thr}

Stability selection example 4/6

- Simulated X r
  covariance ma
  measurement
  biomarkers)
- Outcome Y si
  chosen expos
- CV of lasso re



Universiteit Utrecht

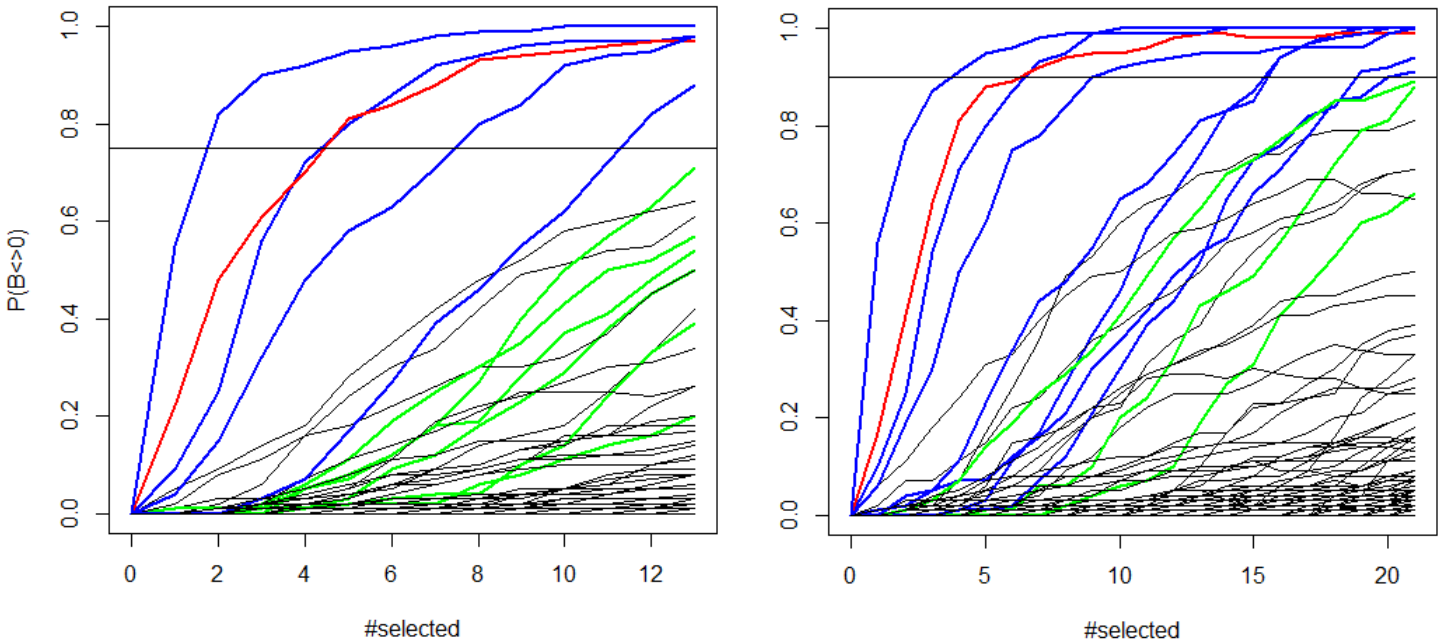


Stability selection example 5/6

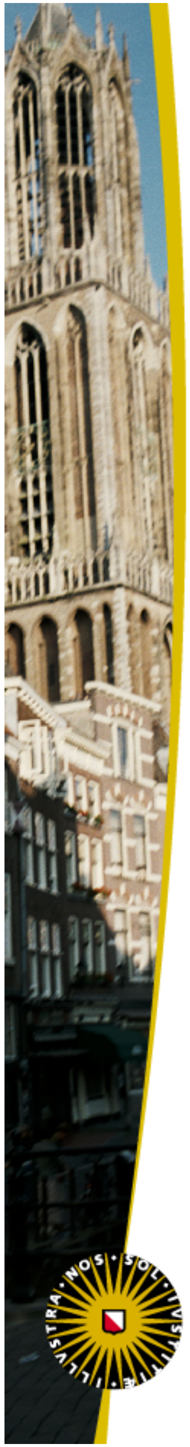
- Stability selection implemented in R packages "hdi" (Meinshausen) and "stabs" (Shah). Need to indicate X and Y, a variable selection

hr

tq)



Universiteit Utrecht



Stability selection example 6/6

- Stability selection does not result in a model (we can obviously fit one with the selected variables only)
- Subsampling needs to be 1:1 for the upper error bound (used here) to hold
- The theoretical conditions for the bound to hold exactly are rather restrictive.
- The bound has been found to be conservative in some practical situations and alternatives have been proposed (e.g. Shah et al. 2013).


Thank you!