Feature Selection via Detecting Ineffective Features

Kris De Brabanter[1,2] **László Györfi**[3]

Dep. of Electrical Engineering, KU Leuven, Belgium

[2]Dep. of Statistics & Computer Science, Iowa State University, Ames, USA

[3]Dep. Computer Science & Information Theory, Budapest University of Technology and Economics, Hungary

July, 8 2013

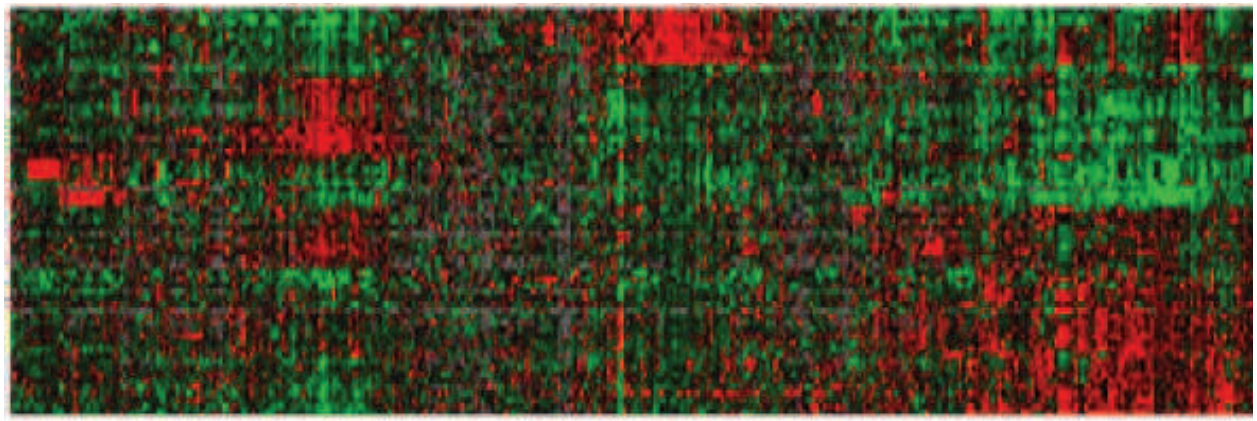
ROKS 2013

Outline

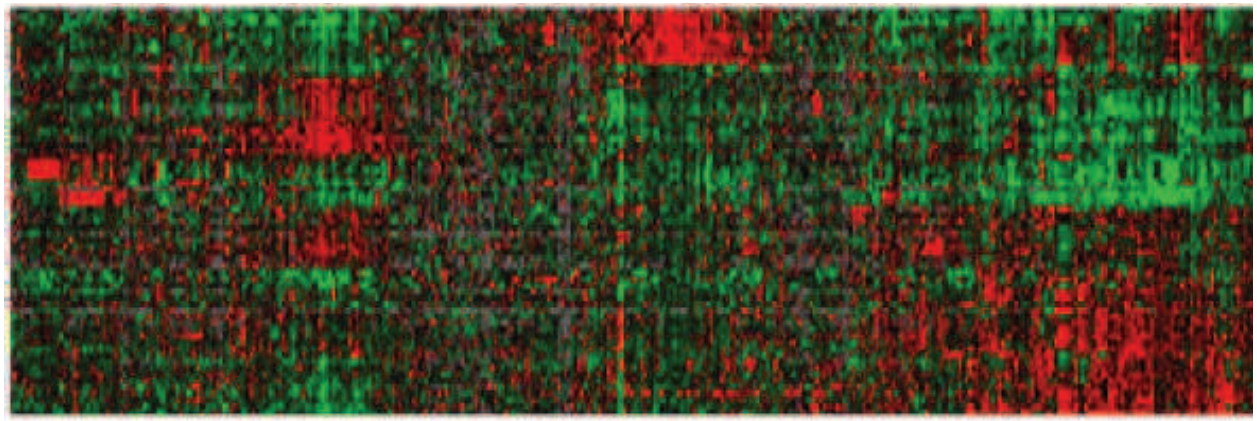
Outline

Problem description

Application & methods

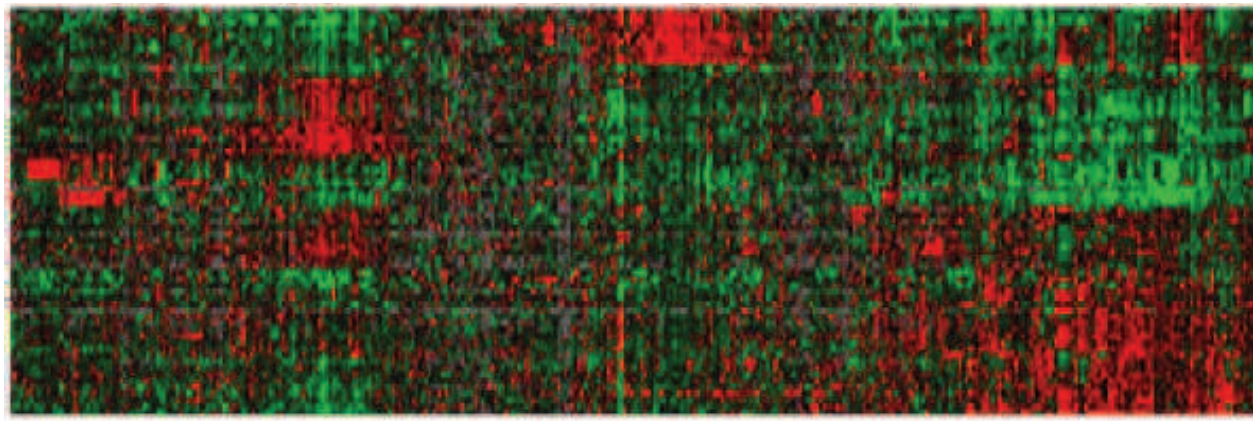


Application & methods



Find important genes

Application & methods



Find important genes

- LASSO (Tibshirani, 1996)
- Prediction Analysis for Microarrays (Tibshirani *et al.*, 2002)
- t-test (Chun & Wen, 2010)
- Variable selection via additive models (Antoniadis *et al.*, 2012)
- LARS (Efron *et al.*, 2004)
- ...

Application & methods (cont'd)

- Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$
- samples are i.i.d
- Model: $Y_i = \sum_{j=1}^{d} \beta_j X_i^{(j)} + \varepsilon_i \quad 1, \ldots, n$
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. and independent of X_i
- $\mathbf{E}[\varepsilon_i \mid X_i] = 0$

LASSO

When $d > n$, standard LS estimator is not unique and will overfit data. Therefore, some kind of regularization is needed.

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1$$

Some asymptotic theory for LASSO: $Y = X\beta + \varepsilon$ is correct

$$\boxed{\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1}$$

Some asymptotic theory for LASSO: $Y = X\beta + \varepsilon$ is correct

$$\boxed{\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1}$$

1. Consistency prediction (Greenshtein & Ritov, 2004; Bartlett *et al.*, 2009): If
   - $\lambda = \lambda_n = O\big(\sqrt{\log(d)/n}\big)$
   - $\|\beta\|_1 = o\big((n/\log n)^{1/4}\big)$
   - then $(\hat{\beta} - \beta)^T \Sigma_X (\hat{\beta} - \beta) = o_P(1)$ as $n \to \infty$

Some asymptotic theory for LASSO: $Y = X\beta + \varepsilon$ is correct

$$\boxed{\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1}$$

1. Consistency prediction (Greenshtein & Ritov, 2004; Bartlett *et al.*, 2009): If
   - $\lambda = \lambda_n = O\big(\sqrt{\log(d)/n}\big)$
   - $\|\beta\|_1 = o\big((n/\log n)^{1/4}\big)$
   - then $(\hat{\beta} - \beta)^T \Sigma_X (\hat{\beta} - \beta) = o_P(1)$ as $n \to \infty$
2. Consistency variable selection (Meinshausen & Bühlmann, 2006)
   - Let $\hat{S}(\lambda) = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, d\}$
   - if $\inf_{j \in S} |\beta_j| \gg \sqrt{(\mathrm{card}(S) \log d)/n}$ and $\lambda = \lambda_n \gg O\big(\sqrt{\log(d)/n}\big)$
   - Design matrix $\mathbf{X}$ is not *ill posed* and no strong linear dependence within sub-matrices
   - then $\mathbf{P}[\hat{S}(\lambda) = S] \to 1$ as $d \geq n \to \infty$

Some asymptotic theory for LASSO: $Y = X\beta + \varepsilon$ is correct

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1$$

1. Consistency prediction (Greenshtein & Ritov, 2004; Bartlett *et al.*, 2009): If
   - $\lambda = \lambda_n = O\big(\sqrt{\log(d)/n}\big)$
   - $\|\beta\|_1 = o\big((n/\log n)^{1/4}\big)$
   -

λ is larger for variable selection than prediction

2. Cons
   - Let $\hat{S}(\lambda) = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, d\}$
   - if $\inf_{j \in S} |\beta_j| \gg \sqrt{(\mathrm{card}(S)\log d)/n}$ and $\lambda = \lambda_n \gg O\big(\sqrt{\log(d)/n}\big)$
   - Design matrix $\mathbf{X}$ is not *ill posed* and no strong linear dependence within sub-matrices
   - then $\mathbf{P}[\hat{S}(\lambda) = S] \to 1$ as $d \geq n \to \infty$

Parameter selection for variable selection

- BIC, AIC, CV $\rightarrow$ no theoretical justification for variable selection
- CV: LASSO often selects too many variables
- λ close to zero $\rightarrow$ LARS until end of regularization path
- Other ways: stability selection, hypothesis testing

Parameter selection for variable selection

- BIC, AIC, CV $\rightarrow$ no theoretical justification for variable selection
- CV: LASSO often selects too many variables
- λ close to zero $\rightarrow$ LARS until end of regularization path
- Other ways: stability selection, hypothesis testing
- What can be theoretically supported?

Parameter selection for variable selection

- BIC, AIC, CV $\rightarrow$ no theoretical justification for variable selection
- CV: LASSO often selects too many variables
- λ close to zero $\rightarrow$ LARS until end of regularization path
- Other ways: stability selection, hypothesis testing
- What can be theoretically supported?

Substantial relevant covariates (Bühlmann & van de Geer, 2011)

Let

$$S^{\text{relevant}(C)} = \{j : |\beta_j| \geq C, j = 1, \ldots, d\},$$

then for any fixed $0 < C < \infty$

$$\mathbf{P}[\hat{S}(\lambda) \supset S^{\text{relevant}(C)}] \rightarrow 1, \quad n \rightarrow \infty.$$

For $C_n > O(\text{card}(S)\sqrt{\log(d_n)/n})$ and $\lambda_n = O\big(\sqrt{\log(d_n)/n}\big)$: with high probability $\hat{S}(\lambda_n) \supset S^{\text{relevant}(C)} \rightarrow \hat{S}(\lambda_{CV}) \supset S^{\text{relevant}(C)}$

Summary LASSO: Pros vs. Cons

- Easy to implement
- Fast algorithms available (Matlab/R)
- Results are easy to interpret
- Solid theoretical foundations (if model is correct)

Summary LASSO: Pros vs. Cons

- Easy to implement
- Fast algorithms available (Matlab/R)
- Results are easy to interpret
- Solid theoretical foundations (if model is correct)

- Model based
- Model restrictions are not necessarily met
- No theoretical results if model does not hold

Outline

Theoretical aspects

- Let $Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$
- $L^\star := \mathbf{E}\{(Y - m(\mathbf{X}))^2\} = \min_f \mathbf{E}\{(Y - f(\mathbf{X}))^2\}$ be minimum MSE
- $L^\star = \mathbf{E}\{Y^2\} - \mathbf{E}\{m(\mathbf{X})^2\}$

Theoretical aspects

- Let $Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$
- $L^\star := \mathbf{E}\{(Y - m(\mathbf{X}))^2\} = \min_f \mathbf{E}\{(Y - f(\mathbf{X}))^2\}$ be minimum MSE
- $L^\star = \mathbf{E}\{Y^2\} - \mathbf{E}\{m(\mathbf{X})^2\}$

$$\Downarrow$$

$$\boxed{L_n^\star := \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{i,1}}$$

- $Y_{i,1}$ is value of first nearest neighbor of $\mathbf{X}_i$ from $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$

Theoretical aspects (cont'd)

Theorem (Almost sure convergence & universal consistency)

Assume bounded Y*, then*

$$L_n^\star \to L^\star \quad a.s.$$

Moreover, if $\mathbf{E}\{Y^2\} < \infty$

$$\frac{1}{n}\sum_{i=1}^{n} L_i^\star \to L^\star \quad a.s.$$

Theoretical aspects (cont'd)

Theorem (Almost sure convergence & universal consistency)

Assume bounded Y, then

$$L_n^\star \to L^\star \quad a.s.$$

Moreover, if $\mathbf{E}\{Y^2\} < \infty$

$$\frac{1}{n}\sum_{i=1}^{n} L_i^\star \to L^\star \quad a.s.$$

Theorem (Rate of convergence & consistency)

Assume X and Y bounded and the regression function m is Lipschitz continuous. Further, assume that ties occur with probability zero. Then, for $d \geq 2$

$$\mathbf{E}\{|L_n^\star - L^\star|\} \leq c_1 n^{-1/2} + c_2 n^{-2/d}.$$

Further, if $n \to \infty$ faster than d

$$\mathbf{E}\{|L_n^\star - L^\star|\} = o_p(1).$$

Detection of inefficient features

- Denote $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d)$

Detection of inefficient features

- Denote $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d)$
- Corresponding minimum MSE

$$L^{\star(-k)} := \mathbf{E}\left\{Y - \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\}\right\}^2$$

Detection of inefficient features

- Denote $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d)$
- Corresponding minimum MSE

$$L^{\star(-k)} := \mathbf{E}\left\{Y - \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\}\right\}^2$$

- Test (null) hypothesis:

$$\mathcal{H}_k : L^{\star(-k)} = L^\star$$

Leaving out k-th component does not increase minimum MSE

Detection of inefficient features

- Denote $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d)$
- Corresponding minimum MSE

$$L^{\star(-k)} := \mathbf{E}\left\{Y - \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\}\right\}^2$$

- Test (null) hypothesis:

$$\mathcal{H}_k : L^{\star(-k)} = L^\star$$

Leaving out k-th component does not increase minimum MSE

- Under $\mathcal{H}_k$:

$$m(\mathbf{X}) := \mathbf{E}\{Y \mid \mathbf{X}\} = \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\} =: m^{(-k)}(\mathbf{X}^{(-k)}) \quad a.s.$$

Detection of inefficient features (cont'd)

- Data without k-th component:

$$D_n^{(-k)} = \{(\mathbf{X}_1^{(-k)}, Y_1), \ldots, (\mathbf{X}_n^{(-k)}, Y_n)\}$$

Detection of inefficient features (cont'd)

- Data without k-th component:

$$D_n^{(-k)} = \{(\mathbf{X}_1^{(-k)}, Y_1), \ldots, (\mathbf{X}_n^{(-k)}, Y_n)\}$$

- Estimate $L^{\star(-k)}$ by

$$L_n^{\star(-k)} := \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{i,1}^{(-k)}$$

Detection of inefficient features (cont'd)

- Data without k-th component:

$$D_n^{(-k)} = \{(\mathbf{X}_1^{(-k)}, Y_1), \ldots, (\mathbf{X}_n^{(-k)}, Y_n)\}$$

- Estimate $L^{\star(-k)}$ by

$$L_n^{\star(-k)} := \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{i,1}^{(-k)}$$

- Test statistic:

$$L_n^{(-k)} - L_n = \frac{1}{n}\sum_{i=1}^{n} Y_i (Y_{i,1} - Y_{i,1}^{(-k)})$$

Detection of inefficient features (cont'd)

- Data without k-th component:

$$D_n^{(-k)} = \{(\mathbf{X}_1^{(-k)}, Y_1), \ldots, (\mathbf{X}_n^{(-k)}, Y_n)\}$$

- Estimate $L^{\star(-k)}$ by

$$L_n^{\star(-k)} := \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{i,1}^{(-k)}$$

- Test statistic:

$$L_n^{(-k)} - L_n = \frac{1}{n}\sum_{i=1}^{n} Y_i (Y_{i,1} - Y_{i,1}^{(-k)})$$

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*

Issue with the test statistic

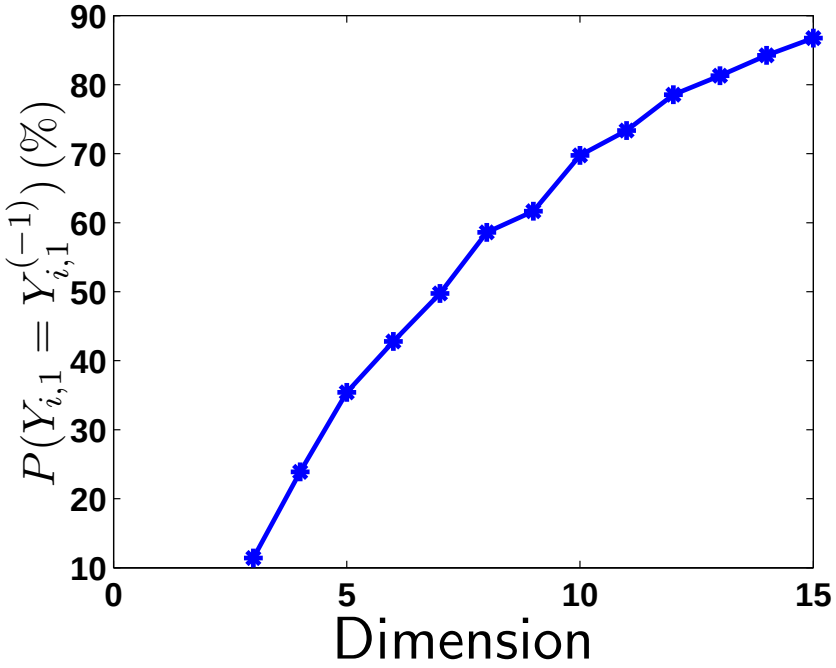
- $\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_{i,1} - Y_{i,1}^{(-k)})$ is small even when $\mathcal{H}_k$ is not true

Issue with the test statistic

- $\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_{i,1} - Y_{i,1}^{(-k)})$ is small even when $\mathcal{H}_k$ is not true
- Large probability that first N.N. of $\mathbf{X}_i$ and $\mathbf{X}_i^{(-k)}$ are the same

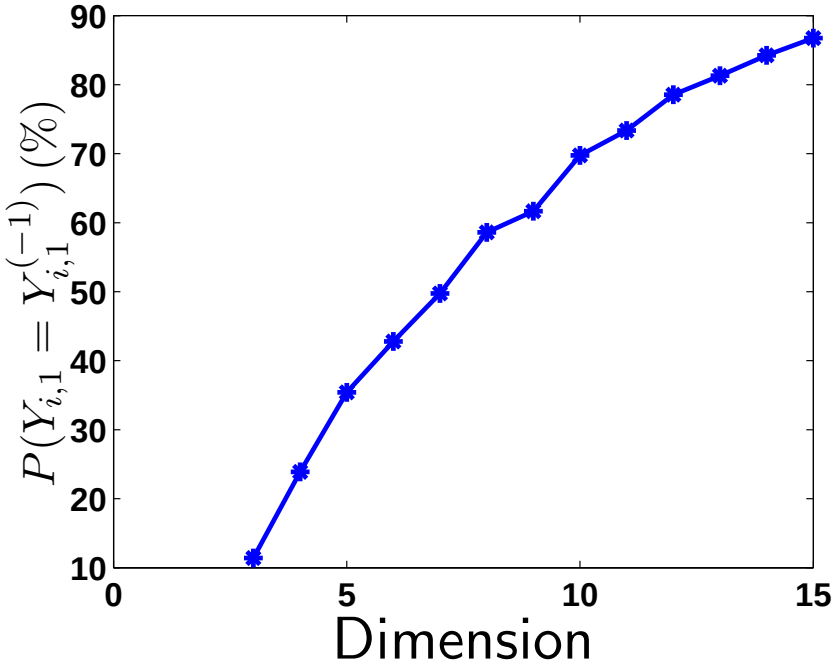
Issue with the test statistic

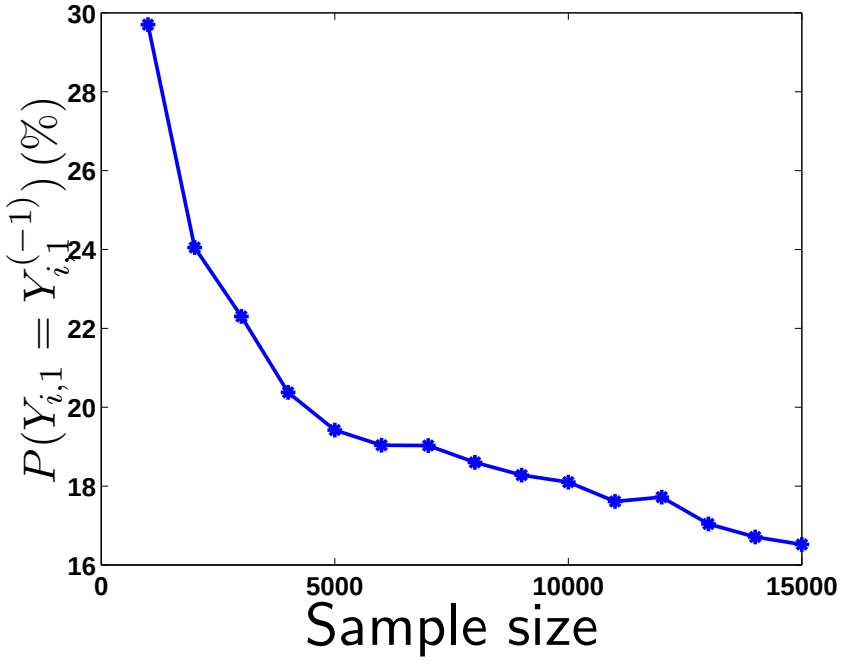
- $\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_{i,1} - Y_{i,1}^{(-k)})$ is small even when $\mathcal{H}_k$ is not true
- Large probability that first N.N. of $\mathbf{X}_i$ and $\mathbf{X}_i^{(-k)}$ are the same



Issue with the test statistic

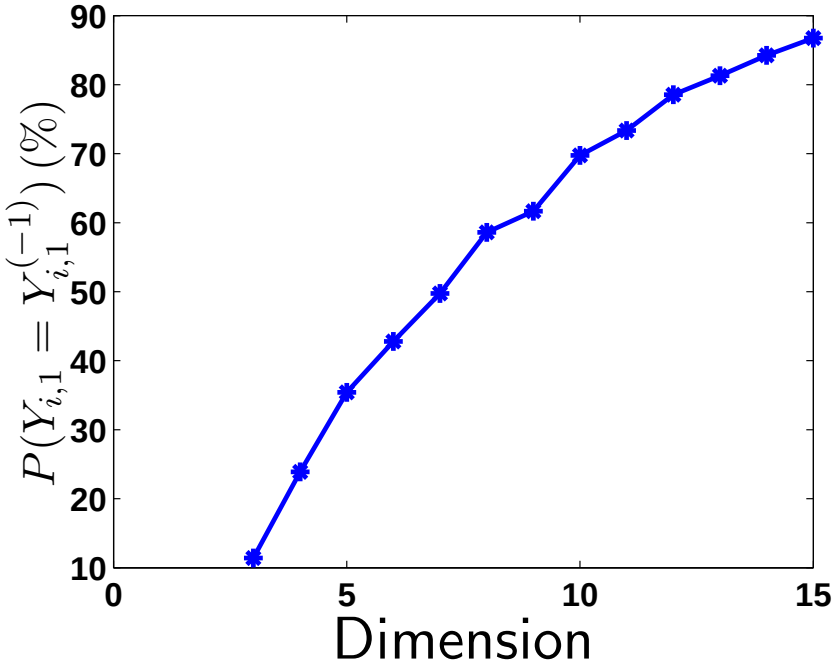
- $\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_{i,1} - Y_{i,1}^{(-k)})$ is small even when $\mathcal{H}_k$ is not true
- Large probability that first N.N. of $\mathbf{X}_i$ and $\mathbf{X}_i^{(-k)}$ are the same

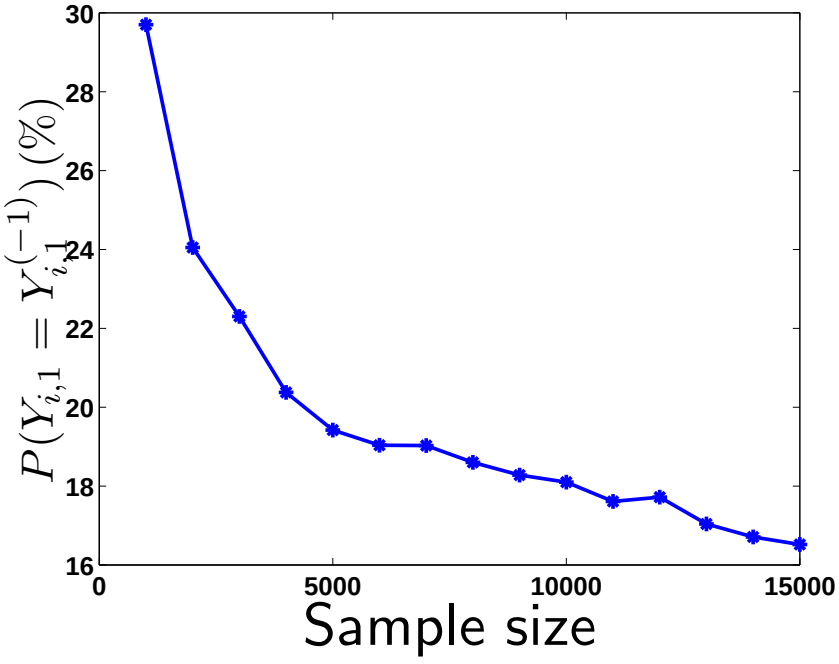




Issue with the test statistic

- $\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_{i,1} - Y_{i,1}^{(-k)})$ is small even when $\mathcal{H}_k$ is not true
- Large probability that first N.N. of $\mathbf{X}_i$ and $\mathbf{X}_i^{(-k)}$ are the same





Modification of test statistic needed

Modification of test statistic

- Modify test statistic such that

$$(\hat{Y}_{i,1}, \hat{Y}_{i,1}^{(-k)}) = \begin{cases} (Y_{i,1}, Y_{i,1}^{(-k)}) & \text{if} \quad Y_{i,1} \neq Y_{i,1}^{(-k)} \\ \\ I_i(Y_{i,2}, Y_{i,1}^{(-k)}) + (1 - I_i)(Y_{i,1}, Y_{i,2}^{(-k)}) & \text{otherwise,} \end{cases}$$

with

$$I_i = \begin{cases} 0 & \text{with probability } 1/2, \\ 1 & \text{with probability } 1/2, \end{cases}$$

Limit distribution of the test statistic

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*

Limit distribution of the test statistic

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*
- But what is small? $L_n^{(-k)} - L_n$ is r.v.

Limit distribution of the test statistic

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*
- But what is small? $L_n^{(-k)} - L_n$ is r.v.
- In general, use CLT $\rightarrow$ not possible here

Limit distribution of the test statistic

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*
- But what is small? $L_n^{(-k)} - L_n$ is r.v.
- In general, use CLT $\rightarrow$ not possible here
- $L_n^{(-k)} - L_n$ is average of dependent r.v.

Limit distribution of the test statistic

- Accept $\mathcal{H}_k$ if $L_n^{(-k)} - L_n$ is *small*
- But what is small? $L_n^{(-k)} - L_n$ is r.v.
- In general, use CLT $\rightarrow$ not possible here
- $L_n^{(-k)} - L_n$ is average of dependent r.v.

Definition (Exchangeable random variables)

A triangular array $V_{n,i}, n = 1, \ldots,$ and $i = 1, \ldots, n$ is called (row-wise) exchangeable if for any fixed n and for any permutation $\rho(1), \ldots, \rho(n)$ of $1, \ldots, n$, the distributions of

$$(V_{n,1}, \ldots, V_{n,n})$$

and

$$(V_{n,\rho(1)}, \ldots, V_{n,\rho(n)})$$

are equal.

Limit distribution of the test statistic (cont'd)

Theorem (Blum *et al.* (1958), De Brabanter & Györfi (2012))

Let

$$V_{n,i} = \frac{Y_i(Y_{i,1} - Y_{i,1}^{(-k)}) - \mathbf{E}\{Y_i(Y_{i,1} - Y_{i,1}^{(-k)})\}}{\sqrt{\mathbf{Var}(Y_i(Y_{i,1} - Y_{i,1}^{(-k)}))}}.$$

Then for

- $\mathbf{E}\{V_{n,1}V_{n,2}\} = o(1/n)$
- $\lim_{n\to\infty} \mathbf{E}\{V_{n,1}^2 V_{n,2}^2\} = 1$
- $\mathbf{E}\{|V_{n,1}|^3\} = o(\sqrt{n})$,

it follows that

$$\sqrt{n}(L_n^{(-k)} - L_n) \xrightarrow{d} N(0, 2L^\star \, \mathbf{E}\{Y^2\})$$

under $\mathcal{H}_k$.

Limit distribution of the test statistic (cont'd)

Theorem (Blum *et al.* (1958), De Brabanter & Györfi (2012))

Let

$$V_{n,i} = \frac{Y_i(Y_{i,1} - Y_{i,1}^{(-k)}) - \mathbf{E}\{Y_i(Y_{i,1} - Y_{i,1}^{(-k)})\}}{\sqrt{\mathbf{Var}(Y_i(Y_{i,1} - Y_{i,1}^{(-k)}))}}.$$

Then for

- $\mathbf{E}\{V_{n,1}V_{n,2}\} = o(1/n)$
- $\lim_{n\to\infty} \mathbf{E}\{V_{n,1}^2 V_{n,2}^2\} = 1$
- $\mathbf{E}\{|V_{n,1}|^3\} = o(\sqrt{n})$,

it follows that

$$\sqrt{n}(L_n^{(-k)} - L_n) \xrightarrow{d} N(0, 2L^\star \, \mathbf{E}\{Y^2\})$$

under $\mathcal{H}_k$.

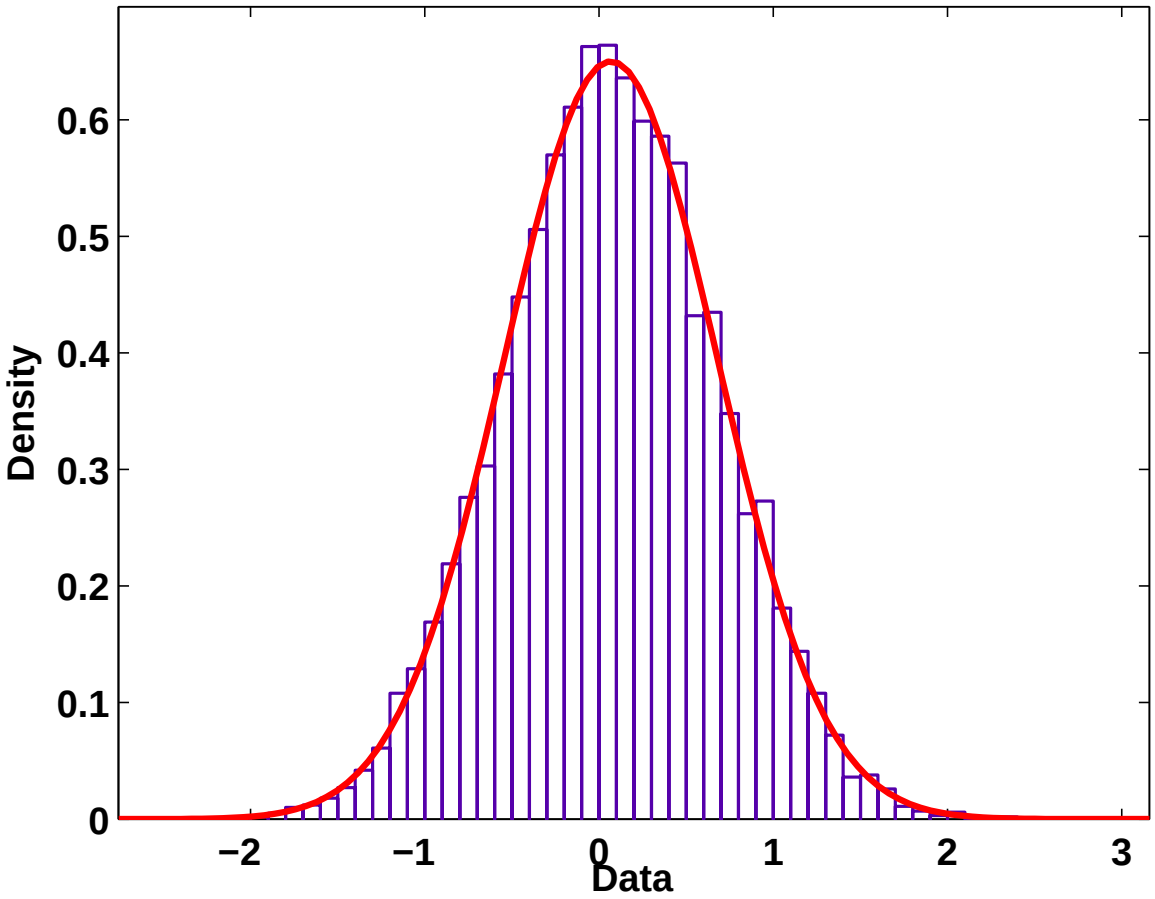
!Smoothness of regression function and dimension do not count under $\mathcal{H}_k$!

Empirical verification of the theorem

- $Y = \sum_{i=1}^{5} c_i X^{(i)} + \varepsilon$, $c_1 = 0$ and $c_i = 1$ for $i = 2, \ldots, 5$.
- $\mathbf{X}$ uniform on $[0,1]^5$ and $\varepsilon \sim N(0, 0.05^2)$.
- Bootstrap (10,000 replications)

Empirical verification of the theorem

- $Y = \sum_{i=1}^{5} c_i X^{(i)} + \varepsilon$, $c_1 = 0$ and $c_i = 1$ for $i = 2, \ldots, 5$.
- $\mathbf{X}$ uniform on $[0,1]^5$ and $\varepsilon \sim N(0, 0.05^2)$.
- Bootstrap (10,000 replications)



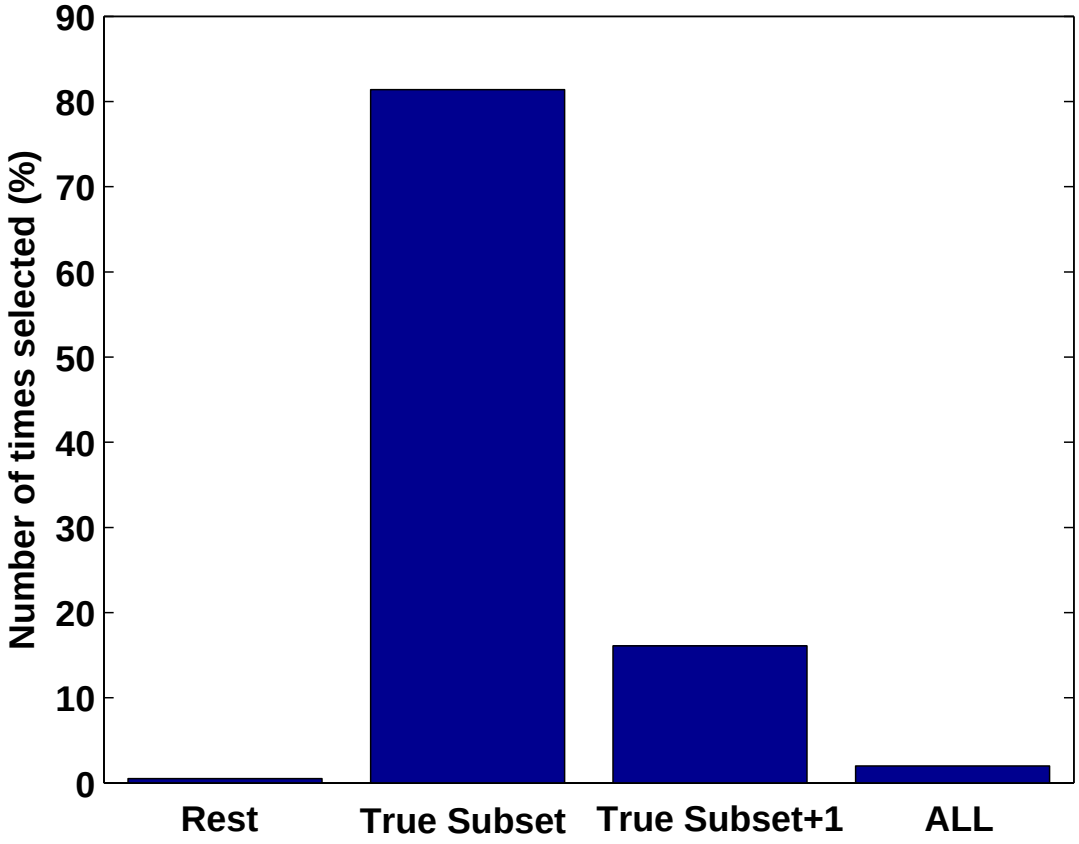
Outline

Toy example

- $Y = \sin(\pi X^{(1)}) \cos(\pi X^{(4)}) + \varepsilon$ with $\varepsilon \sim N(0, 0.1^2)$.
- $\mathbf{X}$ uniform on $[0, 1]^5$
- $n = 1000$
- 1000 repetitions, significance level $\alpha = 0.05$

Toy example

- $Y = \sin(\pi X^{(1)}) \cos(\pi X^{(4)}) + \varepsilon$ with $\varepsilon \sim N(0, 0.1^2)$.
- $\mathbf{X}$ uniform on $[0,1]^5$
- $n = 1000$
- 1000 repetitions, significance level $\alpha = 0.05$



Bosting housing data

- Data set: $n = 506$, $d = 13$
- Goal: predict median value of owner-occupied homes in USD

Bosting housing data

- Data set: $n = 506$, $d = 13$
- Goal: predict median value of owner-occupied homes in USD

Variable	description
1	per capita crime rate by town
2	proportion of residential land zoned for lots over $25,000$ sq.ft
3	proportion of non-retail business acres per town
4	Charles River dummy variable
5	nitric oxides concentration
6	average number of rooms per dwelling
7	proportion of owner-occupied units built prior to 1940
8	weighted distances to five Boston employment centres
9	index of accessibility to radial highways
10	full-value property-tax rate per USD
11	pupil-teacher ratio by town
12	proportion of African-Americans by town
13	percentage of lower status of the population

Boston housing data (cont'd)

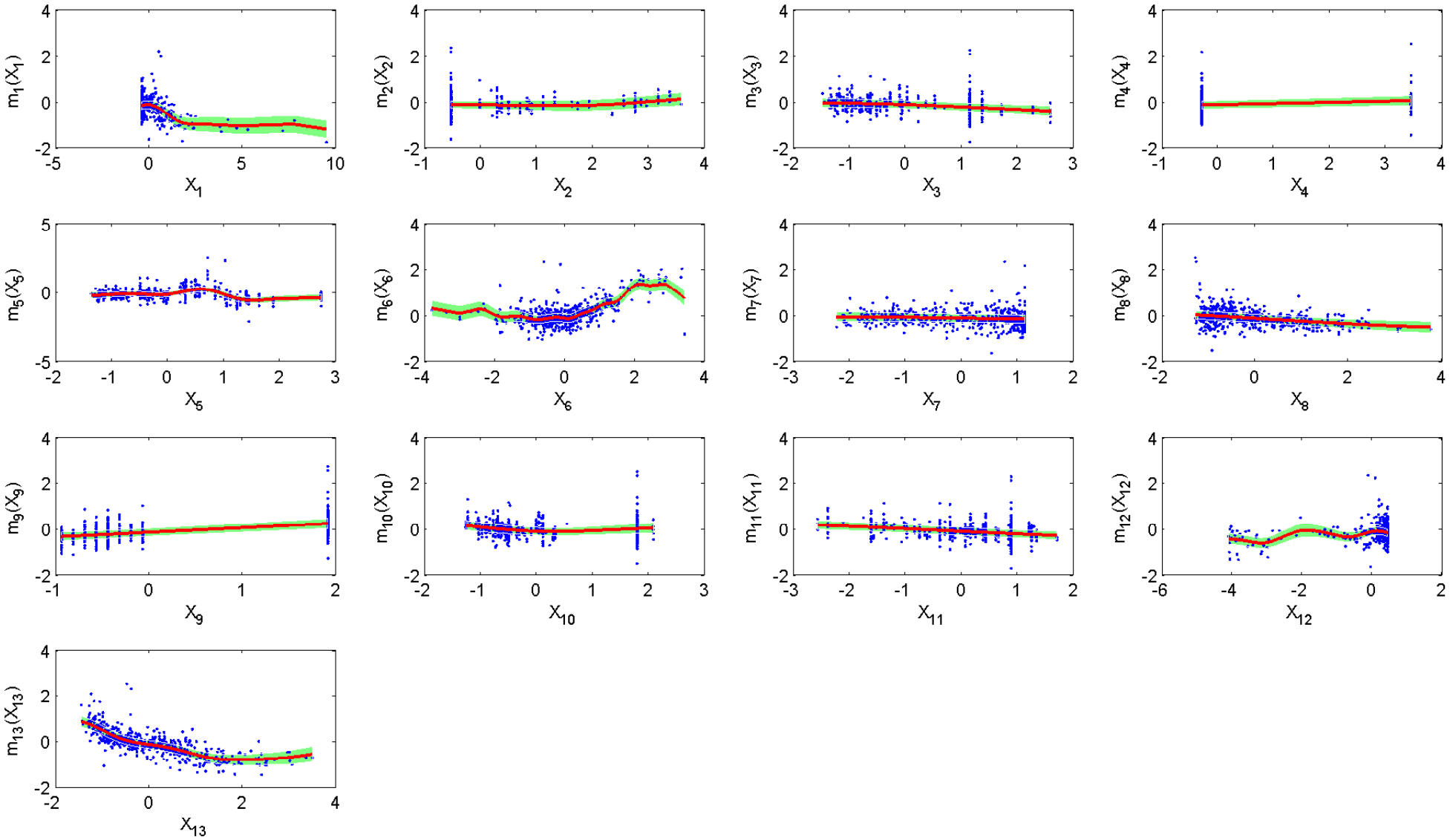
- Repeated over 500 times
- Split random in training and test samples (80/20)
- $\text{MSE}_{\text{full}} = 0.22$ and $\text{MSE}_{\text{reduced}} = 0.21$

Boston housing data (cont'd)

- Repeated over 500 times
- Split random in training and test samples (80/20)
- $\text{MSE}_{\text{full}} = 0.22$ and $\text{MSE}_{\text{reduced}} = 0.21$

Variable	description
1	per capita crime rate by town
2	proportion of residential land zoned for lots over $25,000$ sq.ft
3	proportion of non-retail business acres per town
4	Charles River dummy variable
5	nitric oxides concentration
6	average number of rooms per dwelling
7	proportion of owner-occupied units built prior to 1940
8	weighted distances to five Boston employment centres
9	index of accessibility to radial highways
10	full-value property-tax rate per USD
11	pupil-teacher ratio by town
12	proportion of African-Americans by town
13	percentage of lower status of the population

boston housing data (cont'd)



Outline

Conclusion

- Feature selection technique is easy to implement
- Results are easy to interpret
- Solid theoretical foundations
- Model-free methodology

Conclusion

- Feature selection technique is easy to implement
- Results are easy to interpret
- Solid theoretical foundations
- Model-free methodology

- No theoretical result if d grows faster than n due to the model-free character

Thank you for your attention

Any questions...?

References

- Tibshirani R. (1996). Regression analysis and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, 58:267–288
- Tibshirani R., Hastie T., Narasimhan B. & Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression (2002). *PNAS*, 99(10):6567-6572
- Chun L.H. & Wen C.L. (2010). Detecting differentially expressed genes in heterogeneous disease using half Student's t-test. *Int.I.Epidemiol*, 10:18
- Antoniadis, A., Gijbels, I. & Verhasselt, A. (2012). Variable selection in additive models using P-splines. *Technometrics*, 54(4):425–438
- Greenshtein E. & Ritov Y. (2004). Persistence in high-dimensional predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988
- Bartlett P., Mendelson S. & Neeman J. (2009). l_1-regularized regression: persistence and oracle inequality. Manuscript
- Meinshausen N. & Bühlmann P. (2006). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B*, 72:417–473
- Efron B., Hastie T., Johnstone I. & Tibshirani R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, 32:407–451
- Bühlmann P. & van de Geer S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Springer

References

Blum J.R., Chernoff M., Rosenblatt M. & Teicher H. (1958). Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics*, 10:222-229

De Brabanter K. & Györfi L. (2012). Feature selection via detecting ineffective features. *International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: theory and applications.*