High-dimensional variable selection by decorrelation: Introducing the CAR-score

Verena Zuber

joint work with
Korbinian Strimmer

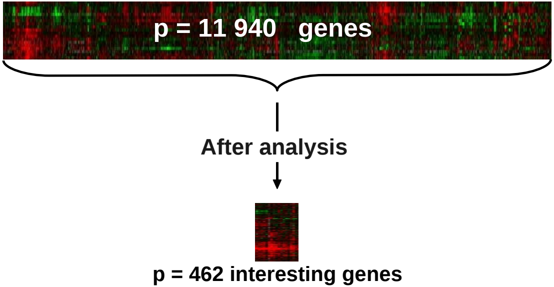
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), University of Leipzig

Workshop on
Validation in Statistics and Machine Learning
Oktober 6th, 2010

"It is a very sad thing that nowadays there is so little useless information."

Oscar Wilde
published in Saturday Review (1894)

Today: Analysis of gene-expression data⋆



⋆ Lu et al. (2004): " Gene regulation and DNA damage in the ageing human brain"

I. The Linear Model: Focus on Variable Selection and Importance

The linear model (population level)

$$\underbrace{Y}_{1\times 1} = \underbrace{\beta^t}_{1\times p} \underbrace{\boldsymbol{X}}_{p\times 1} + \underbrace{\epsilon}_{1\times 1}$$
$$= \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

- Y: 1 dependent variable or response (1×1, with $E(Y) = 0$)
- $\boldsymbol{X}$: p explaining variables ($p \times 1$, with $E(\boldsymbol{X}) = 0, Var(\boldsymbol{X}) = \boldsymbol{V}$)
- $\boldsymbol{\beta}$: p regression coefficients ($1 \times p$)
  - Interpretation: β_i, with $i \in \{1, \ldots, p\}$, gives the influence of X_i on Y conditional on all the other $p-1$ variables
  - The residual sum of squares is optimized by:

$$\boldsymbol{\beta} = \quad cov(\boldsymbol{X})^{-1} cov(\boldsymbol{X} Y) = \Sigma_{XX}^{-1} \Sigma_{XY}$$

- ϵ: Irreducible error with $E(\epsilon) = 0$

Strategies for determining a "good" model

1. Variable selection
    - Information criteria based on penalized residual sum of squares [George (2000)], e.g. AIC, BIC, C_p, RIC etc.
    - Penalized regression models like Lasso [Tibshirani (1996)], Elastic Net [Zou and Hastie (2005)], SCOUT [Witten and Tibshirani (2009)] etc.
2. Variable importance
    - Marginal Correlation between $\boldsymbol{X}$ and Y
      e.g. Sure Independence Screening [Fan and Lv (2008)]
    - Metrics for relative importance, like squared standardized β, Pratt's metric or other decompositions of R^2
      For a comprehensive overview see Grömping (2006).

Decorrelation offers a new quantity for
variable selection <u>and</u> variable importance.

How decorrelation leads to a new tool
for variable selection and quantifying variable importance:

II. Presenting Correlation Adjusted CoRRelation, the CAR-score

Definition of the CAR-score

We define the p-dimensional CAR-score vector (Correlation Adjusted CoRRelation) $\boldsymbol{\omega}$ as:

$$\underbrace{\boldsymbol{\omega}}_{p\times 1} = \underbrace{\boldsymbol{P}^{-1/2}}_{p\times p} \underbrace{\boldsymbol{P}_{\boldsymbol{X}Y}}_{p\times 1}$$

- $\boldsymbol{P}$: Correlation of $\boldsymbol{X}$
- $\boldsymbol{P}_{\boldsymbol{X}Y}$: Vector of marginal correlations between $\boldsymbol{X}$ and Y

Criterion for variable importance:
We propose to use $\omega^2(i)$ to quantify the importance of variable X_i in the linear model, with $i \in 1, ..., p$

Properties of the CAR-score I

1. Deduction from the best linear predictor:
   The CAR-score quantifies the influence of a decorrelated and standardized variable on the best linear predictor $Y^\star$.
2. Reformulating the decomposition of variance:
   The CAR-score leads to a coherent additive decomposition of the proportion of variance explained (on the sample level: coefficient of determination, R^2).
3. The CAR-score as a quantity for variable importance

1. The best linear predictor

$$Y^\star = \underbrace{\beta^t}_{\Sigma_{XY}^t \Sigma_{XX}^{-1}} \boldsymbol{X}$$

After some simple transformations the standardized best linear predictor $Y^\star$ simplifies to the following decomposition:

$$Y^\star/\sigma_Y = \omega^t\, \delta(\boldsymbol{X})$$

- The decorrelated and standardized data $\delta(\boldsymbol{X})$; $Cov(\delta(\boldsymbol{X})) = diag(1)$

$$\underbrace{\delta(\boldsymbol{X})}_{p\times 1} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X}$$

- The correlation between $\boldsymbol{X}$ and Y adjusted for the correlation among $\boldsymbol{X}$:

$$\underbrace{\omega}_{p\times 1} = \boldsymbol{P}^{-1/2}\boldsymbol{P}_{XY}$$

2. Decomposition of the proportion of variance explained

- Total variance: $Var(Y) = \sigma_Y^2$
- Explained variance:

$$\begin{aligned} Var(Y^\star) &= \sigma_Y^2 Var(\boldsymbol{\omega}^t \boldsymbol{\delta}(\boldsymbol{X})) \\ &= \sigma_Y^2 \boldsymbol{\omega}^t \underbrace{Var(\boldsymbol{\delta}(\boldsymbol{X}))}_{diag(1)} \boldsymbol{\omega} \\ &= \sigma_Y^2 \boldsymbol{\omega}^t \boldsymbol{\omega} \end{aligned}$$

- The decomposition of variance rewritten in CAR-scores:

$$\begin{aligned} \overbrace{Var(Y)}^{\text{Total variance}} &= \overbrace{Var(Y^\star)}^{\text{Explained variance}} + \overbrace{Var(Y - Y^\star)}^{\text{Unexplained variance}} \\ \sigma_Y^2 &= \sigma_Y^2(\boldsymbol{\omega}^t \boldsymbol{\omega}) + \sigma_Y^2(1 - \boldsymbol{\omega}^t \boldsymbol{\omega}) \end{aligned}$$

2. Decomposition of the proportion of variance explained II

Proportion of variance explained:

$$\frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\sigma_Y^2 \boldsymbol{\omega}^t \boldsymbol{\omega}}{\sigma_Y^2} = \boldsymbol{\omega}^t \boldsymbol{\omega} = \sum_{i=1}^{p} \omega_i^2$$

- The sum of squared CAR-scores adds up to the proportion of variance explained.
- Note: In the set-up of discriminant analysis: The sum of squared correlation adjusted t (CAT)-scores [Zuber and Strimmer (2009)] adds up to Hotelling's T.

3. The CAR-score as quantity for variable importance

1. Proper decomposition of the proportion of variance explained:

$$\frac{\text{Explained Variance}}{\text{Total Variance}} = \sum_{i=1}^{p} \omega_i^2$$

2. Non-negativity: $\omega_i^2 \geq 0$
3. Inclusion-Property: $\omega_i^2 \neq 0$ if $\beta_i \neq 0$
4. Exclusion-Property: $\omega_i^2 = 0$ if $\beta_i = 0$
   The CAR-score fulfills the Exclusion-Property only
   if there is no correlation between the null variables with $\beta = 0$
   and non-null variables with $\beta \neq 0$

Properties of the CAR-score II

4. Connections to other quantities for variable importance:

$$\overbrace{\boldsymbol{P}_{\boldsymbol{X}Y}}^{\text{Correlation}} \quad \xrightarrow{\boldsymbol{P}^{-1/2}} \quad \overbrace{\boldsymbol{P}^{-1/2}\boldsymbol{P}_{\boldsymbol{X}Y} = \boldsymbol{\omega}}^{\text{CAR-score}} \quad \xrightarrow{\boldsymbol{P}^{-1/2}} \quad \overbrace{\boldsymbol{P}^{-1/2}\boldsymbol{\omega} = \boldsymbol{\beta}_{\text{std}}}^{\text{Std. Regression Coeff.}}$$

5. Oracle CAR-score: *If we know*
   - which variables are null or non-null and
   - that there is no correlation between null and non-null variables

then any consistent estimate of the CAR-score $\boldsymbol{\omega} = \boldsymbol{P}^{1/2}\boldsymbol{\beta}_{\text{std}}$ equals 0 for the null variables:

$$\boldsymbol{\omega} = \underbrace{\begin{pmatrix} \boldsymbol{P}_{\text{non-null}} & 0 \\ 0 & \boldsymbol{P}_{\text{null}} \end{pmatrix}^{1/2}}_{\boldsymbol{P}^{1/2}} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_{\text{std,non-null}} \\ 0 \end{pmatrix}}_{\boldsymbol{\beta}_{\text{std}}} = \begin{pmatrix} \boldsymbol{\omega}_{\text{non-null}} \\ 0 \end{pmatrix}$$

Properties of the CAR-score III

6. Distribution of the empirical squared CAR-score under H_0:
   $\hat{\omega}^2(i)$ follows $Beta(\frac{1}{2}, \frac{n-2}{2})$
7. Grouping Property: When two variables X_i and X_j are correlated, their CAR-scores ω_i and ω_j tend to be equal:

$$| \rho(X_i, X_j) | \to 1 \quad \Rightarrow \quad \omega_i^2 - \omega_j^2 \to 0$$

8. Orthogonal Property (The CAR-score for a group of variables): The importance of a group of variables $1, ..., g$ is given by:

$$\omega_{group}^2 = \sum_{i=1}^{g} \omega_i^2 = \omega_1^2 + ... + \omega_g^2$$

The CAR-score is a population quantity;
thus it is not tied to any inference framework.
Any kind of "good" estimate can be used.

A simple recipe for variable selection:

1. (If p is too large, a prescreening is advisable.
   Limitation: Estimation of the $p \times p$ correlation matrix $\boldsymbol{P}$)
2. Estimate the CAR-scores:
   - Large sample case ($n >> p$): Empirical estimates
   - Small n, large p: Regularized estimates, like shrinkage procedures or penalized maximum likelihood estimates
3. Rank the variables according to their squared CAR-score
4. Choose a suitable cut-off (a fixed cut-off corresponds to information criteria like AIC, BIC, etc)
5. Refit the linear model based on the remaining variables

IV. Results

All analysis is performed in R

- care: Empirical and shrinkage estimates for the CAR-score
- relaimpo: Relative importance of variables
- scout: Implementation of Lasso and Elastic Net
- fdrtool: False (non) discovery rate

Simulation: The Set-up

- X is (multivariate) Gaussian distributed: $X \sim MvN(0, R)$
- ϵ is Gaussian distributed: $\epsilon \sim N(0, \sigma^2 = 9)$

- Set-up 1:
  - Low dimensional: $p = 8$ and $n = 50 - 100$
  - `beta=c(3,1.5,0,0,2,0,0,0)`
  - Autocorrelation: $\rho(x_i, x_j) = 0.5^{|i-j|}$
  - Signal variance to noise variance: 2.36
- Set-up 2:
  - Large p, small n: $p = 40$ and $n = 10 - 50$
  - $p = 10$ non-null variables with `beta[1:10]=c(3,3,3,3,3,-2,-2,-2,-2,-2)` and $p = 30$ null variables
  - Pairwise correlation of $\rho = 0.9$ among the non-null variables
  - Signal variance to noise variance: 3.22

Simulation: Comparing the results

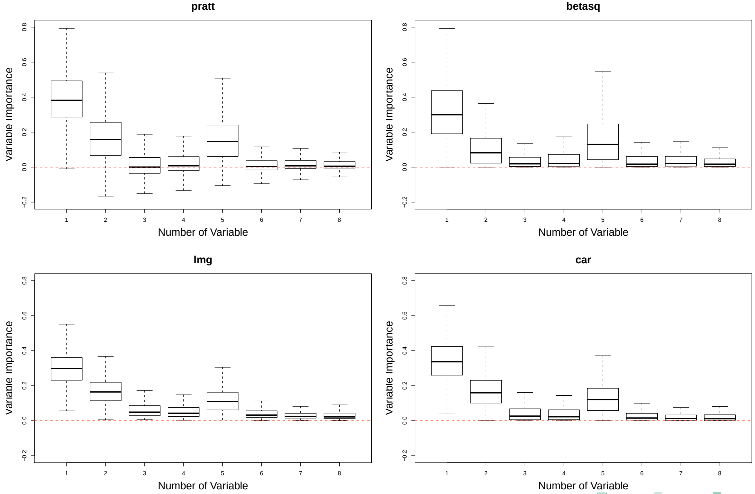
- What to compare?
  1. Variable selection:
     Mean model error, median model size and the β-coefficients
  2. Variable importance:
     Quantity of the different metrics
- The competitors:
  1. Variable selection:
     Elastic Net, Lasso and Ordinary Least Squares
  2. Variable importance:
     Squared β_{std}'s, Pratt's measure and the `lmg`-measure
- Set-up 1: Empirical CAR-score; Set-up 2: Shrinkage CAR-score
- The CAR-scores are used for variable selection,
  then the linear model is refitted

Mean model error with (SE) and median model size

	CAR	Elastic Net	Lasso	OLS
Set-up 1:				
$n = 50$	**119** (7)	130 (6)	148 (6)	230 (9)
	3	5	5	8
$n = 100$	**55** (3)	58 (2)	59 (3)	99 (3)
	3	5	5	8
Set-up 2:				
$n = 10$	**1482** (44)	1501 (45)	1905 (75)	—
	10	13	6	—
$n = 20$	**838** (30)	950 (26)	1041 (29)	—
	9	**10**	6	—
$n = 50$	**358** (11)	571 (10)	608 (8)	5032 (214)
	10	7	5	40

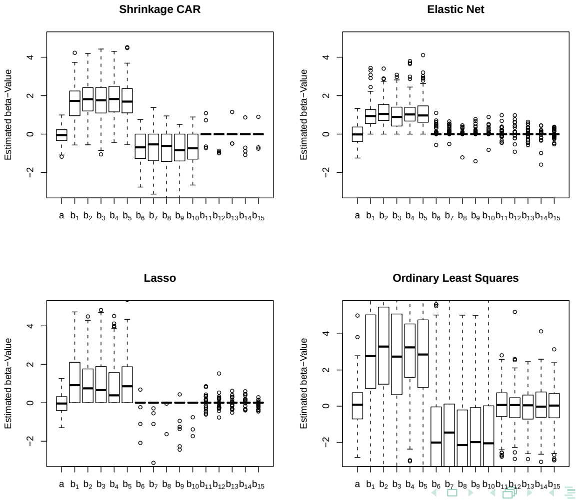
Set-up 1: Boxplots of the estimated variable importance

```
beta=c(3,1.5,0,0,2,0,0,0)
```



Set-up 2: Boxplots of the first 15 estimated β-values

```
beta=c(3,3,3,3,3,-2,-2,-2,-2,-2, 0,0,...)
```

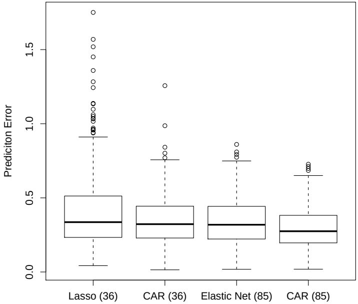


"Gene regulation and DNA damage in the ageing human brain"

from Lu et al. in Nature (2004)

- The data is available on the Gene Expression Omnibus ("GSE1572")
- $n = 30$ and $p = 12\,625$
- Y: Age of the individual (26-106 years)
- $\boldsymbol{X}$: Gene expression of postmortem brain tissue (frontal cortex) (Platform: Affymetrix Human Genome U95 Version 2 Array)
- A prescreening is performed using the empirical marginal correlations and FNDR control: Remaining size $p = 403$
- Model size of the competing procedures:
  - Lasso: 36 genes
  - Elastic Net: 85 genes
  - CAR-score: $50 - 60$ genes
- All procedures include different variables.

Ageing: The cross-validated prediction error



	Model Size	Mean Prediction error (SE)
Lasso	36	0.4006 (0.0011)
CAR	36	0.3357 (0.0070)
Elastic Net	85	0.3417 (0.0068)
CAR	85	0.2960 (0.0059)

IV. Conclusion

Summary

1. We introduce a remarkable simple way of quantifying variable importance and selecting variables in the linear model:
   The CAR-score
2. The CAR-score is embedded elegantly in the theoretical framework of the linear model:
   - The CAR-score quantifies the influence of a decorrelated variable on the best linear predictor.
   - It leads to a coherent decomposition of the proportion of variance explained.
3. Simulations show that the CAR-score achieves a lower model error than Lasso and Elastic Net and identifies the correct model size.
4. In the analysis of real data the CAR-score achieves a lower prediction error than competing procedures.

The preprint of Zuber and Strimmer (2010):
"Variable importance and model selection by decorrelation"
is available on:

- http://arxiv.org/abs/1007.5516
- http://www.uni-leipzig.de/~zuber/

care(CAR-Estimation)-package available from CRAN:

- cran.r-project.org/web/packages/care/index.html

Thank You Very Much
For Your Attention!

- Fan and Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties." JASA (96) 1348-1360
- Fan and Lv (2008): "Sure Independence Screening for Ultra-High Dimensional Feature Space." JRSS Series B (70) 849-911
- Grömping (2006): "Relative Importance for Linear Regression in R" Journal of Statistical Software 17 (1)
- Lu et al. (2004): " Gene regulation and DNA damage in the ageing human brain." Nature (429) 883-891
- Tibshirani (1996): "Regression shrinkage and selection via the lasso." JRSS Series B. (58) 267-288
- Witten and Tibshirani (2009): "Covariance-regularized regression and classification for high-dimensional problems." JRSS Series B.
- Zou and Hastie (2005): "Regularization and variable selection via the elastic net." JRSS Series B. (67) 301-320
- Zuber and Strimmer (2009): "Gene ranking and biomarker discovery under correlation." Bioinformatics (25) 2700-2707