Variable selection for correlated data in high dimension using decorrelation methods

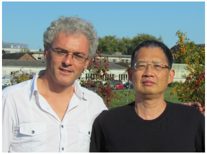
Emeline Perthame, David Causeur, Ching-Fan Sheu, Chloé Friguet

Variable selection for correlated data in high dimension using decorrelation methods

Emeline Perthame
INRIA, team MISTIS, Grenoble

Joint work with

David Causeur — *Agrocampus, Rennes*
Ching-Fan Sheu — *NCKU, Tainan, Taiwan*
Chloé Friguet — *UBS, Vannes*



StatLearn, Vannes, April 2016

Outline

1. Introduction

2. Impact of dependence and dependence modeling

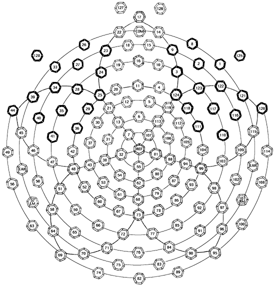
3. Disentangling signal from noise ...

- ... for a multiple testing issue
- ... for a supervised classification issue

4. Conclusion

The instrument: a 128-channel geodesic sensor net

- Electroencephalography (EEG) is the recording of electrical activity at scalp locations over time.
- The recorded EEG traces, which are time locked to external events, are averaged to form the event-related (brain) potentials (ERPs).



Auditory oddball experiment

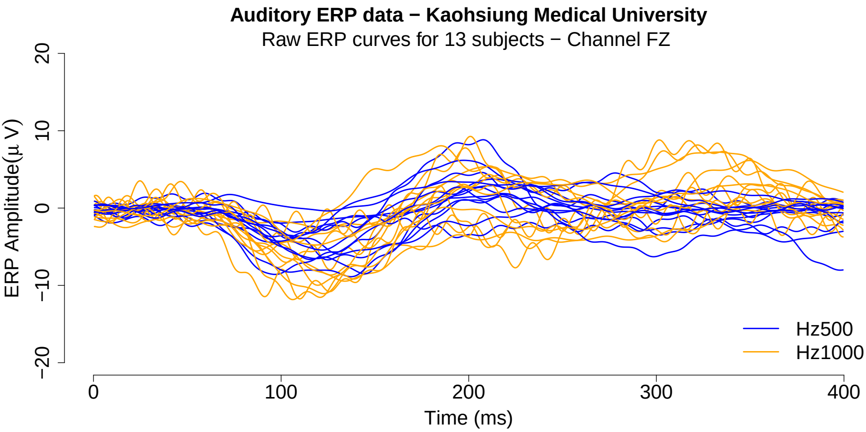
A very commonly used experimental task

- Two auditory stimuli are presented to subjects
  - A stimulus (500Hz) occurring frequently
  - A stimulus (1000Hz) occurring infrequently
- ERPs are recorded on a 400 ms interval after the onset.

Motivations

- Auditory evoked potential (AEP): elicited by auditory stimulus
- Mismatch negativity (MMN): elicited by any change in the stimulus (odd/frequent)
- AEP and MMN are electrophysiological marker candidates for psychiatric disorders such as schizophrenia

ERP curves



→ Signal detection: is there any difference between the two conditions ?

→ Signal identification: when does the difference occur ?

Linear model framework for ERP curves

At time t for subject i in condition j

- Multivariate analysis of variance model

$$Y_{ijt} = \mu_t + \alpha_{it} + \gamma_{jt} + \varepsilon_{ijt}$$

- Functional analysis of variance model

$$Y_{ijt} = \sum_{s=1}^{S} m_s \varphi_s(t) + \sum_{s=1}^{S} a_{is} \varphi_s(t) + \sum_{s=1}^{S} g_{js} \varphi_s(t) + \varepsilon_{ijt}$$

where $\varphi_s(.)$, $s = 1, \ldots, S$ are B-splines.

Linear model framework for ERP curves

At time t for subject i in condition j

$$Y_{ijt} = \mu_t + \alpha_{it} + \gamma_{jt} + \varepsilon_{ijt}$$

Signal detection

- Is there any difference between the two conditions ?

$$H_0: \text{ for } t = 1, \ldots, T \text{ and } j = 1, 2, \gamma_{jt} = 0$$

- Is it relevant to predict the label from ERP curves ?

$\rightarrow$ High dimension: need for variable selection

Signal identification

$$\text{For } t = 1, \ldots, T, H_{0t}: \text{ for } j = 1, 2, \gamma_{jt} = 0$$

Some approaches

Detection

- F-test for multivariate (or functional) ANOVA [1]
- Optimal detection (Higher Criticism [2])

Supervised classification

- Ignoring correlations: Naive approaches [3]
- Introducing sparsity: Lasso, Sparse LDA [4]

Identification

- FDR controlling: Benjamini-Hochberg ...

→ Efficient under independence

1. Bugli and Lambert, 2006, Stat Med
2. Donoho and Jin, 2004, AOS
3. Bickel and Levina, 2004, Bernoulli ; Tibshirani et al., 2003, Stat Sc
4. Tibshirani, 1996, JRSS ; Clemmensen et al., 2011, Technometrics

Guthrie-Buchwald procedure[5]

- Assumes an auto-regressive process with auto-correlation ρ
- Distribution of L_ρ under the null

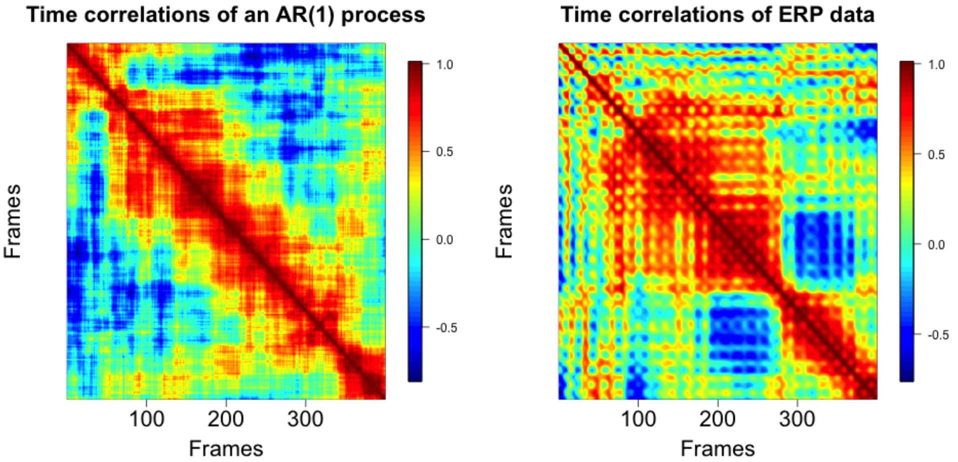
$$L_\rho = \#\{t, p_t \leq \alpha\}$$

where $(p_1, \ldots, p_T)$ are p-values and α is a preset level

- A time interval is rejected if it is significant at the preset level and longer than usual time intervals

5. Guthrie and Buchwald, 1991, Psychophysiology

Strong and complex temporal dependence structure



→ Dependence affects the stability of selection procedures

Outline

1. Introduction

2. Impact of dependence and dependence modeling

3. Disentangling signal from noise ...

   - ... for a multiple testing issue
   - ... for a supervised classification issue

4. Conclusion

Rare and Weak paradigm [6]

- Two components mixture for test statistics

$$\mathcal{T} = \mu + \varepsilon, \varepsilon \sim \mathcal{N}(0, \mathbb{I}_T)$$

- Where signal is
  - Rare

$$\eta = T^{-\beta}, \beta \in (\frac{1}{2}, 1)$$

  - Weak

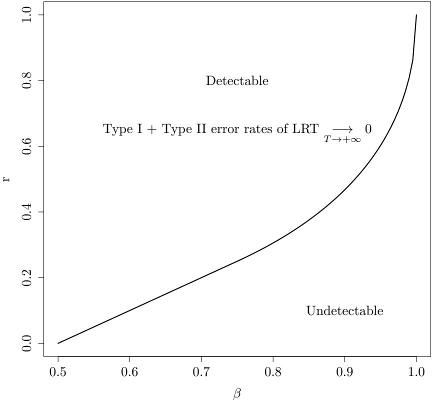
$$A = \sqrt{2r\log(T)}, r \in (0, 1)$$

6. Donoho and Jin, 2004, AOS ; 2008, PNAS

Phase diagram under independence[7]

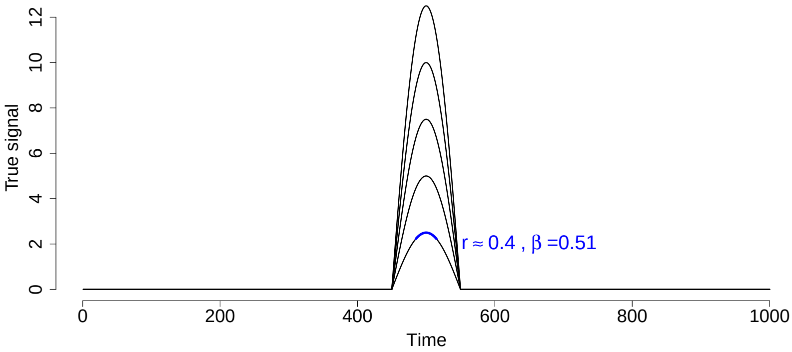
- Signal is detectable when $r > \rho^*(\beta)$:

$$\rho_D^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} < \beta < 1. \end{cases}$$



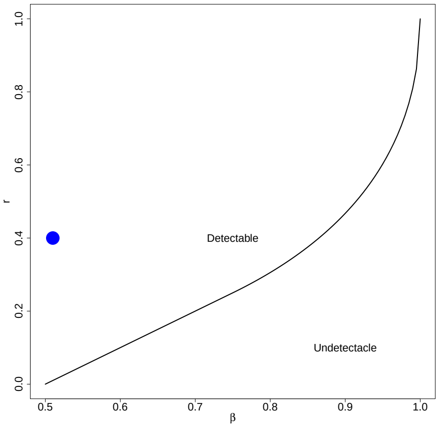
7. Ingster, 1999, Math Meth of Stat ; Donoho and Jin, 2004, AOS

Impact of dependence - Signal identification



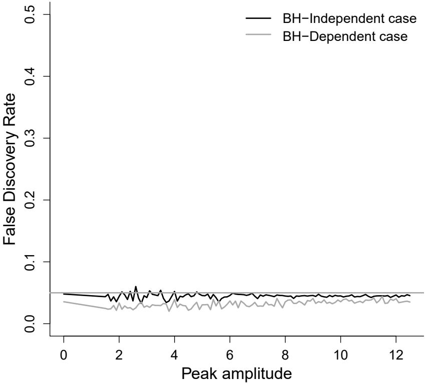
- Independence and ERP time dependence pattern
- 1000 datasets for each amplitude
- Benjamini Hochberg correction

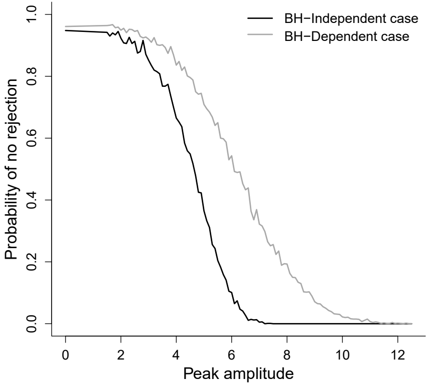
Impact of dependence - Signal identification



- Independence and ERP time dependence pattern
- 1000 datasets for each amplitude
- Benjamini Hochberg correction

Impact of dependence - Signal identification

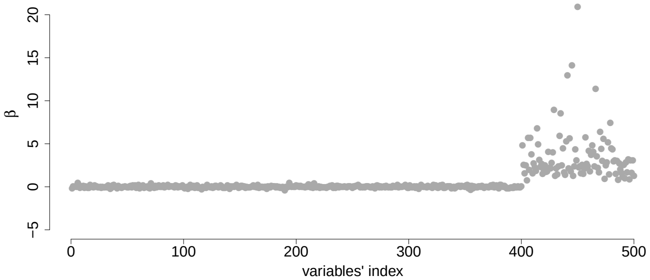




- Instability of multiple testing procedures

$$\text{FDR} = \text{pFDR}(1\text{-PNR})$$

Impact of dependence - Variable selection

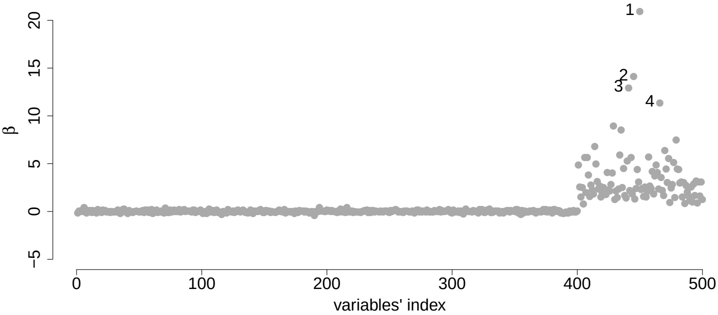


$$\log \frac{\mathbb{P}(Y=2|X)}{\mathbb{P}(Y=1|X)} = \beta_0 + \beta' x$$

- Independence and ERP time dependence pattern
- 1000 datasets for each dependence structure
- Variable selection performed by Lasso [8]

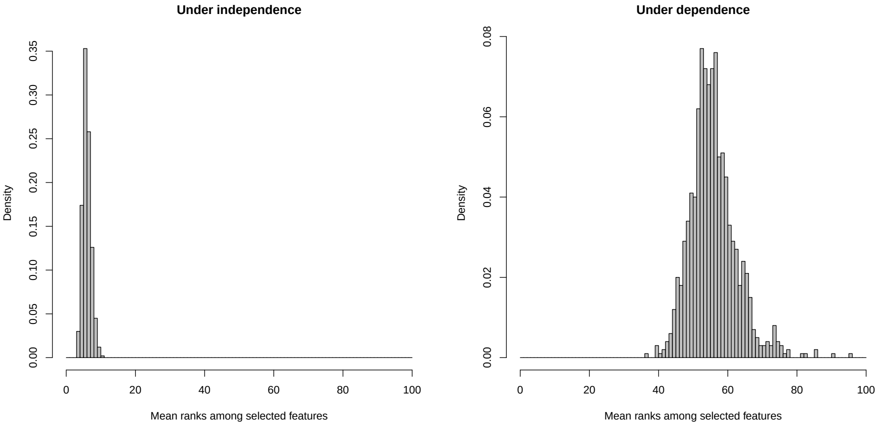
8. `glmnet` `R` package, Friedman et al., 2010, JSS

Impact of dependence - Variable selection



- Predictor X_t is assessed by its rank r_t deduced from its regression coefficient
- Relevance of a selected set $\mathcal{S}$ is given by the mean rank in $\mathcal{S}$: $r_{\mathcal{S}} = \frac{1}{\#\mathcal{S}} \sum_{t \in S} r_t$

Impact of dependence - Variable selection



- Relevance: the most predictive variables are not selected under dependence
- Stability: selected subsets are not reproducible

Impact of dependence - Improving stability

- Bootstrap
  - Bolasso [9]
  - Stability selection [10]
- Dependence modeling
  - Surrogate variable analysis [11]
  - Latent effect adjustment after primary projection [12]
  - Factor analysis for multiple testing [13]

9. Bach, 2008, Proceedings ICML
10. Meinshausen and Bühlmann, 2010, JRSS
11. Leek and Storey, 2007, PLoS Genetics
12. Sun, Zhang and Owen, 2012, AOAS
13. Friguet, Kloareg and Causeur, 2009, JASA

Factor modeling of dependence

- Distribution of ERP curves

$$X = (X_1, \dots, X_T) | Y = y \sim \mathcal{N}_T(\mu_y, \Sigma)$$

- Latent factor modeling

$$X = \mu_y + BZ + e \text{ with } e \sim \mathcal{N}_T(0, \Psi)$$

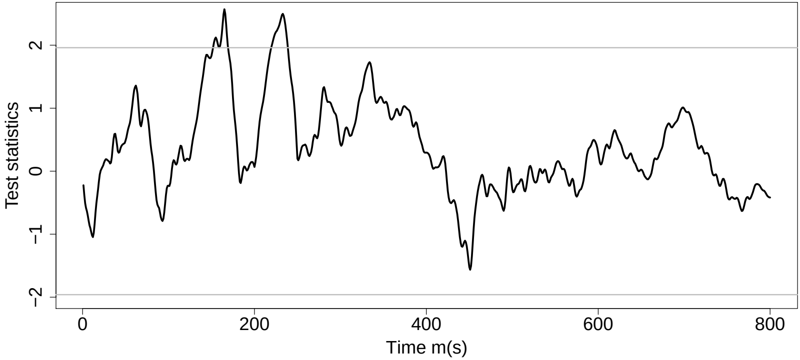
$$\Psi \text{ diagonal, } \operatorname{rank}(B) = q,$$

$$Z \sim \mathcal{N}_q(0, \mathbb{I}_q),$$

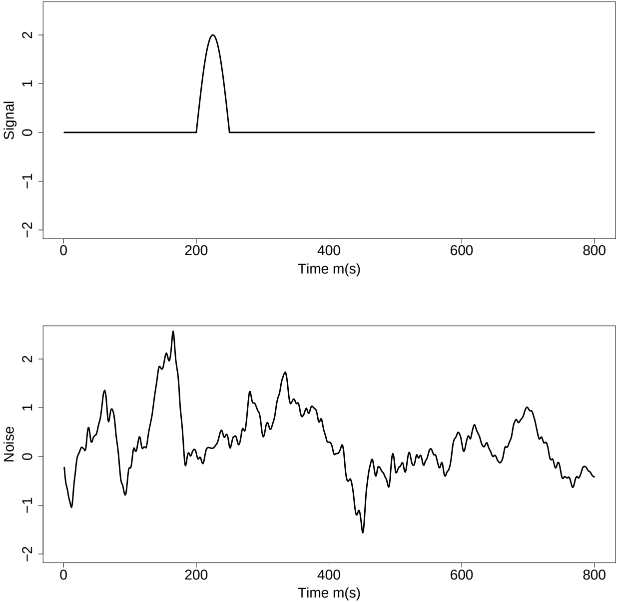
- Decomposition of covariance matrix

$$\Sigma = \Psi + BB'$$

Signal is hidden by noise



Signal is hidden by noise



Outline

Multiple testing issue

- ERP measure at time t, for subject i,

$$Y_{ijt} = \mu_t + \alpha_{it} + \gamma_{jt} + \varepsilon_{ijt}$$

- In matrix notations

$$Y_t = \mu_t + X_0\alpha_t + X\gamma_t + \varepsilon_t$$

with $\mathbb{V}(\varepsilon_1, \ldots, \varepsilon_T) = \Sigma$

- Multiple testing for $t = 1, \ldots, T$

$$H_{0,t} : \gamma_t = 0$$

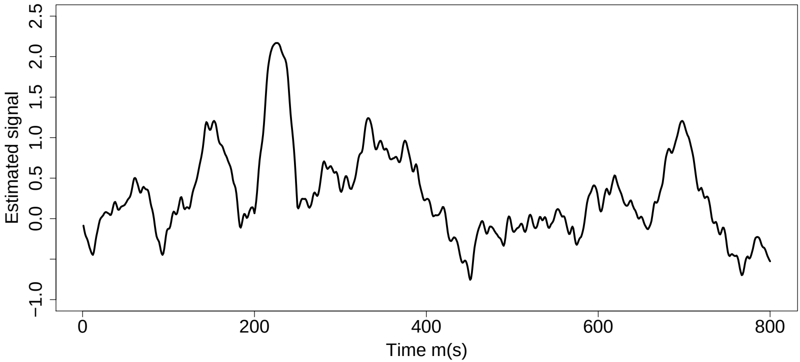
- Dependence among tests

A prior knowledge of the signal

- OLS signal estimation of $\gamma = (\gamma_1, \ldots, \gamma_T)$

$$\hat{\gamma} = \gamma + \delta$$

with $\delta \sim \mathcal{N}(0, \widetilde{\Sigma})$ and $\widetilde{\Sigma} \propto \Sigma$

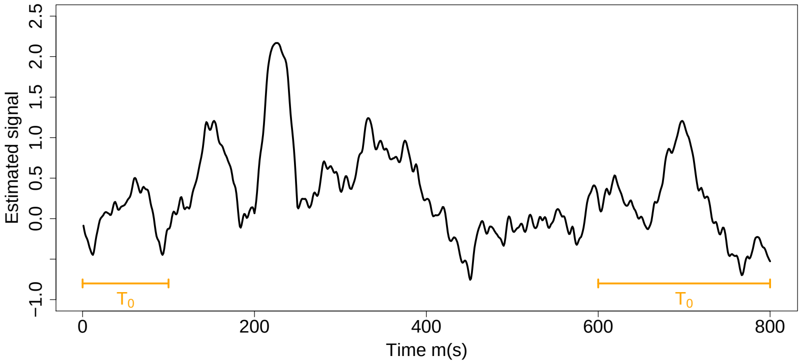


A prior knowledge of the signal

- OLS signal estimation of $\gamma = (\gamma_1, \ldots, \gamma_T)$

$$\hat{\gamma} = \gamma + \delta$$

with $\delta \sim \mathcal{N}(0, \widetilde{\Sigma})$ and $\widetilde{\Sigma} \propto \Sigma$



A prior knowledge of the signal

- OLS signal estimation of $\gamma = (\gamma_1, \dots, \gamma_T)$

$$\hat{\gamma} = \gamma + \delta$$

with $\delta \sim \mathcal{N}(0, \widetilde{\Sigma})$ and $\widetilde{\Sigma} \propto \Sigma$

- Noise is somewhere observed without signal

$$\begin{pmatrix} \delta_0 \\ \delta_{-0} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \widetilde{\Sigma}_{0,0} & \widetilde{\Sigma}'_{-0,0} \\ \widetilde{\Sigma}_{-0,0} & \widetilde{\Sigma}_{-0,-0} \end{pmatrix}\right]$$

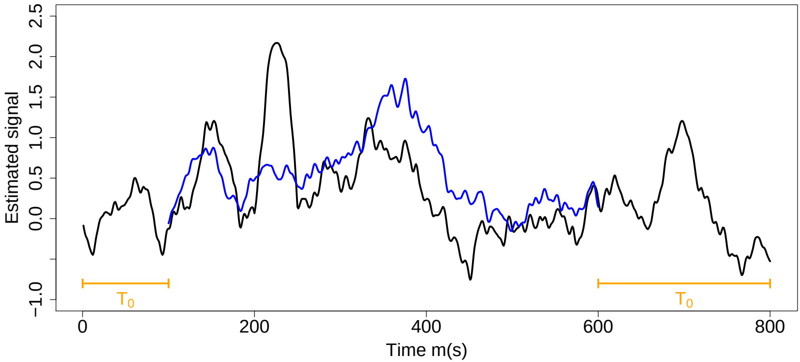
- And can be estimated elsewhere

$$\hat{\delta}_{-0} = \hat{\Sigma}_{-0,0}\hat{\Sigma}_{0,0}^{-1}\hat{\delta}_0$$

A prior knowledge of the signal

- And can be estimated elsewhere

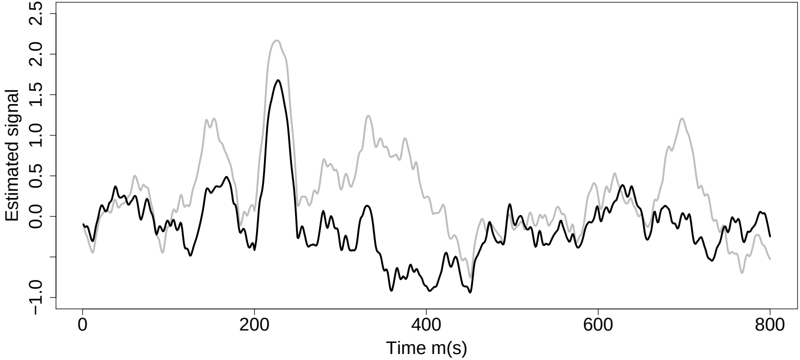
$$\hat{\delta}_{-0} = \hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1} \hat{\delta}_0$$



A prior knowledge of the signal

- New estimation of the signal

$$\hat{\gamma}^{\text{new}} = \hat{\gamma} - \hat{\delta}$$



Iterative algorithm

- New estimation of the signal

$$\hat{\gamma}^{\text{new}} = \hat{\gamma} - \hat{\delta}$$

- Update of residual errors $\hat{\varepsilon}^{\text{new}} = Y_t - (\hat{\mu}_t + \hat{\alpha}_{it} + \hat{\gamma}_t^{\text{new}})$
- New estimation of covariance matrix
- Alternates estimation of signal and covariance structure
- Until convergence of test statistics
- Update of T_0

Choice of T_0

Prior knowledge

- ERP: psychologists may know that signal does not occur before/after some time points
- Genomics: biologists may know that some genes are not involved in a biological process

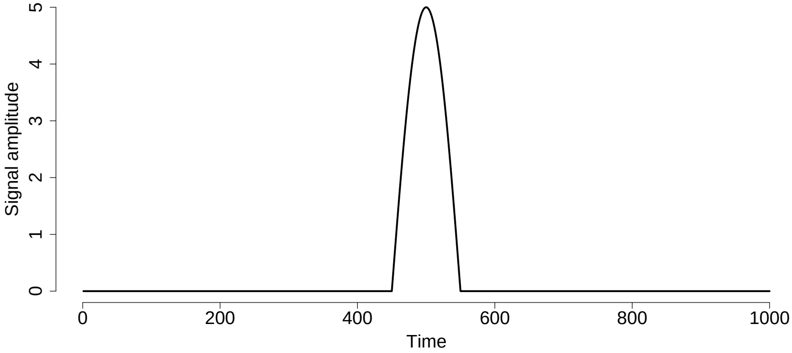
No prior knowledge

- Conservative approach

$$T_0 = \{t, p_t \geq t_0\}$$

where $(p_1, \dots, p_T)$ are p-values

Simulations - Adaptive factor analysis procedure



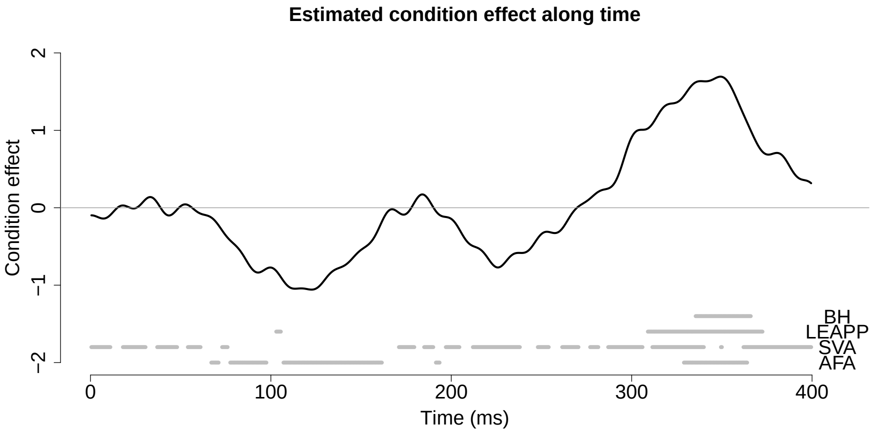
- Dependence structure of ERP experiment
- 1000 generated datasets

Simulations - Adaptive factor analysis procedure

Method	FDR [14]	TDR [15]	PD [16]
Benjamini-Hochberg	0.031	0.057	0.281
Benjamini-Yekutieli	0.009	0.011	0.101
Guthrie-Buchwald	0.086	0.233	0.538
SVA	0.088	0.151	0.599
LEAPP	0.151	0.304	0.847
AFA	0.034	0.498	1.000

14. False Discovery Rate
15. True Discovery Rate
16. Probability of Detecting the peak

Application to auditory data



80 - 120 ms: Auditory evoked potential

100 - 200 ms: Mismatch negativity for the difference curve

Conclusion

- Adaptive estimation of signal and factor model parameters
- Designed for strong dependence
- Efficient multiple testing procedure
  - FDR is controlled
  - Good detection power
- `ERP` package available on CRAN [17]

17. Causeur and Sheu, 2014, R package version 1.0.1

Outline

1. Introduction

2. Impact of dependence and dependence modeling

3. Disentangling signal from noise ...

- ... for a multiple testing issue
- ... for a supervised classification issue

4. Conclusion

Supervised classification issue

- Prediction of a label → Hz500 or Hz1000 frequency
- From ERP curves profiles $X = (X_1, \dots, X_T)$

$$(X|Y = y) \sim \mathcal{N}_p(\mu_y, \Sigma)$$

- Among linear classification rule

$$LR(x) = \log \frac{\mathbb{P}(Y = 2|X)}{\mathbb{P}(Y = 1|X)} = \beta_0 + x'\beta$$

- The best one is Bayes' rule

$$\begin{aligned} \beta &= \Sigma^{-1}(\mu_2 - \mu_1) \\ \beta_0 &= \log \frac{p_2}{p_1} - 0.5(\mu_2 + \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) \end{aligned}$$

- Theoretical misclassification rate π

Some estimation methods

Logistic regression

- Minimizing the deviance

$$(\hat{\beta}_0, \hat{\beta}) = \text{argmin}_{\beta_0,\beta} - 2\sum_{i=1}^{n} \log[1 + \exp(-V_i(\beta_0 + x_i'\beta))]$$

where $V_i = \pm 1$

- High dimension
  - ℓ_2-penalization: Ridge [18]
  - ℓ_1-penalization: Lasso [19]

18. Hoerl and Kennard, 1970, Technometrics
19. Tibshirani, 1996, JRSS

Some estimation methods

Linear Discriminant Analysis

- OLS estimate → Method of moments

$$(\hat{\beta}_0, \hat{\beta}) = \text{argmin}_{\beta_0,\beta} \sum_{i=1}^{n} [V_i - (\beta_0 + x_i'\beta)]^2, \text{ where } V_i = \pm 1$$

- High dimension
  - Ignoring correlations: Diagonal Discriminant Analysis (DDA)[18], Nearest Shrunken Centroids[19]
  - Shrinkage Discriminant Analysis[20] (SDA)
  - Sparse linear discriminant analysis[21](SLDA)

18. Bickel and Levina, 2004, Bernoulli
19. Tibshirani et al., 2003, Stat Sc
20. Ahdesmäki and Strimmer, 2010, AOAS
21. Clemmensen et al., 2011, Technometrics

Conditional classification rule

- Under factor model assumption ($\Sigma = \Psi + BB'$)

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & B \\ B' & I_q \end{pmatrix}\right]$$

- Among classification rules linear in (x, z)
- The best one is the conditional Bayes' classifier

$$LR(x, z) = \log \frac{\mathbb{P}(Y = 2|X, Z)}{\mathbb{P}(Y = 1|X, Z)} = \beta_0^* + (x - Bz)'\beta^*$$

$$\text{with } \beta^* = \Psi^{-1}(\mu_2 - \mu_1)$$

$$\beta_0^* = \log \frac{p_2}{p_1} - 0.5(\mu_2 + \mu_1)'\Psi^{-1}(\mu_2 - \mu_1)$$

- Analytical expression of misclassification rate π_Z^*

Conditional classification rule

- Bayes rule error π
- Under factor model assumption

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & B \\ B' & I_q \end{pmatrix}\right]$$

- Conditional Bayes rule error π_Z^*
- One can show that $\pi \geq \pi_Z^*$

$\rightarrow$ Theoretical superiority of conditional approach based on decorrelated data $\widetilde{X} = X - BZ$

Iterative decorrelation of data

- Estimation of μ_1 and μ_2
- Computation of centered profiles
- Estimation of factor model parameters[22] (Ψ, B)
- Decorrelation of data using generalized Thompson's formula

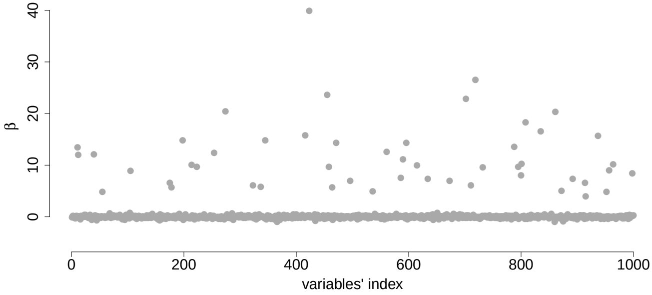
$$\tilde{x} = x - \hat{B}\hat{z}'$$

Generalized Thompson's formula

$$\widehat{Z} = \mathbb{E}_X(Z) = (I_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}\Big(x - \big[\mu_1 \mathbb{P}_X(1) + \mu_2 \mathbb{P}_X(2)\big]\Big)$$

22. Friguet, Kloareg and Causeur, 2009, JASA

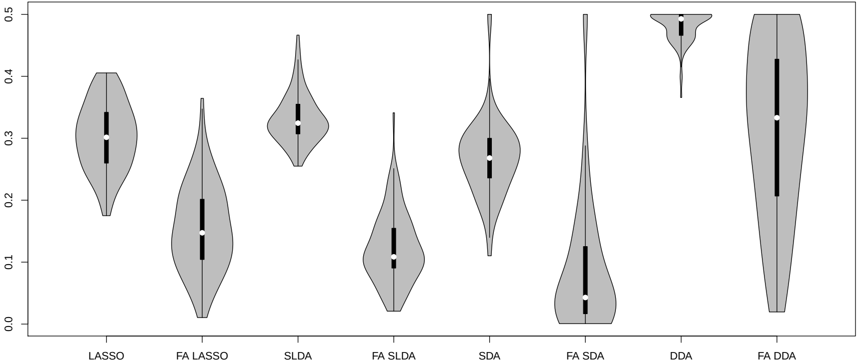
Simulations



- $n_0 = n_1 = 13$
- Various dependence structures [23]
- 1000 learning datasets
- 1 testing dataset

23. Meinshausen and Bühlmann, JRSS, 2010

Simulations - Prediction error rates



→ Variable selection methods compared to their factor-adjusted version

Simulations - Selection accuracy

Method	Nb of selected var.	Accuracy
LASSO [24]	13.10	62.36
Factor-adjusted LASSO	8.03	93.02
SLDA [25]	10.00	62.50
FA SLDA	10.00	90.90
SDA [26]	57.20	75.07
FA SDA	68.22	67.93
DDA [27]	149.42	15.58
FA DDA	97.65	48.76

24. Tibshirani, 1996, JRSS ; Friedman et al., 2010, JSS
25. Clemmensen et al., 2011, Technometrics
26. Ahdesmäki and Strimmer, 2010, AOAS
27. Bickel and Levina, 2004, Bernoulli

Conclusion

- Decorrelation method designed for prediction issues
- Preprocessing of the data which enables the use of usual selection methods
- `FADA` package available on CRAN [28]
- Application in genomics
- Adjustment for batch effect [29]

28. Perthame, Friguet and Causeur, 2014, R package version 1.2
29. Hornung, Boulesteix and Causeur, submitted

Outline

Take home message

→ Whatever the statistical analysis, it would be efficient to account for dependence because it is a *blessed* situation [30]

→ Accounting for dependence introduces hyper-parameters

- Risk of overfitting
- Results depend on the estimation of the dependence model
  - Need for robust models
  - With few parameters
  - To guarantee reproducible results

30. Hall and Jin, 2010, AOS

References

D. Causeur and C.-F. Sheu.
ERP: Significance analysis of Event-Related Potentials data, 2014.
R package version 1.0.1.

E. Perthame, C. Friguet, and D. Causeur.
FADA: Variable selection for supervised classification in high dimension, 2014.
R package version 1.2.

E. Perthame, C. Friguet, and D. Causeur.
Stability of feature selection in classification issues for high-dimensional correlated data.
Statistics and Computing, pages 1–14, 2015.

C. Sheu, E. Perthame, D. Causeur, and Y. Lee.
Accounting for time dependence in large-scale multiple testing of event-related potential data.
AOAS, 10(1):219–245, 2016.