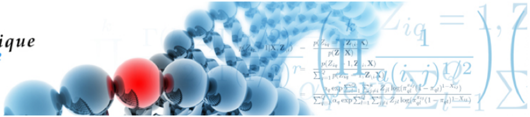


Estimation of a non-parametric variable importance measure of a continuous exposure

with Antoine Chambaz and Mark J. van der Laan
Electron. J. Stat. vol. 6 (2012) 1059–1099

Pierre Neuvial

Laboratoire Statistique et Génome
Université d'Évry Val d'Essonne
UMR CNRS 8071 – USC INRA

Clermont-Ferrand, Journées MAS 2012

Outline

1. Association between DNA copy number and gene expression

2. Targeted minimal loss estimation of association

3. Results

Outline

1. Association between DNA copy number and gene expression

2. Targeted minimal loss estimation of association

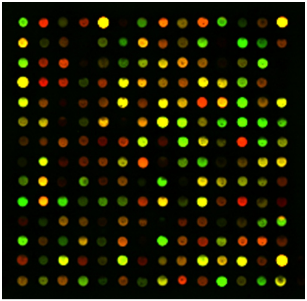
3. Results

Cancers are characterized by changes at the molecular level

Different levels of biological information

- DNA copy number
- gene expression
- DNA methylation

Quantitative measurements can be obtained from DNA microarrays

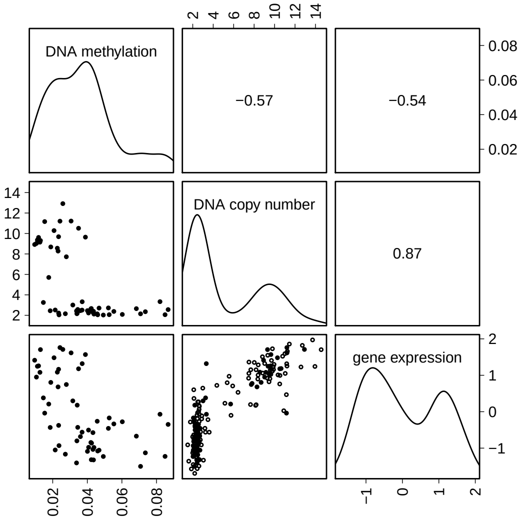


Goal : find genes that **drive** tumorigenesis

- to better understand cancer cells
- to help find new treatments

What gene-level data look like

187 GBM (brain cancer) samples from the Cancer Genome Atlas (TCGA)



Note : genomic data are complex

In the preceding plot, "gene-level data" refers to :

DNA methylation (W) : proportion of "methylated" signal at a CpG locus in the gene's promoter region.

DNA copy number (X) : smoothed normalized total copy number relative to a set of reference samples.

Expression (Y) : "unified" gene expression level across 3 platforms

Which genes are drivers ?

"Driver genes" are expected to show some **association** between DNA copy number and gene expression

⇒ **Test** for association, and **quantify** it

Methods for genome-wide scanning for gene-level associations

- linear correlations
- differential expression (T-tests) between copy number states
- canonical correlation analyses

Issues with existing methods

- they essentially identify genes that were already known to be implied
- associations may be non linear
- DNA methylation may down-regulate gene expression

Outline

1. Association between DNA copy number and gene expression
2. Targeted minimal loss estimation of association
3. Results

Definition of a parameter of interest

Observation $O = (W, X, Y) \sim P \in \mathcal{M}$ for a given gene :

- W : DNA methylation
- X : DNA copy number ; $X = 0$: **copy neutral state** (2 copies)
- Y : gene expression
- $\mathcal{M}$: non-parametric set of candidate data-gen. distributions of O

Parameter of interest (defined for all $P \in \mathcal{M}$)

$$\Psi(P) = \arg\min_{\beta \in \mathbb{R}} E_P \left[(E_P(Y|X, W) - E_P(Y|X = 0, W) - \beta X)^2 \right]$$

A non-parametric variable importance measure

Parameter of interest (defined for all $P \in \mathcal{M}$)

$$\Psi(P) = \arg\min_{\beta \in \mathbb{R}} E_P \left[(E_P(Y|X,W) - E_P(Y|X=0,W) - \beta X)^2 \right]$$

- $\Psi(P)$ **is a variable importance measure** of the "effect" of X (continuous) on Y (continuous) accounting for W
- $\Psi(P)$ **is non-parametric :** in a semi-parametric model where $E_P(Y|X,W) = E_P(Y|X=0,W) + \beta X$, we have $\Psi(P) = \beta$. By contrast, $\Psi : \mathcal{M} \to \mathbb{R}$ is defined **universally**

Interpretation on the parameter of interest

Let $\theta(P)(X,W) = E_P(Y|X,W)$, then

$$\Psi(P) = corr(X, r_P(X,W))\sqrt{\frac{E_P[r_P(X,W)^2]}{E_P[X^2]}},$$

where $r_P(X,W) = \theta(P)(X,W) - \theta(P)(0,W)$

Case where X is binary

If $X \in \{0,1\}$, then

$$\Psi(P) = E_P[(\theta_P(1,W) - \theta_P(0,W))h(W)]$$

with weight $h(W) = P(X=1|W)/P(X=1)$

Targeted minimal loss estimation (TMLE) : motivation

Goal : estimate a parameter $\Psi(P)$ from observations arising from a distribution P. **Ψ is a known, smooth functional**.

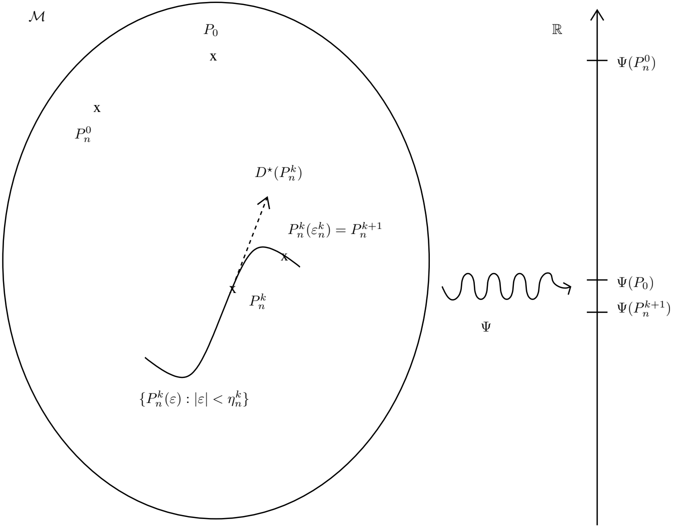
Naive strategy

1. Estimate P using $\hat{P}$
2. Plug-in : $\Psi(\hat{P})$

Our target parameter is $\Psi(P)$, not P !

- $\hat{P}$ aims at balancing bias and variance for the whole distribution
- $\Psi(\hat{P})$ **does not** balance bias and variance for $\Psi(P)$

Targeted minimal loss estimation : illustration



Targeted minimal loss estimation : algorithm

From an initial estimate P_n^0 :

1. Create a model $P_n^0(\varepsilon)$ parametrized by $\varepsilon \in \mathbb{R}$ whose score is the **efficient influence curve** of Ψ at P_n^0
2. Estimate ε using maximum likelihood : ε_n^0
3. Update accordingly : $P_n^1 = P_n^0(\varepsilon_n^0)$

Repeat as many times as necessary : $\forall k \geq 0, P_n^{k+1} = P_n^k(\varepsilon_n^k)$

... hence **final estimate** $P_n^\star$

Statistical properties

P_0 : true distribution of O

Consistency (double robustness)

TMLE is consistent if one of the following conditions holds :

- $\theta(P_n^\star)(0,\cdot)$ consistently estimates true $\theta(P_0)(0,\cdot)$
- $E_{P_n^\star}(X|W)$ and $P_n^\star(X=0|W)$ consistently estimate $E_{P_0}(X|W)$ and $P_0(X=0|W)$

Asymptotic normality

Under the same conditions with higher rates of convergence, TMLE is asymptotically Gaussian.
We can compute asymptotic *p*-values and thus **rank genes**

Note : comparison with a non-parametric estimator

Letting $\theta(P)(X, W) = E_P(Y|X, W)$, we have

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_P \left[\theta(P)(X, W) - \theta(P)(0, W) - \beta X)^2\right]$$

An obvious substitution estimator of $\Psi(P_0)$ is

$$\hat{\psi}_n = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_{P_n} \left\{ \left(\hat{\theta}_n(X, W) - \hat{\theta}_n(0, W) - \beta X\right)^2 \right\},$$

where $\hat{\theta}_n(X, W)$ is the Nadaraya-Watson estimator of $\theta(P_0)(X, W)$.

Such an estimator has non-parametric convergence rates ;
it **may not achieve $\sqrt{n}$-consistency**.

Outline

1. Association between DNA copy number and gene expression
2. Targeted minimal loss estimation of association
3. Results

Simulation strategy

Assumptions :

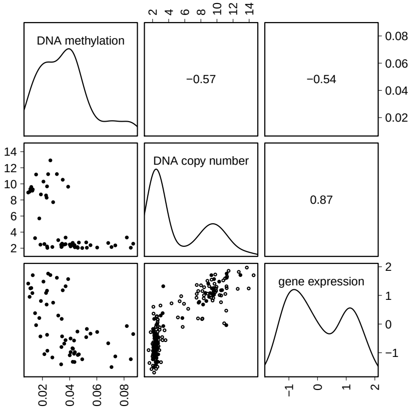
- up to 3 copy number classes : normal regions, and regions of copy number gains and losses
- in normal regions, expression is negatively correlated with methylation
- in regions of copy number alteration, copy number and expression are positively correlated

GBM data used as a baseline for simulation :

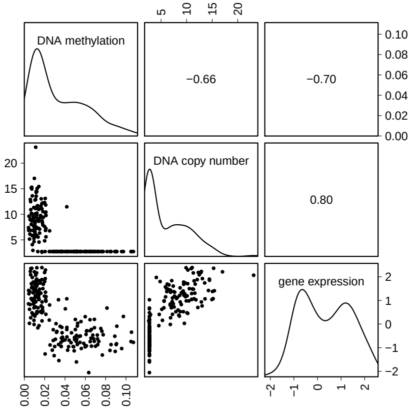
Sample name	Methylation	Copy number	Expression
TCGA-02-0001	0.05	2.72	-0.46
TCGA-02-0003	0.01	9.36	1.25

Simulated data set mimics real data set

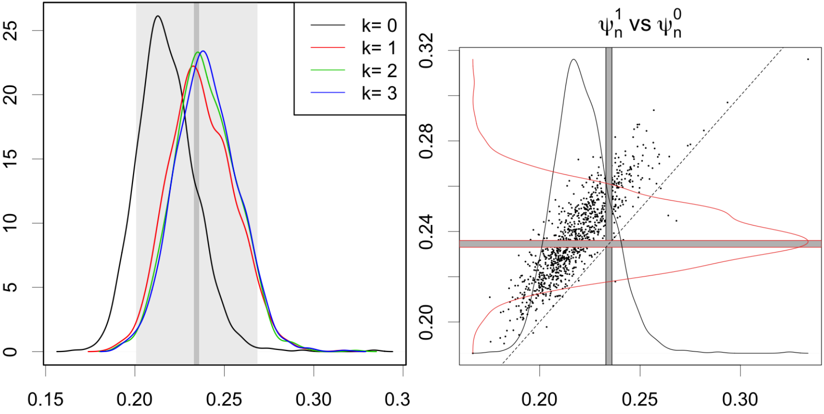
Real data (GBM, n=187)



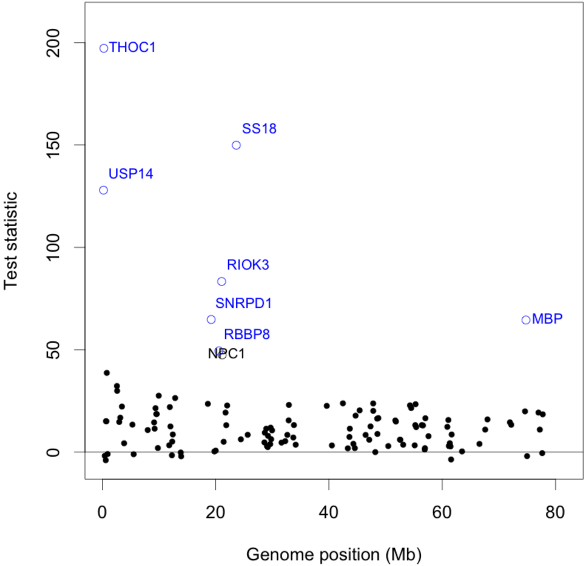
Simulated data (n=200)



Simulated data : TMLE corrects initial estimation



Real data analysis : TCGA OV data set



Thanks

- **Antoine Chambaz**
- Mark van der Laan
- Terry Speed

The Cancer Genome Atlas Research Network