Bioinformatics, 31(18), 2015, 3054–3056
doi: 10.1093/bioinformatics/btv320
Advance Access Publication Date: 21 May 2015
Applications Note

OXFORD

Gene expression

tmle.npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation

Antoine Chambaz[1] and Pierre Neuvial[2,*]

[1]Modal'X, Université Paris Ouest Nanterre, Nanterre, France and [2]Laboratoire de Mathématiques et Modélisation d'Évry, Université d'Évry val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA, Évry, France

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on February 5, 2015; revised on April 13, 2015; accepted on May 18, 2015

Abstract

Summary: We describe the implementation of the method introduced by Chambaz *et al.* in 2012. We also demonstrate its genome-wide application to the integrative search of new regions with strong association between DNA copy number and gene expression accounting for DNA methylation in breast cancers.

Availability and implementation: An open-source `R` package `tmle.npvi` is available from CRAN (http://cran.r-project.org/).

Contact: pierre.neuvial@genopole.cnrs.fr

1 Introduction

Looking for genes whose DNA copy number is 'associated with' their expression level in a cancer study can help pinpoint candidates implied in the disease and enhance our understanding of its molecular bases. Genomic covariates may play an important role in the biological process and should therefore be taken into account. For instance, DNA methylation is known to regulate gene expression. To quantify the association between DNA copy number and expression, accounting for such relevant genomic covariates, Chambaz *et al.* (2012) have crafted a new parameter and built a method to infer it upon the targeted minimum loss-based inference principle (TMLE). Coined by van der Laan and Rubin (2006), TMLE has been applied in a variety of contexts (van der Laan and Rose, 2011). In this note, we describe the implementation of this method and its genome-wide, integrative application.

Considering associations between DNA copy numbers and expression levels in genes is not new (Andrews *et al.*, 2010; Louhimo and Hautaniemi, 2011; Pollack *et al.*, 2002; Sun *et al.*, 2011; van Wieringen and van de Wiel, 2008). In contrast to these earlier contributions, ours does explicitly exploit that DNA copy number measurements feature both a reference level and a continuum of other levels, instead of discretizing them or considering them as purely continuous. Moreover, we naturally handle multidimensional, continuous covariates without discretization. We do not need to assume that they are normally distributed, nor that their true effect on DNA copy number and gene expression is linear.

Methods

2.1 Presentation

We first define the statistical parameter that we wish to target. Let $O = (W, X, Y)$ be a generic observation, where W, X and Y are respectively the covariates (e.g. DNA methylation), DNA copy number and expression of a gene of interest in a randomly picked biological sample. Let x_0 be a reference value for X, corresponding to the normal state of 2 DNA copies. We assume that the probability to observe $X = x_0$ is bounded away from 0 and 1. We assume without loss of generality that $x_0 = 0$. In the absence of additional solid knowledge regarding the law of O, we decide to focus on the following nonparametric variable

importance measure Ψ. It is a mapping from $\mathcal{M}$, the set of all laws compatible with the definition of O, to $\mathbb{R}$ given by

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

Its counterpart neglecting W is a different mapping from $\mathcal{M}$ to $\mathbb{R}$ given by

$$\mathcal{F}(P) = \underset{\beta\in\mathbb{R}}{\arg\min}\ E_P\left\{(Y-\beta X)^2\right\} = \frac{E_P\{XY\}}{E_P\{X^2\}}.$$

The parameter Ψ is pathwise differentiable in the sense that, for every $P \in \mathcal{M}$, there exists a function $\nabla\Psi_P$ of O such that for any bounded $s : O \mapsto s(O)$ and all $|\varepsilon| < ||s||_\infty^{-1}$, if we characterize $P_\varepsilon \in \mathcal{M}$ by setting $dP_\varepsilon/dP = 1 + \varepsilon s$ then $\Psi(P_\varepsilon) = \Psi(P) + \varepsilon E_P\ \{\nabla\Psi_P(O)\ s(O)\} + o(\varepsilon)$. Likewise, $\mathcal{F}$ is pathwise differentiable and associated with $\nabla\mathcal{F}$. In semiparametrics, $\nabla\Psi$ and $\nabla\mathcal{F}$ are known as the efficient influence curves of Ψ and $\mathcal{F}$. Both are known in closed-form. The expression of $\nabla\Psi$ involves finite- and infinite-dimensional features of P, including $\theta_P(X,W) = E_P(Y|X,W)$, $g_P(W) = P(X=0|W)$, $\nu_P(W) = E_P(X|W, X\neq 0)$ and $\sigma_P^2 = E_P\{X^2\}$.

2.2 TMLE Algorithm

Say that we observe n independent random variables $O_1, \ldots, O_n$ drawn from P. They yield the estimator $f_n = \sum_{i=1}^n X_i Y_i / \sum_{i=1}^n X_i^2$ of $\mathcal{F}(P)$. Regarding the estimation of $\Psi(P)$, it consists in

1. initially estimating P with P_n^0 and setting $k = 0$;
2. iteratively (a) constructing a one-dimensional model $\{P_n^k(\varepsilon) : |\varepsilon| < ||s||_\infty^{-1}\}\mathcal{M}$ by setting $dP_n^k(\varepsilon)/dP_n^k = 1 + \varepsilon s$ with $s = \nabla\Psi_{P_n^k}$, (b) computing the corresponding maximum likelihood estimator ε_n^k, and (c) defining $P_n^{k+1} = P_n^k(\varepsilon_n^k)$ and updating $k \leftarrow k+1$ unless a stopping criterion is met;
3. finally forming, by substitution, the TMLE estimator $\psi_n = \Psi(P_n^K)$ where K is the last value of k.

Mild assumptions on P guarantee that $n^{-1}\sum_{i=1}^n \nabla\Psi_{P_n^k}(O_i) \approx 0$. This is the pivotal property upon which the statistical study of ψ_n relies.

2.3 Inference

Under additional assumptions on P and on P_n^K, (f_n, ψ_n) satisfies a central limit theorem (see Chambaz *et al.*, 2012). Furthermore, it is possible to estimate the corresponding asymptotic covariance matrix, hence the asymptotic variances of ψ_n and of $(\psi_n - f_n)$. This paves the way to the testing of '$\Psi(P) = 0$' and of '$\Psi(P) = \mathcal{F}(P)$' against some alternatives.

3 Implementation

Computing $\Psi(P_n^k)$ by Monte-Carlo requires simulating a large number of couples (X, W) under P_n^k and evaluating $\theta_{P_n^k}$ at the simulated values. A naive approach would rely on calls to all $g_{P_n^l}, \nu_{P_n^l}, \theta_{P_n^l}$ for $l = 1, \ldots, k$. This would be computationally expensive due to nestedness. Instead, we construct a law Π_n^k with $g_{\Pi_n^k} = g_{P_n^k}$, $\nu_{\Pi_n^k} = \nu_{P_n^k}$, $\sigma^2_{\Pi_n^k} = \sigma^2_{P_n^k}$ and such that, under Π_n^k, $X \in \{x_1, \ldots, x_m\}$, where m is user-supplied (defaults to 30), $x_1, \ldots, x_m$ are empirical quantiles of X, and $W \in \{W_1, \ldots, W_n\}$. The support of (X, W) consists of $n \times m$ points known beforehand. Storing iteratively the vectors $(g_{P_n^k}(W_j))_{j\leq n}$, $(\nu_{P_n^k}(W_j))_{j\leq n}$ and matrix $(\theta_{P_n^k}(x_i, W_j))_{i\leq m, j\leq n}$ is sufficient. Thus, we reduce the time complexity while keeping the space complexity in $O(n \times m)$ instead of $O(n^2)$. Finally, computing $\Psi(P_n^k)$ requires simulating under Π_n^k and retrieving the values of $\theta_{P_n^k}$ from the stored matrix.

The series of updates is interrupted if the total variation distance $d_{TV}(P_n^k, P_n^{k+1})$, $|\Psi(P_n^k) - \Psi(P_n^{k+1})|$, or $|\sum_{i=1}^n \nabla\Psi_{P_n^k}(O_i)|$ is small. Sensible data-dependent renormalizations and default values quantifying what 'small' means are given for each criterion in `tmle.npvi`.

Running `library(tmle.npvi); example(tmle.npvi);` in `R` demonstrates the use of `tmle.npvi`. The simulation relies on synthetic data generated by perturbing real observations (Chambaz *et al.*, 2012).

4 An application to TCGA data

As an illustration, we study a breast cancer data set from The Cancer Genome Atlas (TCGA) Network (2012). We downloaded DNA methylation (W), DNA copy number (X) and expression (Y) of 11 314 genes for $n = 463$ patients from https://tcga-data.nci.nih.gov/docs/publications/brca_2012. The dimension of W is the number of CpG loci in the gene's promoter region, which can vary from one gene to another. Conveniently, our implementation handles multi-dimensional covariates.

Figure 1 presents the $(-\log_{10})$ p-values of the bilateral tests of '$\Psi(P) = \mathcal{F}(P)$' for the 10 246 genes without missing data against their genomic position. A pattern emerges of regions featuring very small P-values, among which chromosomes 1q, 8, 16q. The pattern is not correlated to the marginal distribution of X, represented in the background. We also compute the partial correlation of X and Y given W for each gene (not shown). No pattern emerges. This suggests that the approach we propose may be useful to identify novel regions worthy of interest.

For this data set, the typical run time of the method for a single gene with the default options of the package is 10 seconds on a standard laptop. The total run time for analyzing all 10 246 genes was 42 CPU hours. Obviously, this analysis can be parallelized very easily, as each gene is treated independently of all the other ones.

5 Conclusion

We have described the implementation of the method introduced by Chambaz *et al.* (2012). We have also demonstrated its genome-wide application to the integrative search of new regions with strong association between DNA copy number and gene expression accounting for DNA methylation in breast cancers. The interest of the parameter Ψ goes beyond the specific genomic context of this note, as it is designed to quantify the effect of any continuous exposure X with a reference value x_0 on any continuous outcome Y accounting for possibly multi-dimensional, relevant covariates W.

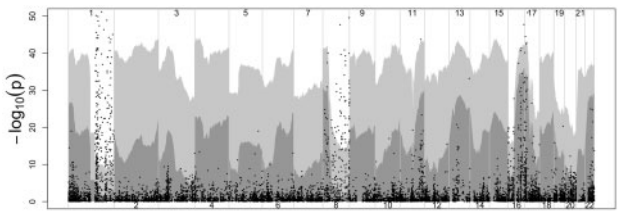

Fig. 1. Each dot corresponds to the genomic position and $(-\log_{10})$ P-value of '$\Psi(P) = \mathcal{F}(P)$' against '$\Psi(P) \neq \mathcal{F}(P)$' for one of the 10 246 genes without missing data. The chromosomes are delimited by vertical gray lines. The background image represents, gene by gene, the proportions of the 463 samples for which $X < 0$ (gray), $X > 0$ (light gray) and $X = 0$ (white)

Acknowledgement

The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov.

Funding

This work was partially supported by the French National Cancer Institute (INCa) *via* Cancéropôle Île-de-France (INCa Grant 2011-1-LABEL- CNRS 3-1).

Conflict of Interest: none declared.

References

Andrews,J. *et al*. (2010) Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number. *PLoS ONE*, **5**, e8665.

Chambaz,A. *et al*. (2012) Estimation of a non-parametric variable importance measure of a continuous exposure. *Electron. J. Statist.*, **6**, 1059–1099.

Louhimo,R. and Hautaniemi,S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, **27**, 887.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA*, **99**, 12963–12968.

Sun,Z. *et al*. (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, **6**, e17490.

The Cancer Genome Atlas (TCGA) Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

van der Laan,M.J. and Rose,S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Verlag, New York.

van der Laan,M.J. and Rubin,D. (2006) Targeted maximum likelihood learning. *Int. J. Biostat.*, **2**, Article 11.

van Wieringen,W.N. and van de Wiel,M.A. (2008) Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **5**, 19–29.