One year postdoc for "A new method for the detection of gene-environment interactions in cancer studies"

Olivier Bouaziz
MAP5, Université de Paris,
45 Rue des Saints-Pères, 75006 Paris
email: olivier.bouaziz@parisdescartes.fr

In recent years, the detection of heterogeneity in epidemiology has attracted increasing interest. One of the main reasons comes from the fact that taking into account heterogeneity in data analysis makes it possible to gain statistical power. In addition, the detection of gene-environment interactions is of major interest in epidemiology because it makes it possible to identify subgroups with high risks in the population. Although several methods have already been proposed for this problem the detection of gene-environment effects remains difficult, in particular because the causal effect is generally not directly observed (as for some treatments for example) and only proxi variables (such as BMI for example) are accessible.

In a recent article by Alarcon *et al.* (2016) it has been shown that conventional methods for detecting gene-environment interactions are ineffective when the exposure variable is not directly observed. In this context, the aim for the postdoc will be to develop a new statistical method for detecting gene-environment effects associated with the occurrence of cancer. The proposed approach will detect groups of individuals characterized by their environmental factors with different cancer risks. It is based on a breakpoint method that has been proposed by Modibo Diabaté (work still ongoing) and is being implemented in an R package. The method is described for logistic models in Alarcon and Nuel (2018) or survival analysis in Bouaziz and Nuel (2018). In these two papers, it has been shown that the breakpoint model approach is particularly powerful compared to conventional gene-environment detection approaches.

The method works as follows: each patient is positioned in a proximity space using multiple covariates (personal, clinical, environmental, etc.). The approach then seeks to exploit the fact that two neighboring patients in this proximity space are more likely to be exposed to latent (and therefore not necessarily observed) common factors. Then a Principal Component Analysis (PCA) is applied to the proximity space and a smoothing curve (called "principal curve", see Hastie and Stuetzle, 1989; Kégl *et al.*, 2000) is applied

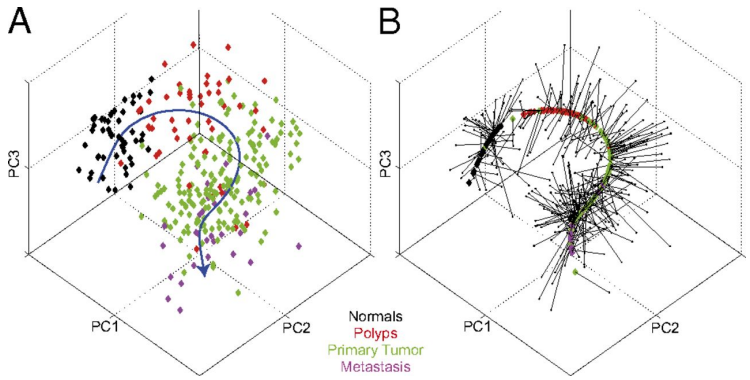

Figure 1: Example of a principal curve based on the three principal components.

to it. This makes it possible to project each individual on this curve and to obtain an order on the individuals. Thus, close individuals on this curve will share similar exposure profiles. An example of a principal curve construction on the three main components of a dataset is shown in Figure 1. Once the main curve is constructed and the individuals projected on it, it is then possible to apply the breakpoint method developed by Modibo Diabaté in either a context of survival data or in a logistic framework.

For example, in survival analysis, the model assumes that, for $i = 1, \ldots, n$ representing an individual:

$$\begin{aligned}\lambda(t|\boldsymbol{X}_i, R_i = k) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t \mid T_i \geq t, \boldsymbol{X}_i, R_i = k)}{\Delta t} \\ &= \lambda_k(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}_k),\end{aligned}$$

where the T_is are the times to cancer occurrence, $\boldsymbol{X}_i$ the SNPs and R_i the unobserved cluster index, with $k = 1, \ldots, K$. In this model, we aim at estimating $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_K, \lambda_1, \ldots, \lambda_K)$ where the $\boldsymbol{\beta_k}$s represent the effects of SNPs in each cluster and the λ_ks represent the effect in the absence of SNPs. By looking at these clusters in terms of environmental variables we will be able to determine complex phenomena of gene-environment interaction. In Bouaziz and Nuel (2018), the estimation method used to estimate the parameters of the model is based on the use of the EM algorithm where computation of $\mathbb{P}(R_i = k|\text{data}; \boldsymbol{\theta})$ is performed using hidden Markov models.

The postdoc student will be funded by the French National League against Cancer (LNCC), it will take place at the MAP5 laboratory and can start from September 1st 2021. A good part of the postdoc will be devoted to applying the method using R on the EPIC (European Prospective Investigation into Cancer and Nutrition) and the UK Biobank data. The postdoc will lead to two applied papers and potentially to one methodological paper. The specific objectives for the postdoc are the following:

- Applying the method to the EPIC and the UK Biobank data. At first, the postdoc will have to get familiar with the method developed by Modibo Diabaté. Then

he/she will investigate how packages using principal curves work and which one is the most appropriate for our problem. Before applying the principal curves approach, a pre-processing step of the data will be probably needed, using a dimension reduction technique (such as PCA) and a dedicated method to deal with categorical variables. Of note, the UK Biobank data are of high dimension (40 Go) and the post-doc will have to use specific tools to deal with such voluminous data.

- Investigating other approaches not based on principale curves. For example, the problem could be tackled using *minimum spanning trees* (see for instance Grygorash *et al.*, 2006).

The EPIC cohort is a multi-centric European study that includes more than $500,000$ individuals, recruited in the 1990, in 10 European countries (see Riboli *et al.*, 2002). This dataset contains $7,491$ genotyped women ($3,831$ cases of breast cancer and $3,623$ controls) with different clinical and environmental information on patients: for example, socio-economic status, height, weight, BMI, smoking status, alcohol consumption, eating habits (obtained from a questionnaire), the status of menopause, the use of hormonal treatment (contraception or for menopause) etc.

The UK Biobank data come from a prospective cohort on $488,377$ British individuals, all genotyped. Recruitment took place from 2006 to 2010, for individuals aged 40 to 69 years. In May 2018, $79,000$ cases of cancer were diagnosed. The main ones are melanoma, breast cancer, uterine cancer, prostate cancer and colon cancer. These data also contain several environmental and clinical informations about these individuals: lifestyle, biological measurements, biomarkers in the blood and urine, also images of the brain and heart as well as repeated measures of physical activity.

Contact	`olivier.bouaziz@parisdescartes.fr`
Salary	3,195 € per month (before tax)
Dates	September 2021 - August 2022
Postdoc place	MAP5, Université de Paris 45 Rue des Saints Pères, Paris 75006
Required skills	Algorithmic methods such as EM A good experience with the R software
Publications objectives	Two applied papers

References

Alarcon, Flora and Nuel, Gregory. (2018). Detecting latent exposure in genome-wide association studies using a breakpoint model for logistic regression. *Statistical Methods in Medical Research*.

Alarcon, Flora, Perduca, Vittorio and Nuel, Gregory. (2016). Is it possible to detect g× e interactions in gwas when causal exposure is unobserved? *Journal of Epidemiological Research* **2**(1), 109–117.

Bouaziz, Olivier and Nuel, Grégory. (2018). A change-point model for detecting heterogeneity in ordered survival responses. *Statistical methods in medical research* **27**(12), 3595–3611.

Grygorash, Oleksandr, Zhou, Yan and Jorgensen, Zach. (2006). Minimum spanning tree based clustering algorithms. In: *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*. IEEE. pp. 73–81.

Hastie, Trevor and Stuetzle, Werner. (1989). Principal curves. *Journal of the American Statistical Association* **84**(406), 502–516.

Kégl, Balázs, Krzyzak, Adam, Linder, Tamás and Zeger, Kenneth. (2000). Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence* **22**(3), 281–297.

Riboli, E, Hunt, KJ, Slimani, N, Ferrari, P, Norat, T, Fahey, M, Charrondiere, UR, Hemon, B, Casagrande, C, Vignat, J *et al.* (2002). European prospective investigation into cancer and nutrition (epic): study populations and data collection. *Public health nutrition* **5**(6b), 1113–1124.