Université Lyon 1

Habilitation à diriger des recherches

présentée par

Franck Picard

- A statistical tour of genomic data -

Soutenance le jeudi 4 décembre 2014, à Lyon,
devant le jury composé de

I. Eckley	Université de Lancaster	rapporteur
A.-L. Fougères	Université Lyon 1	présidente
W. Huber	EMBL Heidelberg	rapporteur
C. Matias	CNRS, Paris	examinatrice
J.-M. Poggi	Université Paris 5	examinateur
M. Rattray	Université de Manchester	examinateur
J.-P. Vert	Mines ParisTech	rapporteur

Introductory acknowledgements

I (scientifically) grew up in Paris, being (well?) educated by a circle of brilliant statisticians who were working on sequence analysis, motifs exceptionality, and then microarrays. At that time, computers were invading biology wet-lab benches, and our colleagues were getting nervous about not being able to keep and analyze their data on their own. I worked on the modelling of copy-number data that had just come out. We proposed a method to detect chromosomal abnormalities using change-points in Gaussian series that were 2,500 point long for the whole genome. The first part of this presentation will end up with projects involving high-density microarray data that are 500,000 long, and that are available for thousands of patients, illustrating the flood of genomic data we have faced the past years. I defended my PhD in 2005 (date from which this manuscript begins), and then started a postdoc, in the Statistics & Genome Lab of Evry (I would like to thank here B. Prum for welcoming me during these two years). We kept on working on copy-number analysis and generalized our approach to the multisample case, as more and more studies were going clinical. Multiple change-point models were still at the core of these works, and surprisingly, our approaches were also of interest for other scientists working on climate change detection, as I will explain later. During my postdoc (2005-2007), I also started working on a new project on biological networks. "Complexity" and "networks" were undoubtedly among the fanciest (scientific) keywords at that time, and I must confess, I succumbed. I learned many things working on random graphs, and I am very grateful to S. Robin & J.-J. Daudin, my former PhD advisors for inviting me to join them on the MixNet project, which will be the subject of the second part of this manuscript. This part will be complemented by recent projects on networks that involve my eternal office-mate, V. Miele, and also V. Viallon and S. Lambert-Lacroix, who introduced me to penalized estimation and to the "lasso" (that is a technical term, despite what non-initiated ones might think).

In 2007 I had the unique chance to get a permanent position from the Centre National de la Recherche Scientifique (CNRS), and I went to the laboratoire de biométrie et biologie évolutive (LBBE) in Lyon were I am currently doing my research. I would like to thank the group leaders who welcomed me, M.-F. Sagot and M. Gouy, and most importantly D. Mouchiroud as the leader of this incredible lab, for her constant support and encouragements. Among the infinite number of exciting questions that were possible to tackle, I chose what I thought would be an "easy" one, which is the study of the inter-individual variability of copy-number data. Then I realized that multiple change-point models would not be the most appropriate to tackle this issue: they are very

efficient in detecting breakpoints but the associated models are highly irregular and do not handle the whole signal as a series on which such statistical questions could be easily addressed. Thanks to the proximity with my shadow advisor[1] A. Antoniadis, I got involved in functional data analysis, and I started to consider copy-number studies as high-dimensional longitudinal analysis, with the genome being the observation grid. We started a collaboration with S. Lambert-Lacroix on functional mixed models and curve clustering, and we co-supervised our first graduate student on the subject (2009-2013). The question of functional inter-individual variations has been much more complex than I thought, and Part 1 ends with current projects on the subject.

As the LBBE is a real Biology lab (contrary to the other labs I worked in), I quickly got involved in next-generation sequencing data analysis which has become inevitable nowadays. Thanks to the request of my dearest colleagues L. Duret and M.-N. Prioleau we started working on the "replication" project. This sequence will be the subject of Part 3, and I must say that this project has been one of the most inspiring and exciting to work on. I quickly realized that clever computational biologists like L. Duret usually do not need complicated statistics to answer their questions. It was cruel to me, but I learnt to focus my energy on questions that could not be answered by simple statistics, such as the differential analysis of peak-like data or the modelling of interacting epigenetic marks along the genome. The last chapters of this manuscript will be dedicated to these new projects. They deal with the modelling of NGS data by either multivariate analysis, point processes or functional models for counts, with A.-L. Fougères, G. Marais, P. Reynaud-Bouret, V. Rivoirard and E. Roquain as my new research-mates.

This manuscript is the result of all those wanders in the field of statistical modelling. Everything that will be written in this document can be hold against me, since I have this incredible chance of conducting my research with no constraint. It is just a snapshot on the 2005-2014 period which resulted in the creation of a new group "Statistics in high dimension for genomics" with the adventurous L. Jacob who recently joined the lab, and whom I thank for having defended this project with me. The "snapshot" also involves that the presented research projects are still ongoing and I am still learning things and (I hope) still making progress on the long path of multidisciplinary research and applied statistics. I am very grateful to the members of my jury who accepted to review and comment this work. It is the result of all these collaborations I mentioned above, scientific and personal encounters, and of the involvement of the students I worked with (I. Bardet, G. Durif, M. Giacofci, S. Ivanoff, F. Mifsud, L. Modolo, A. Muyle) whom I would like to thank for their contributions and patience, and for letting me teach them some things.

$\star$

[1] I have had other shadow advisors whom I would like to thank also, S. Robin of course, A. Bar-Hen, and C. Gautier.

Contents

Part I

From change-points to functional models for copy-number data analysis

Chapter 1

Joint segmentation for copy number profiles

Cancer bioinformatics has received enormous attention in the past ten years, and studying the structure of cancer genomes has been a productive research direction. Linking chromosomal aberrations and cancer is far from new: oncogenes and tumour suppressor genes are known to be frequently amplified or deleted, leading to DNA copy imbalances. In the late 1990s the microarray CGH technology has allowed the investigation of copy number changes at the genome scale in one experiment [112]. To date statistical efforts have mainly focused on the recovery of the segmental structure by segmentation and the unknown discrete copy number values from the raw data with a "calling" step, at the single sample level. During my PhD (2002-2005) we worked on the application of change-point models to the analysis of array CGH data. Briefly, we proposed to model the array CGH signal by a Gaussian process organized at different positions along the genome, such that when a deletion or an amplification is present, the signal is supposed to jump abruptly from a segment to another. In this model, each segment represents a portion on the genome on which the model has detected no change, but the number, position and level values of segments are unknown. We first proposed the application of a Dynamic Programming algorithm to the search of breaks position, and to compare different model selection criteria to select the number of segments [93]. In a second step, we enriched the model by supposing that all segments corresponding to a similar copy number, or "state" (such as amplified once, or deleted twice), were sharing the same level on average, which resulted in what is called a "calling" step, since this model calls the segments into a finite number of states [94]. More than 30 methods have been published on the subject, and reviews concerning array CGH data analysis are now available [89, 121]. The proposed statistical frameworks range from breakpoint detection [88, 93, 97], to smoothing [63, 15] and hidden Markov models [76, 113]. Existing methods have already been compared, and one consistent result is that segmentation methods perform best for the analysis of array CGH data [72, 126].

As the array CGH technology has become more popular, biologists now face the problem of analysing profiles associated with several patients simultaneously. Even though breakpoint detec-

tion can easily be achieved at the single patient level, new modelling and computational challenges arise at the multi-patient level. Many questions need to be addressed such as the joint analysis of chromosomal alterations for a set of profiles [95, 120], the detection of recurrent alterations within this set [102, 107, 105, 101] and the clustering of patients according to their CGH profile [124, 74]. In this chapter, we present our main results [FP10, FP9, FP12, 121], concerning joint analysis issue that we addressed in three main points.

(i) The efficiency of the segmentation approach is based on the use of Dynamic Programming (DP) [93]. However, a drawback of this algorithm is its complexity in $\mathcal{O}(Kn^2)$, with n being the length of the signal and K the number of segments. Consequently, segmenting multiple profiles raises a major computational issue. In Picard et al. (2011) [FP10] we propose a trick to use DP on multiple profiles, whose complexity is reduced thanks to a second layer of DP. In a further work [FP12] we investigated other computational issues, by using parallel programming to deal with the multisample issue, and linearized dynamic programming to analyze high density array signals.

(ii) The calling step consists in the assignment of copy number values to probes to determine which probes are in the "deleted", "amplified" or "normal" state for instance. One limitation of pure segmentation methods is that they do not give information about the copy number values. 'Merging' steps have been proposed to cluster segments into groups of homogeneous copy number values. These strategies are based on statistical tests [126] or on clustering [122, 94]. This downstream step was shown to be of 'paramount importance' when using segmentation for array CGH [126]. But the merging step only constitutes a second-stage procedure, whereas segmentation can also learn from the calling step in a unified model to gain in power in the detection of breaks that correspond to changes in copy number values [94]. Considering multiple profiles gives the opportunity to perform global calling for the whole dataset, since the average signal associated to each copy number change is likely to be common across profiles.

(iii) By normalization we refer to a step that removes or accounts for possible artefacts of the aCGH technology. Performing a joint analysis constitutes an opportunity to correct for this bias which is shared by all signals measured on the same type of arrays. The origins of this bias is unclear, but a consensus exists on its link with GC content [26, 95]. It can be viewed as a heterogeneity between hybridization intensities that would be observed even when dealing with DNA without aberration. When considering one profile only, correcting this bias is dangerous, since there exists an aliasing between copy number changes and wavy patterns. Consequently, this correction may be suitable for single profiles, but not for cancer profiles for which aberrations could be smoothed as well. When considering multiple profiles, a calibration set can be used to estimate this wave bias and to remove it from the data [120]. However, when no calibration set is available this bias can be modelled for by adding a correction term within the segmentation model.

In Picard et al. (2011) [FP10] we proposed a unified statistical framework to correct for those effects. The model we proposed can be viewed as a generalization of existing strategies [95, 120] by the integration of the calling method within the segmentation model. We also proposed a general normalization strategy that may include probe-specific bias correction or account for any exogenous covariate such as GC-content.

1.1 Multiple change points for joint segmentation

Notations for segmentation models In the following chapter, $Y_i(t)$ will denote the log-ratio measured at position t for patient i, each position stands for a probe on the array and thus we consider that t is discrete with $t = 1, \ldots, n_i$. $\mathbf{Y}_i$ will denote the single profile for patient $i = 1, \ldots I$ of size n_i. Then we suppose that the mean of profile $\mathbf{Y}_i$ is subject to $k_i - 1$ abrupt changes at breakpoints $\{t_k^i\}$ (with convention $t_0^i = 0$ and $t_{k_i}^i = n_i$) and is constant between two breakpoints within the interval $]t_{k-1}^i, t_k^i]$. In the following we denote by $K = \sum_i^I k_i$ the total number of segments across profiles, and $N = \sum_i^I n_i$ the total number of observations. Thus we consider the following model:

$$\forall t \in]t_{k-1}^i, t_k^i], \;\; Y_i(t) = \mu_{ik} + E_i(t),$$

where $E_i(t)$ stands for a Gaussian white noise with variance σ^2. In order to use the matricial formulation of linear models, we introduce notations $\text{Block}_{i=1}^I [A_i]$ the diagonal block matrix with A_i the ith diagonal block, and $\mathbb{1}_n$ the column vector filled with n ones. Then we consider the $[N \times K]$-incidence matrix of breakpoints denoted by $\mathbf{T} = \text{Block}_{i=1}^I [\mathbf{T}_i]$ with $\mathbf{T}_i = \text{Block}_{k=1}^{k_i} \left[\mathbb{1}_{n_k^i}\right]$ of size $[n_i \times k_i]$ being the incidence matrix of breakpoints in profile i, and with $n_k^i = t_k^i - t_{k-1}^i + 1$ being the length of segment k for profile i. We also introduce notation $\boldsymbol{\mu} = [\mu_{ik}]$ (of size $[K \times 1]$). Then our model is $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{E}$, where $\mathbf{Y}$ ($[N \times 1]$) stands for the observed data, and where $\mathbf{E}$ is centered Gaussian vector with diagonal covariance matrix $\sigma^2\mathbf{I}$. Further work (posterior to this project) generalized this framework to Poisson and Negative Binomial models [34].

Using Dynamic Programming for joint segmentation The main challenge of the multiple samples strategy is to find the best global segmentation according to the maximum likelihood criterion [93]. For this purpose Dynamic Programming is the computational key ingredient. However, the question of computational efficiency is asked when considering I joint profiles because DP complexity is quadratic with the size of the data. We propose a computational trick to reduce this burden when segmenting multiple profiles. The minimization problem reduces to finding $\{\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}\}$ such that:

$$\{\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}\} = \underset{\{\mathbf{T}, \boldsymbol{\mu}\}}{\text{argmin}}\, RSS_K(\mathbf{T}, \boldsymbol{\mu}),$$

with $RSS_K(\mathbf{T}, \boldsymbol{\mu})$ the residual sum of squares of a segmentation model with K segments such that:

$$\begin{aligned} RSS_K(\mathbf{T}, \boldsymbol{\mu}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}\|^2 &= \sum_{i=1}^I RSS_{k_i}^i(\mathbf{T}_i, \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^I \sum_{k=1}^{k_i} \sum_{t \in]t_{k-1}^i, t_k^i]} (Y_i(t) - \mu_{ki})^2. \end{aligned}$$

When dealing with multiple profiles this minimization must be done under an additional constraint which is : $\sum_i k_i = K$. The computational trick we propose is based on the following breakdown:

$$\min_{\{\mathbf{T},\boldsymbol{\mu}\}} RSS_K(\mathbf{T},\boldsymbol{\mu}) = \min_{k_1+\ldots+k_I=K} \left\{ \sum_{i=1}^{I} \min_{\mathbf{T}_i,\boldsymbol{\mu}_i} RSS^i_{k_i}(\mathbf{T}_i,\boldsymbol{\mu}_i) \right\}.$$

Since the RSS is additive according to the patients and to the number of segments, we propose a double stage Dynamic Programming to solve this optimization problem. Let us introduce a new notation to explain the core of the algorithm, and denote by $\widehat{\mathbf{T}}^i(k_i)$ the set of optimal breaks with k_i segments for profile i.

Stage-1 The first step consists in finding all optimal breakpoints for each profile for $k_i = 1, \ldots, k_{\max}$ segments: $\widehat{\mathbf{T}}^i(k_i)$. This step is done using classical Dynamic Programming.

Stage-2 The second step consists in the allocation of the optimal number of segments to each profile. We aim at determining the optimal sequence $\widehat{k}_1, \ldots, \widehat{k}_I$, such that $\sum_i \widehat{k}_i = K$ for a given K. We denote by $RSS_K(\widehat{\mathbf{T}}^1(k_1), \ldots, \widehat{\mathbf{T}}^I(k_I))$ the total sum of squares for a model with K segments spread over I profiles, each having k_i segments. This step is solved using recursion:

$$\begin{aligned}
&\forall i \in [1:I], \\
\{\widehat{k}_1, \ldots, \widehat{k}_i\} &= \operatorname*{argmin}_{k_1+\ldots+k_i=K} RSS_K\left(\widehat{\mathbf{T}}^1(k_1), \ldots, \widehat{\mathbf{T}}^i(k_i)\right) \\
&= \operatorname*{argmin}_{k'+k''=K} \left\{ \operatorname*{RSS}_{\substack{k'\\ k'_1+\ldots+k'_i=k'}} \left(\widehat{\mathbf{T}}^1(k'_1), \ldots, \widehat{\mathbf{T}}^{i-1}(k'_{i-1})\right) + RSS^i_{k''}\left(\widehat{\mathbf{T}}^i(k'')\right) \right\}.
\end{aligned}$$

At the end of this double-stage Dynamic Programming, we have the optimal breakpoint positions for the optimal number of segments in each profile.

$$\widehat{\mathbf{T}}(K) = \left\{ \widehat{\mathbf{T}}^1(\widehat{k}_1), \ldots, \widehat{\mathbf{T}}^I(\widehat{k}_I) \right\}.$$

Complexity in time. The first stage corresponds to the segmentation of individual profiles into $k_{\max}$ segments each, with complexity $\mathcal{O}(n^2 I k_{\max})$. The complexity of the second stage is $\mathcal{O}((I k_{\max})^2 \times I)$ which makes the overall complexity of order $\mathcal{O}(I n^2 k_{\max} + k^2_{\max} I^3)$. Assuming that the major term is n (which is consistent with the ever increasing density of aCGH) the second term remain negligible and the complexity becomes $\mathcal{O}(I k_{\max} n^2)$. This complexity should be compared with the one of Dynamic Programming applied to the complete dataset with $N = In$ points into $K_{\max} = I k_{\max}$ segments, that is $\mathcal{O}(K_{\max} N^2) = \mathcal{O}(I^3 k_{\max} n^2)$. The 2-stage DP therefore reduces the complexity by a factor I^2.

Model selection. In practice the total number of segments K as well as the number of segments for each profile $\{k_1, \ldots, k_I\}$ are unknown. Based on our computational strategy that is based on a 2-stage Dynamic Programming we choose to select K by model selection, and stage 2 of the DP is used to determine the best combination of numbers of segments among profiles. As discussed by many authors, segmentation models raise a difficult issue in terms of model selection. The question has been studied in the single profile context [93, 127], and the work had to be done for joint segmentation. Zhang and Siegmund (2007) [127] developed a very powerful framework for model selection in change-point models by considering a continuous time version of the model that solves the irregularity issue associated with the discrete nature of the break points. Due to the excellent performance of this strategy, we proposed a generalization of their framework to the joint segmentation setting [FP10].

1.2 Joint normalization and calling

Integrative normalization. The interest in considering many profiles is that if a systematic bias is observed for every profile, considering the joint analysis can help in its correction. In the following, we will denote by $b(t)$ this bias at position t (for probe $t = 1, \ldots, n$), and we suppose that it is present and common accross profiles, such that the segmentation model becomes:

$$\forall t \in]t^i_{k-1}, t^i_k], \;\; Y_i(t) = \mu_{ik} + b(t) + E_i(t).$$

Then we use a unified matricial formulation such that $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\mathbf{B} + \mathbf{E}$, with $\mathbf{X} = (\mathbf{I}_n, \ldots, \mathbf{I}_n)^T$ (I blocks) which spreads the common fixed effect $\mathbf{B} = (b(t_1), \ldots, b(t_n))^T$ over the I patients. We proposed different strategies to model this bias [FP10].

A first model consists in considering a model where $b(t) = \beta_t$ stands for a probe effect, or a reference hybridization intensity as already proposed [95]. This modelling would consider that if a probe shows a systematic bias, it would be detected by this effect. This is the simplest correction that could be made on the data. We get the following model:

$$\forall t \in]t^i_{k-1}, t^i_k], \;\; Y_i(t) = \mu_{ik} + \beta_t + E_i(t).$$

This new part of the model can be estimated using an iterative least-squares algorithm such that $\widehat{\beta}_t^{[h+1]} = \sum_{i=1}^{I} (Y_i(t) - \widehat{\mu}_{ik}^{[h]})/I$, and breaks are updated with dynamic programming on $\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}^{[h+1]}$. We also proposed to introduce some control on the regularity of the bias function using splines and wavelets [FP10].

Multivariate Calling. The principle of the segmentation/clustering model we developed [94] is to integrate in the segmentation model that different segments with the same underlying copy number should share the same mean signal on average. Suppose we observe P distinct states, then the mean of each segment should lie in a restricted set $\mathbf{m} = \{m_1, \ldots, m_P\}$, where m_1 is the inferred mean from all segments in cluster 1 that share the same copy number. This is why

we introduce random indicator variables $[C_{kp}^i]_{(kp)}$ that give the state of each individual's segments, and the joint segmentation/clustering model is such that:

$$\{C_{kp}^i = 1\},\ \forall t \in]t_{k-1}^i, t_k^i],\ \ Y_i(t) = m_p + b(t) + E_i(t).$$

We use a modified version of the EM algorithm to estimate the maximum likelihood parameters adapted from [94]. Briefly, the E-step is used to assess $\{\tau_p^{ik}\}$ the $posterior$ probabilities of membership to clusters for each segment using a classical Bayes rule, and the mean corresponding to each state is estimated in the M-step such that:

$$m_p^{[h+1]} \quad = \quad \frac{\sum_{i=1}^{I}\sum_{k=1}^{k_i}\tau_p^{ik[h]}\sum_{t\in]t_{k-1}^i,t_k^i]}(Y_i(t) - b^{[h]}(t))}{\sum_{i=1}^{I}\sum_{k=1}^{k_i} n_k^i \tau_p^{ik[h]}}.$$

Then the question is to update $b(t)$. If considering the probe-effect model (β_t), the maximum likelihood estimator is $\beta_t^{[h+1]} = \sum_{i,p}\tau_p^{ik(t)[h]}(Y_i(t) - m_p^{[h+1]})/I$ which corresponds to the weighted residuals at position t after segmentation/clustering.

Application to the study of climate change. Multiple change-point models have long been used to detect variations in climate series [27]. Interestingly, the approach of joint segmentation was also of interest for climate scientists that are interested in the detection of changes in grapes harvest dates across several French regions. The aim was to detect changes in terms of agricultural practices that may affect the harvest date in specific stations. As harvest are made earlier in hot years, we typically wanted to distinguish station-specific variations from those due to variations of the climate. For this purpose we developed a joint segmentation model with random effects to model the common climatic effect on temperature variations [FP9, FP3].

1.3 Package and performance

All models that were developed for single or multiple sample analysis were implemented in the `cghseg` `R/C++` package . From a performance point of view, DP-based segmentation methods have always shown the best performance among other methods [72] even in the most recent comparisons [60]. However, the computational burden associated with the quadratic complexity (in signal length) has limited the use of DP on high density arrays (with $\sim 500,000$ points), the issue being even more problematic for multiple samples. Fortunately, two independent studies proposed a linearization of the dynamic programming algorithm with a pruning strategy [99, 70], and the version of [99] has been included in the `cghseg` package. Moreover, in a recent publication [FP12] we proposed a linearization of the dynamic programming algorithm in the case of segmentation-calling, with a modified cost function derived from the k-means algorithm. We also proposed to deal with the joint segmentation issue by providing a parallel version of existing algorithms implemented in the `cghseg` package. With the growing availability of multicore computers (from laptops to many-core servers), it has become essential to provide software that use every available

ressource. R has been a tremendous plateform for package distribution, this is why we worked on a new version of the new cghseg package, a *next generation package* that is adaptive to available computing power. The performance of cghseg are impressive : segmenting 1,000 profiles of length 100,000 can now be done in few hours, which was impossible before. This makes segmentation models a new exact investigation method that can be used in routine for exploratory as well as deep analysis.

Chapter 2

When Genomics goes functional

Functional data analysis has gained increased attention in the past years, with a particular focus on mass-spectrometry data when applied to high-throughput biology. In this context, the aim is to characterize the protein content of biological samples by separating compounds according to their mass to charge ratio (m/z), and mass-spectromety has become standard to improve proteomic profiling of diseases as well as clinical diagnosis. Dedicated methods have been developed to analyze such data for differential analysis, supervised classification and clustering [59]. One central element is the modelling of the inter-individual variability by using functional random effects, since subject-specific fluctuations are known to be the largest source of variability in mass-spec data [41]. In [FP2] we focused on the non supervised task which consists in finding groups of individuals whose proteomic landscape is similar. Surprisingly the clustering task received less attention, and is mainly based on hierarchical clustering on the set of peaks detected across spectra [16, 82]. However such method is known to depend heavily on the peak detection method and has the strong dis-advantage of neglecting the inter-individual variability, whereas this information should be central for subgroup discovery. Thus our main focus was to model and cluster curves of this type in a functional mixed model framework.

However, our first motivation was not to analyze mass-spectrometry data, but copy number profiles! Clustering patients based on their CGH profiles is very promising and has been successfully used to identify molecular subtypes of cancer [44, 124]. However, clustering CGH profiles based on segmentation has the same drawbacks as clustering mass spectra based on detected peaks: results highly depend on the segmentation methods. Moreover the inter-individual variability has never been investigated in this type of data, whereas it is likely to represent an important part of the variability of the data especially for cancer profiles. This point is of particular interest in this chapter. The inter-individual variability of copy-number profiles has been often mentioned but never completely assessed. For instance, when considering cancer data, it has been long known that contamination with normal tissue could modify both the level of the microarray signal (decreasing the log ratio in the case of normal contamination at amplified loci), as well as the location of the break along the genome. Moreover, the field of human genetics has been investigating the impact

of polymorphic copy number variations along the genome, and estimates suggest that 10% of the human genome is subject to polymorphic copy number variation. In other words, this means that when the purpose is to determine chromosomal aberrations that are specific of a given phenotype (like cancer subtype, [32]), there exists a potentially important inter-individual variability in the data that has never been accounted for.

It appears that the framework of change-point detection was not necessarily the most appropriate one to tackle this question. This framework is very powerful to detect the changes, but the discrete modelling of time-points makes the model highly irregular. This is why I chose to focus on functional models using wavelets to consider the observed profiles as functions belonging to some functional space. When dealing with curve clustering in the presence of individual variability, a pioneering work was based on a spline decomposition of the signal [65], which reduces to a linear mixed effect model on which clustering and low-dimensional representation can be performed. However splines show two main drawbacks: i) they are inappropriate when dealing with functions that show peaks and irregularities, ii) they require heavy computational efforts and so are not adapted to high dimensional data. On the contrary, wavelet representations appear to be a natural framework to consider such irregularities through the sequence space of (usually sparse) Besov representation. Recent works have attempted to estimate and infer the functional mixed effects framework based on a wavelet decomposition approach. A fully Bayesian version has been proposed [83], with non-parametric estimates of fixed and random effects as well as between and within-curve covariance matrix estimates to accomodate a wide variety of correlation structures. Estimation and inference have also been preliminary investigated in the frequentist framework [7].

From 2010 to 2013 we co-supervized a graduate student with Sophie Lambert-Lacroix (Pr., Grenoble University) on this subject. We published a first article [FP2] on curve clustering with functional random effects, with application to array CGH data analysis. Following this work, we investigated statistical questions related to dimension reduction in wavelet-based functional models, which will be the purpose of the last section that presents on-going work on the subject.

2.1 Functional modelling of copy number profiles

We consider that we observe I curves $Y_i(t)$ over n equally spaced time points $\mathbf{t} = (t_1, \ldots, t_n)$ in $[0, 1]$, with $n = 2^J$ for some integer J and we model these data by the linear functional model of the form:

$$Y_i(t) = \mu_i(t) + E_i(t), \qquad E_i(t) \sim \mathcal{N}(0, \sigma_E^2). \tag{2.1}$$

In the following we will use notation $\mathbf{Y}_i(\mathbf{t}) = [Y_i(t_1), \ldots, Y_i(t_n)]$. In the preceeding chapter, we focused on the detection of abrupt changes in μ_i using change-point models. Here our aim is not to identify breakpoints precisely, but we rather focus on the global modelling of μ_i as a function that is organized along the genome. If only one group of individuals is observed, then $\forall i \in \{1, \ldots, I\}$, $\mu_i = \mu$, which makes μ the common copy number profile among the samples. In the functional clustering setting we suppose that individuals are spread among L unknown clusters of prior size π_ℓ, $\ell = 1, \ldots, L$, and we denote by $\zeta_{i\ell}$ the indicator variable that equals 1 if the ith individual is in

the ℓth group. Given $\{\zeta_{i\ell} = 1\}$, model (2.1) becomes

$$Y_i(t) = \mu_\ell(t) + E_i(t), \tag{2.2}$$

where $\mu_\ell(t)$ is the principal functional fixed effect that characterizes cluster ℓ. To handle subject-specific random deviations from the cluster average curve we introduce random functions $U_i(t)$ that are modelled as centered Gaussian processes not necessarily stationary but independent from $E_i(t)$. Then given $\{\zeta_{i\ell} = 1\}$, model 2.2 becomes

$$Y_i(t) = \mu_\ell(t) + U_i(t) + E_i(t), \quad U_i(t) \sim \mathcal{N}(0, K_\ell(\bullet, t)). \tag{2.3}$$

Once defined in the functional domain, the classical approach is to convert the original infinite-dimensional clustering problem into a finite-dimensional problem using a functional basis representation of the model. Briefly, we are working with an orthonormal wavelet basis

$$\{\phi_{j_0k}(t), k = 0, 1, \ldots 2^{j_0} - 1; \psi_{jk}(t), j \geq j_0, k = 0, \ldots, 2^j - 1\}$$

generated from a father wavelet ϕ and a mother wavelet ψ of regularity r, ($r \geq 0$). In this basis the response curve $Y_i(t)$ has the following decomposition:

$$Y_i(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{i,j_0k}\phi_{j_0k}(t) + \sum_{j\geq j_0}\sum_{k=0}^{2^j-1} d^*_{i,jk}\psi_{jk}(t).$$

In practice we use the Discrete Wavelet Transform (DWT) which can be performed thanks to Mallat's fast algorithm with $\mathcal{O}(n)$ operations only. We denote by $\mathbf{W}$ the $(n \times n)$-matrix containing filter coefficients of the chosen wavelet basis (full details on the discrete wavelet transform can be found here [84]). The resulting scaling and wavelet coefficients $\mathbf{c}_i = [c_{i,j_0k}]_{(k)}$ and $\mathbf{d}_i = [d_{i,jk}]_{(jk)}$ of the individual curves are empirical coefficients (with $j_0 = 0$). They are related to their theoretical continuous counterparts c^*_{i,j_0k} and $d^*_{i,jk}$ by: $c_{i,j_0k} \approx \sqrt{n}c^*_{i,j_0k}$ and $d_{i,jk} \approx \sqrt{n}d^*_{i,jk}$. When applying the DWT to model (2.3) we have

$$\mathbf{W}\mathbf{Y}_i(\mathbf{t}) = \mathbf{W}\boldsymbol{\mu}_\ell(\mathbf{t}) + \mathbf{W}\mathbf{U}_i(\mathbf{t}) + \mathbf{W}\mathbf{E}_i,$$

which reduces to a linear mixed-effect model in the coefficient domains such that

$$\begin{aligned} \mathbf{c}_i &= \boldsymbol{\alpha}_\ell + \boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{d}_i &= \boldsymbol{\beta}_\ell + \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i. \end{aligned}$$

$(\boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell)$ stand for the scaling and wavelet coefficients of the fixed average curve $\boldsymbol{\mu}_\ell(\mathbf{t})$, and $(\boldsymbol{\nu}_i, \boldsymbol{\theta}_i)$ are the scaling and wavelet random coefficients of Gaussian process $\mathbf{U}_i(\mathbf{t})$ such that

$$\begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{G}_{\boldsymbol{\nu}} & 0 \\ 0 & \mathbf{G}_{\boldsymbol{\theta}} \end{bmatrix}\right),$$

and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}) \perp (\boldsymbol{\nu}, \boldsymbol{\theta})'$, with $\sigma_\varepsilon^2 = \sigma_E^2/n$.

The complete estimation framework of the functional clustering model relies on the EM algorithm and is fully detailed in Giacofci et al. (2013) [FP2]. By using the Maximum a posteriori rule, we provide a prediction of the labels $\widehat{\zeta}_{i\ell}$ which allows us to cluster individuals based on their wavelet coefficients. We also studied model selection issues to select the number of clusters, and performed a simulation study to assess the clustering performance of our method. We developed the `curvclust` package for curve clustering with (or without) functional random effects (available on the `CRAN`).

2.2 Modelling inter-individual functional variations

A very critical point in the functional modelling of random effects is the specification of the distribution of their wavelet coefficients. Indeed, thanks to the whitening properties of the wavelet transform [43], even if U_i is non stationary and dependent, the covariance of its wavelets coefficients $\mathbb{V}(\boldsymbol{\theta}_i) = \mathbf{G}_\theta$ can be supposed to be diagonal. Consequently, rather than defining a precise form for U_i such as Brownian motion for instance, a common strategy consists in imposing a diagonal form for $\mathbf{G}_\theta$ (which is also very convenient from a computational point of view). However this strategy alone is not sufficient. Indeed, if U_i is defined through $\mathbf{G}_\theta$ directly, one must be sure that the induced process shares the same degree of smoothness with the fixed effect μ (ie that U_i and μ belong to the same functional space). In other words, when dealing with functional mixed-effect models, the difficulty is that if the fixed-effect curve μ is supposed to belong to some functional space, then the subject-specific deviations arising from the random functions U_i should be controled so that U_i belongs to the same space. This issue has been investigated in the context of functional mixed-models [1, 7], and this goal is achieved by controlling the exponential decrease of the variances of the random wavelet coefficients such that:

$$\mathbf{G}_{\boldsymbol{\theta}} = \underset{jk}{\text{Diag}} \left(2^{-j\eta}\gamma_\theta^2\right), \ \forall j \in \{0, \ldots, J\}, \ k \in \{0, \ldots, 2^j - 1\}. \tag{2.4}$$

This control requires the introduction of parameter η which is associated with the regularity of process U_i. This assumption has important implications in terms of modelling. Since the U_i functions are governed by their covariance structure, the exponential decay of the variance terms implies that the deviations modelled by the random effect will roughly occur at the same scales as the signal. If random variations around the fixed effect are present at large scales (j close to zero) then they will be related to inter-individual variations. On the contrary, random variations occuring at small scales (j close to J) will be considered as noise with no biological information. In other words, the random effect is supposed to be at least as regular as the fixed effect, and more regular than the noise [7]. Also, it may be necessary to allow variance γ_θ^2 to depend on scale, position and cluster membership ($\gamma^2_{\theta,\ell jk}$) [83]. This modelling can be very powerful to consider different types of random functions U_i.

2.3 Functional clustering of CGH data

We considered the clustering of breast-cancer tumors based on their copy number aberration profiles measured by array-based Comparative Genomic Hydridization [44]. We used our curve clustering framework with the Haar basis (piecewise constant basis) to perform subgroup discovery by considering random effects. In the initial publication, the genomic profiles of 62 samples were analyzed using P1/BAC CGH arrays (2464 genomic clones) [44]. We used the 55 profiles for which additional clinical information were available. The authors identified 3 main subtypes of breast cancer that differ with respect to level of genomic instability. Interestingly, a re-analyzis of the data concluded on the lack of correspondance between the two clustering results [124]. Moreover, they discovered many more subgroups and noticed that "the samples in the study could be more heterogeneous than previously implied".

We also find more subgroups than the original study. First, this suggests an increased power gained from considering the random effect in the selection step. Then we were able to identify the 1q/16p subtype on the complete dataset (with 1 mismatch). This subtype was identified in the first study [44] but not by other clustering methods [124] whereas it is associated to the best patient outcome. Since 2 of the 3 identified clusters in the original paper concern ER (Estrogen Receptor) positive tumors, we also performed our method on this subset of patients and retrieve the 1q/16p subtype without mismatch. In this classification, one cluster was made of 3 tumors also identified as similar in the original paper.

Also we proposed some criteria to quantify the signal to noise ratio as well as the strength of the random effect in functional data such that SNR_μ^2 quantifies the strength of the signal with respect to measurement noise, and λ_U is the ratio of the noise variance to the random effect variance:

$$\begin{aligned} \mathrm{SNR}_\mu^2 &= \frac{1}{n\sigma_E^2}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{\ell,j_0k}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{\ell,jk}^2\right), \\ \lambda_{\mathrm{U}} &= \sigma_E^2/\left(\gamma_\nu^2+\frac{\gamma_\theta^2}{1-2^{(1-\eta)}}\right). \end{aligned}$$

When estimated on real data (Table 2.1), the estimated signal to noise ratio appears to be low contrary to the strength of the random effect ($\widehat{\lambda}_{\mathrm{U}} \sim 10^{-4}$) which indicates that the inter-individual variability is ultra-high in these data. As a consequence, finding clusters with biological significance and outcome prediction will require rather hundreds/thousands of patients compared with 55 in the original study.

2.4 Dimension reduction in functional models

Following our work on curve clustering with random effects, we wanted to address the question of dimension reduction that is specific to the use of wavelets. In a first step, using a wavelet representation of the functional model has allowed us to characterize different types of smoothness

Table 2.1: Estimated SNR_μ^2 and λ_{U} for the breast tumor dataset of [44].

	Complete dataset			ER+ dataset	
cluster ID	$\widehat{\text{SNR}}_\mu^2$	$\widehat{\lambda}_{\text{U}}$	cluster ID	$\widehat{\text{SNR}}_\mu^2$	$\widehat{\lambda}_{\text{U}}$
1	2.1e-4	3.9e-04	1	2.1e-3	2.2e-04
2	2.3e-3	3.8e-05	2	7.8e-3	1.9e-05
3	1.3e-3	6.4e-04	3	1.1e-2	3.8e-05
4 (1q/16p)	1.5e-3	1.3e-04	4 (1q/16p)	4.4e-3	4.4e-04
5	9.3e-4	4.3e-05			

conditions assumed on the response curves $Y_i(t)$ by the mean of their wavelet coefficients. But the wavelet representation offers another advantage, that is a sparse representation for a wide variety of functional spaces, which is crucial when dealing with high dimensional data. The underlying idea of thresholding is to take advantage of the compression properties of wavelets coming from their spatially adaptive characteristics [40]. On one hand the main fixed effect μ is assumed to have a certain regularity and hence its representation in the wavelet domain will be supported in a (relatively) few number of "large" coefficients. On the other hand, the noise in the data will uniformly contaminate all coefficients. The goal is then to recover those containing information about the estimated functions by shrinking coefficients, eventually to zero if they contain only noise. Interestingly, even if it is well known that wavelets offer a sparse representation in the coefficients domain, little methodology has been proposed so far in the case of functional data analysis with multiple individuals, except in the Bayesian setting [83]. In addition, the presence of random effect is likely to influence the thresholding of fixed-effects coefficients: if inter-individual variability is present on some coefficients, should they be "more" thresholded for instance ?

In our first article on curve clustering [FP2] the procedure we proposed was based on the union of non-null coefficients across individuals obtained after individual-wise thresholding, but it was heuristic and our aim was to propose an integrated framework for wavelets thresholding in the presence of repeats and random effects. The following work is still on-going and we will present some main directions we have investigated. Facing the diversity of possible models (Table 2.2), we first focused on models without clustering in order to study dimension reduction for functional mixed models "only".

Heteroscedastic thresholding

The following results were obtained by M. Giacofci during her PhD [49], and concern the reconstruction properties of the functional fixed effect in the case of heteroscedastic functional regression. We first focused on the heteroscedastic version of the mixed model, that can be written in two

$Y_i(t)$	Without Clusters	With Clusters
Without random effect	$\mu(t) + E_i(t)$	$\mu_\ell(t) + E_i(t)$
With random effect	$\mu(t) + U_i(t) + E_i(t)$	$\mu_\ell(t) + U_i(t) + E_i(t)$

Table 2.2: Different functional models that can be used depending on the presence of clusters and/or random effects. i stands for the ith curve, and ℓ stands for cluster structure.

forms:

$$Y_i(t) = \mu(t) + U_i(t) + E_i(t),$$

with $U_i(t) \sim \mathcal{N}(0, K_U(\bullet, t))$ independent of $E_i(t) \sim \mathcal{N}(0, \sigma_E^2)$. Equivalently we can write:

$$Y_i(t) = \mu(t) + F_i(t), \quad F_i(t) \sim \mathcal{N}(0, K_F(\bullet, t)),$$

which reduces to a heteroskedastic functional regression model. Thanks to the whitening property of wavelets mentioned above, the covariance matrix of the coefficients of F_i is still diagonal with variance terms depending on the scale j and on the location k such that:

$$\forall i \in \{1 \dots, I\},\ d_{i,jk} = \beta_{jk} + f_{i,jk},\ \ f_{i,jk} \sim \mathcal{N}(0, \sigma_{jk}^2).$$

In the mixed version of the model, σ_{jk}^2 would be further written as the sum of $2^{-j\eta}\gamma_{\theta,jk}^2 + \sigma_\varepsilon^2$, however, when the purpose is to assess the reconstruction properties of an estimator of function μ, the mixed version (with the random effect specification) is not mandatory in a first step. Then we considered thresholding procedures like soft thresholding [40] that consists in respectively shrinking or thresholding, i.e setting to zero, coefficients whose absolute value are below a suitably chosen threshold. In the case of heteroscedastic regression with replicates, the soft thresholding functions that can be expressed formally such as:

$$\widehat{\beta}_{jk}(\lambda_{jk}) = \text{sign}(d_{\bullet,jk}) \left(|d_{\bullet,jk}| - \lambda_{jk}\right)_+, \tag{2.5}$$

with $d_{\bullet,jk} = 1/I \sum_{i=1}^{I} d_{i,jk}$ the averaged empirical coefficients, $(\bullet)_+ = \max(\bullet, 0)$. We adapt λ_{jk} to the heteroscedastic model such that:

$$\lambda_{jk} = \widehat{\sigma}_{jk}\sqrt{2\log n},\ \text{with}\ \widehat{\sigma}_{jk}^2 = \frac{1}{I}\sum_{i=1}^{I}(d_{i,jk} - d_{\bullet,jk})^2. \tag{2.6}$$

In the traditional framework (without replicates), the thresholding parameter would be set to $\widehat{\sigma}\sqrt{2\log(n)}$, with $\widehat{\sigma}$ a MAD (Median Absolute Deviation) estimator of the error variance. Thanks to the I replicates, we can estimate σ_{jk}^2 in the parametric framework. We also considered the Smoothly Clipped Absolute Deviation (SCAD) method [6] as a thresholding rule making a continuous trade-off between hard and soft thresholding. During her PhD [49] M. Giacofci studied the asymptotic properties of the quadratic risk of $\widehat{\mu}$. By considering replicated, two different asymptotic trends emerge:

when the variances are known, the risk associated with the fixed effect depends on the signal size n and is bounded from above by the minimax convergence rate in $\mathcal{O}(n^{-2s/2s+1})$. At each position (j,k), I repeated measurements are available to parametrically estimate the parameters σ^2_{jk}. The expected quadratic risk is then the parametrical one in $\mathcal{O}(I^{-1})$.

Penalized maximum likelihood for wavelets thresholding

Parallel to the investigation of theoretical properties of the mixed model in its heteroscedastic form, we have also investigated computational strategies to propose a thresholding procedure for the functional mixed model and to investigate its performance, even if theoretical results were not yet available. For this purpose we considered the penalized estimation framework that has been the subject of many developments in the past years with the explosion of high dimensional statistics. Thanks to historical connections between ℓ_1-based penalization strategies and thresholding [30], we propose to embed the functional model in the framework of maximum likelihood estimation. First considering the model without cluster and without random effects, in the coefficients domain we have (for wavelet coefficients):

$$\forall i \in \{1\ldots,I\},\, \mathbf{d}_i = \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0,\sigma^2_\varepsilon \mathbf{I})$. Then we consider $\log \mathcal{L}(\boldsymbol{\beta})$ the log-likelihood of this model, and using the Lasso to get thresholded coefficients, we solve:

$$\widehat{\boldsymbol{\beta}}(\lambda_\beta) \in \arg\max_{\boldsymbol{\beta}} \left\{\log \mathcal{L}(\boldsymbol{\beta}) - \lambda_\beta \|\boldsymbol{\beta}\|_1\right\},$$

which allows us to give a very simple solution to the thresholding issue in the multiple sample case,

$$\widehat{\beta}_{jk}(\lambda_\beta) = \text{sign}(d_{\bullet,jk})\left(|d_{\bullet,jk}| - \lambda_\beta\right)_+, \tag{2.7}$$

with $d_{\bullet,jk} = 1/I \sum_{i=1}^I d_{i,jk}$ the averaged empirical coefficients and $\lambda_\beta = \widehat{\sigma}_\varepsilon \sqrt{2\log(n)/I}$ the universal threshold adapted to multiple measurements. This strategy confirms empirical results on baseline estimation in the context of noisy repeated measurements [4] that pointed out the advantage of applying the thresholding procedure to the averaged wavelet coefficient instead of averaging the thresholded individual coefficients in order to use all information available in the regularization stage and enable better performance in finite sample situation.

In a second step we considered functional random effects with the following model on coefficients

$$\forall i \in \{1\ldots,I\},\, \mathbf{d}_i = \boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i,$$

with $\boldsymbol{\theta}_i \sim \mathcal{N}(0,\mathbf{G}_\theta)$, and $\mathbf{G}_\theta = \text{Diag}_{jk}(2^{-j\eta}\gamma^2_{jk})$. We propose to estimate the parameters of the model $\boldsymbol{\varphi} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma^2_\varepsilon)^T$ by penalized maximum likelihood using the EM algorithm [39] that is now well established in the context of parameter estimation for mixed linear models [123].

The key ingredient of this estimation sheme is the conditional expectation of the complete-data log-likelihood given $\mathbf{d}$:

$$\mathbb{E}_{\boldsymbol{\varphi}^{[h]}}\left[\log \mathcal{L}(\mathbf{d}, \boldsymbol{\theta}; \boldsymbol{\varphi})|\mathbf{d}\right] = Q(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{[h]}),$$

with $\mathbb{E}_{\boldsymbol{\varphi}}[\cdot]$ standing for the expectation operator using $\boldsymbol{\varphi}$ as the parameter value and $\mathbb{V}_{\boldsymbol{\varphi}}[\cdot]$ the corresponding variance.

Our proposition is to introduce two lasso penalties on both fixed and random effects: the first one is to ensure the sparse representation of the fixed effect, the second one is based on the hypothesis that among all scales and positions (j, k), only few are subject to inter-individual variations, such that $\boldsymbol{\gamma} = [\gamma_{jk}]_{(jk)}$ is a sparse vector. Then a reparametrization is needed for the penalized likelihood of the model to be convex in both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ [31], such that

$$\forall i \in \{1 \ldots, I\},\ \mathbf{d}_i = \boldsymbol{\beta} + \mathbf{G}_{\boldsymbol{\theta}}^{1/2}\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i,$$

with $\boldsymbol{\theta}_i \sim \mathcal{N}(0, \mathbf{I})$, which only modifies the interpretation of the parameters: "variances" γ_{jk}^2 become regression coefficients, but the statistical properties of the estimators remain unchanged. The penalized quantity to maximize is then:

$$Q(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{[h]}) - \lambda_\beta\|\boldsymbol{\beta}\|_1 - \lambda_\gamma\|\boldsymbol{\gamma}\|_1.$$

We also developed a similar strategy using the SCAD penalty instead of the lasso (not shown here)

Thresholding the coefficients. In the M-step of the algorithm, the fixed-effects coefficients are updated such that:

$$\boldsymbol{\beta}^{[h+1]}(\lambda_\beta) \in \arg\max_{\boldsymbol{\beta}} \left\{ Q(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{[h]}) - \lambda_\beta\|\boldsymbol{\beta}\|_1 \right\}.$$

Introducing notations $\widehat{\boldsymbol{\theta}}_i^{[h]} = \mathbb{E}_{\boldsymbol{\varphi}^{[h]}}\left[\boldsymbol{\theta}_i|\mathbf{d}_i\right]$ and $\widehat{\mathbf{r}}_i = \mathbf{d}_i - \widehat{\mathbf{G}}_{\boldsymbol{\theta}}^{1/2}\widehat{\boldsymbol{\theta}}_i$, this steps reduces to updating $\widehat{\beta}_{jk}(\lambda_\beta)$ such that

$$\widehat{\beta}_{jk}^{[h+1]}(\lambda_\beta) = \text{sign}\left(\widehat{r}_{\bullet,jk}^{[h]}\right)\left(|\widehat{r}_{\bullet,jk}^{[h]}| - \lambda_\beta\right)_+, \tag{2.8}$$

with $\widehat{r}_{\bullet,jk} = 1/I\sum_i \left(d_{i,jk} - 2^{-j\eta/2}\widehat{\theta}_{i,jk}\right)$, which corresponds to the thresholding of the observed wavelet coefficients corrected by the predictors of the random effects. As for the selection of the variance parameters, it is based on the thresholding of:

$$\rho_{jk} = \frac{\sum_{i=1}^I 2^{-j\eta/2}\widehat{\theta}_{i,jk}(d_{i,jk} - \widehat{\beta}_{jk})}{\sum_{i=1}^I 2^{-j\eta}(\widehat{\theta}_{i,jk}^{2[h]} + \mathbb{V}(\theta_{i,jk}|d_{i,jk}))}$$

that can be interpreted as a correlation coefficient between the random effect predictor and the residuals of the fixed-effects. Other derivations concern the use of the SCAD penalty that we use in practice [49] .

Where are we on this project ?

Everything works ! *But*, as any method with tuning parameters, the calibration of λ_β and λ_γ is of central importance and has currently limited our progression. This step is even more complicated by the necessity to tune two parameters instead of one. Basically, two practical choices are possible. The first one consists in the exploration of a bi-dimensional grid to select the most appropriate couple $(\lambda_\beta, \lambda_\gamma)$ adapted to the data. The second choice consists in using the special form of $\lambda_\beta = \widehat{\sigma}_\varepsilon \sqrt{2\log(n)/I}$, and explore a 1D grid for λ_γ. In any case, we will need a criterion or a procedure for selection. Traditional strategies consist in using either the BIC or cross validation. However, the first option is known to be inefficient in high dimensional models, and the second is computationally intensive (given that the estimation framework is iterative itself). Another option would be to derive theoretical forms for these tuning parameters. We will show in the last Part that this option has been very successful in the Poisson case for which the theoretical form of the penalty constant has lead to excellent selection and estimation performance without much computational effort. However, in the case of mixed functional models, given the complexity of the model, it is not sure if this theoretical lead would be feasible. Another very interesting direction would be to study the theoretical reconstruction properties of $\widehat{U}_i$, the functional predictor of the random process U_i (analog of the Best Linear Unbiased Predictor in the linear case). Results exist in the spline framework only [62]. The step forward will be to include the clustering step, in order to answer the question we first asked, that is dimension reduction for curve clustering in the presence of inter-individual variations. The last step will be to study the application of such strategy to copy-number data, but we believe that this methodology will be general enough to be applicable to many situations in practice.

Part II

Some statistical aspects of the analysis of biological networks

Chapter 3

Mixture Models for random graphs

With the increasing power of high throughput technologies and storage capacities, it is now possible to explore datasets which are in the form of complex networks. One characteristics of interest when studying complex networks is their topology or the way particles, proteins or social agents interact [114]. More generally, studying the topology is crucial to understand the organization of networks, as structure often affects function. Since networks show complex structural patterns, one common task is to find an appropriate way to summarize their structure. Many indicators have been proposed for this purpose: the degree distribution [14], the clustering coefficient [86, 3], and the small world property [114] are among the most popular. Clustering methods that have been proposed are mainly focused on community detection, *i.e.* they aim at finding groups of nodes that are highly intra-connected and poorly inter-connected [55]. However, when performing exploratory data analysis, it may be difficult to search for a particular structure. Real networks may not show community structure for instance, or may be characterized by various connectivity patterns among which community is only one feature.

Model-based clustering is a powerful alternative to those methods, as the model underlying the algorithm allows the blind search of connectivity structure without any *a priori* [87, 85, 37]. The basics of this strategy is to consider that nodes are spread among an unknown number of connectivity classes which are unknown themselves. Many names have been proposed for this model, and in the following, it will be denoted by MixNet, which is equivalent to the Stochastic Block model [87]. In the following Chapter we present the MixNet model with its associated inference framework, and some examples of applications.

3.1 Presentation of MixNet and its applications

Our method belongs to the general framework of model-based clustering of network data. This family of models includes the Stochastic Block Model [2] and the MixNet approach we developed [37, 75]. In our framework, we consider data in the form of networks, that are modelled by a random graph $G = (V, E)$ made of a set of n vertices V and a set of random edges $E =$

$\{(i,j) \in V^2, i \leftrightarrow j\}$ ($i \leftrightarrow j$ standing for nodes "i and j are connected"). We denote by $\mathbf{Y} = \{Y_{ij}, (i,j) \in E\}$ the observed measures of these interactions. When the graph is only made of binary connections, Y_{ij} is a random indicator variable that equals 1 when i and j are connected and 0 otherwise, but Y_{ij} can also be quantitative when modelling valued graphs ($Y_{ij} \in \mathbb{R}$ or $\mathbb{N}$ for instance).

Our model is based on the group-membership of nodes based on their connections characteristics. We denote by $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ the matrix of labels of nodes $1, .., n$, *i.e.* $Z_{iq} = 1$ if node i belongs to group q and 0 otherwise. In the context of unsupervised clustering, this matrix is unknown and we aim to recover these labels using the observed information contained in $\mathbf{Y}$. Model-based clustering hypotheses that if labels were known, the distribution of the interaction data would be completely determined. Hence, we start by assuming that there are Q groups with proportions $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_Q)$ such that the distribution of labels $\mathbf{Z}_i = (Z_{i1}, ...Z_{iQ})$ is Multinomial with parameter $\boldsymbol{\alpha}$. The number of groups Q is unknown and will be estimated by model selection. The distribution of the interaction data is specified conditionally to the labels:

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha}), \quad \mathbf{Z}_j \sim \mathcal{M}(1, \boldsymbol{\alpha}), \quad Y_{ij} | \{Z_{iq} Z_{j\ell} = 1\} \sim f(\bullet, \theta_{ql}), \tag{3.1}$$

where distribution $f(\bullet, \theta_{ql})$ can be Bernoulli to model presence-absence data [2, 37], Gaussian or Poisson to model fluxes or abundance data [75]. The parameters of this model are the proportions of the groups ($\boldsymbol{\alpha}$) and the parameters governing the conditional distribution of the observations ($\boldsymbol{\theta} = (\theta_{q\ell})$). In the following, we note $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$.

From a modelling point of view, MixNet constitutes a very flexible framework that can catch different topological structures, and thus is not reduced to a method that searches for modules only. The simplest situation arises with binary connections that can be modelled by Bernoulli random variables such that:

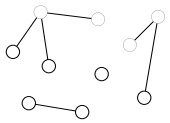
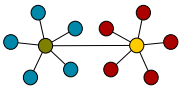
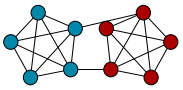
$$Y_{ij} | \{Z_{iq} Z_{j\ell} = 1\} \sim \mathcal{B}(\theta_{ql}).$$

In this model θ_{ql} describes the frequency of interactions between clusters q and ℓ, and $\boldsymbol{\theta}$ becomes a summary of the connectivity matrix. According to the form of $\boldsymbol{\theta}$, different connectivity structures can be detected in the network, as shown in Figure 3.1. Moreover, MixNet can be used as a generative model under which theoretical characteristics of random graphs can be derived, such as network motifs [FP8].

3.2 Inference and adaptation to high dimensional datasets

The objective is to estimate $\boldsymbol{\gamma}$ and to recover the unobserved labels of the data using the posterior expectation of membership $\mathbb{E}(\mathbf{Z}|\mathbf{Y})$. This is achieved using the EM-algorithm to maximize the observed-data log-likelihood denoted by $\log \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$. Unfortunately, the direct maximization of this likelihood is untractable due to the total number of possible partitions ($\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) =$

Table 3.1: Some typical network configurations and their formulation in the MixNet framework for networks with binary connections. Each class is represented by a different color.

Description	Network	Q	$\boldsymbol{\theta}$
Random		1	θ
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Clusters (affiliation networks)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$

$\sum_{\mathbf{Z}} \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\gamma})$). Hence, we use an iterative algorithm that maximizes the complete-data likelihood:

$$\begin{aligned} \log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\gamma}) &= \log \mathcal{L}(\mathbf{Z}; \boldsymbol{\alpha}) + \log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\theta}) \\ &= \sum_{iq} Z_{iq} \log(\alpha_q) + \sum_{ij,q\ell} Z_{iq} Z_{j\ell} \log f(Y_{ij}; \theta_{q\ell}) \end{aligned} \tag{3.2}$$

The labels being unknown, the algorithm proceeds as follows: the E-Step computes the conditional expectation of the complete-data log-likelihood defined as:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{[h]}) &= \mathbb{E}_{\boldsymbol{\gamma}^{[h]}} \left\{ \log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\gamma}) | \mathbf{Y} \right\} \\ &= \sum_{iq} \mathbb{E}_{\boldsymbol{\gamma}^{[h]}}(Z_{iq}|\mathbf{Y}) \log(\alpha_q) + \sum_{ij,q\ell} \mathbb{E}_{\boldsymbol{\gamma}^{[h]}}(Z_{iq} Z_{j\ell}|\mathbf{Y}) \log f(Y_{ij}; \theta_{q\ell}), \end{aligned} \tag{3.3}$$

for a current value of the parameters ($\boldsymbol{\gamma}^{[h]}$). Then the M-step maximizes $\mathcal{Q}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. Computational difficulties often arise at the E-step, mainly due to complex dependency structures that can govern the *posterior* distribution of labels given the data. This issue has motivated many methodological developments, in particular in the context of network data, with the use of variational methods [67] to approximate this *posterior* distribution [37, 75]. In Daudin et al. (2008) [37] we propose to approximate the *posterior* distribution of $\mathbf{Z}|\mathbf{Y}$ by a factorized distribution, which results in the so-called mean-field approximation. This allows us to compute an approximation of $\mathbb{E}_{\boldsymbol{\gamma}^{[h]}}(Z_{iq}|\mathbf{Y})$ from which we infer the labels $\widehat{\mathbf{Z}}_i$ by using a *maximum a posteriori* rule.

In Zanghi et al. (2010) [FP16] we adapted the estimation framework to large or growing networks, by considering so-called *online* strategies which are suitable when data arise sequentially, or when the network is so big that it can not be downloaded at once.

The last step of the inference procedure is to select the number of clusters that is unknown. By using an integrated-likelihood strategy [19], we provided a BIC-like criterion (called ICL for Integrated Classification Likelihood) that is adapted to the case of networks. For a model with Q groups:

$$\text{ICL}_Q = \max_{\boldsymbol{\gamma}} \log \mathcal{L}(\mathbf{Y}, \widehat{\mathbf{Z}}; \boldsymbol{\gamma}) - \frac{1}{2}\frac{Q(Q+1)}{2} \log\left(\frac{n(n-1)}{2}\right) - \frac{1}{2}(Q-1)\log(n).$$

Interestingly, we end up with a two-term pernalty, the first one focusing on the estimation of connection parameters $\boldsymbol{\theta}$, with $n(n-1)/2$ edges as statistical units, and the second one for proportion parameters, with n nodes as statistical units.

3.3 Applications of MixNet

In Vernoux et al. (2011) [FP14] we used the MixNet framework to investigate the global structure of the interaction network of the auxin plant hormone. The control of gene expression in response to auxin involves a complex network of over 50 potentially interacting transcriptional activators and repressors, the auxin response factors (ARFs) and Aux/IAAs (Auxin Indole-3-Acetic Acid). Our colleagues performed a large-scale analysis of the Aux/IAA-ARF pathway in the shoot apex of Arabidopsis, where dynamic auxin-based patterning controls organogenesis. The global structure of the Aux/IAA-ARF network was investigated using a high-throughput yeast two-hybrid approach, and MixNet was used to explore the organization of this network (to determine sets of proteins with similar interactors, Figure 3.1). Three well separated clusters, characterized by contrasting probabilities for within- and between-cluster connectivity, were uncovered. Our results indicate that the topology of the network relies on three principal features: (i) Aux/IAA proteins interact with themselves, (ii) Aux/IAA proteins interact with ARF activators and (iii) ARF repressors have no or very limited interactions with other proteins in the network. This topological study was the first step towards a deeper understanding of the Auxin signalling in shoot meristem. In Picard et al. (2009) [FP11] we also proposed different fields of application of MixNet to trophic, metabolic, and brain connectivity networks, field in which MixNet has recently been emphasized [90]. Finally, the MixNet framework has also been derived to perform sequence clustering [FP4]. Parallel to the statistical developments we published we also developed the `mixnet` software, that has been also adapted to `R` with the `Mixer` package.

3.4 Spatial clustering and application to ecological data

Our first motivation in the development of MixNet was the analysis of networks arising in molecular biology. However in many ecological studies, researchers also analyze data describing the

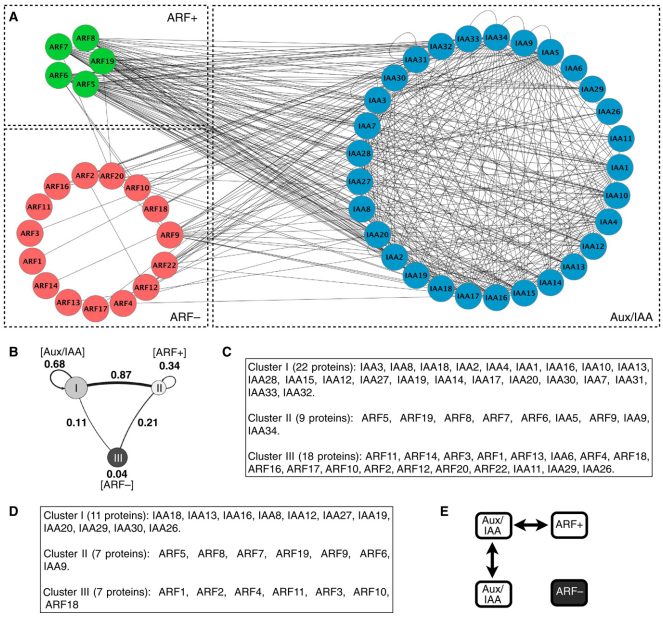


Figure 3.1: Structure the auxin signalling network as proposed in Vernoux et al. (2011). (A) Visual representation of the Aux/IAA-ARF interactome using Cytoscape (`http://www.cytoscape.org`). The proteins are grouped according to their biological identity as indicated. Note the global differences in connectivity of the three biological groups (BD) Connectivity graph and clusters identified by the MixNet algorithm. The probabilities associated with the connectivity structure for the global network are indicated in (B). The three clusters are mainly composed of Aux/IAA (I), ARF activators (II) and ARF repressors (III) as indicated in brackets in (B). The identity of the proteins in these clusters for both the global network (C) and the SAM-specific network (D) is shown. The proteins are ordered from the most to the least central in each cluster based on the distance of the protein to the cluster. (E) The topology of the network relies on stereotypic interaction capacities for the different classes of proteins as represented. ARF+: ARF activators;ARF-: ARF repressors.

interactions between individuals, populations, species or communities. These interactions can be directly observed like trophic relationships in a food-web [71] or they can be inferred from computed distance/similarity measures like genetic distances that have been developed to summarize allele frequency differences between populations [68]. Describing and summarizing these sets of pairwise interactions is an important step to better understand the functioning of ecological systems. These interactions would be denoted by $\mathbf{Y}$ to keep notation homogeneous with the MixNet framework.

Moreover some ecological data have also explicit geographic locations transforming ecological networks into *spatial networks* [35]. When available, this spatial information is often used *posterior* to the identification of groups to improve their ecological interpretation [81, 8, 36]. However, if the aim of a study is to identify spatially-coherent groups (e.g., habitat patches), this indirect approach may not be optimal as it considers the spatial aspect only after summarizing the network structure. We choose to use what we called a structural network that records the spatial proximity between ecological entities of the ecological network $\mathbf{Y}$ (see Figure 3.2). Structural networks are sometimes directly available such as road networks [111], but they are usually constructed using geographical data with *ad-hoc* techniques such as maximum spanning trees [9], k-nearest neighbors [56], distance thresholding [91] or edge-thinning [119, 69]. In the following we suppose that the structural network is given and fixed. It is denoted by $\mathbf{X} = (X_{ij})$ such that $(X_{ij} > 0)$ is the geographical proximity between entities i and j and $X_{ij} = 0$ if they are not connected. We assume $\mathbf{X}$ is symmetric. The entities are the same as in the ecological network $\mathbf{Y}$.

In Miele et al. (2014) [FP5] we propose to embed the geographical information within a regularization framework by adding some constraints in the maximum likelihood estimation of parameters. In regularization techniques, a constraint defined by a network can be introduced using the graph Laplacian [64]. For a network with connection matrix $\mathbf{X} = (X_{ij})$, the Laplacian is defined by $\mathbf{L_X} = \mathbf{D} - \mathbf{X}$ where $\mathbf{D}$ is the diagonal matrix of degrees with diagonal terms $d_i = \sum_j X_{ij}$. The Laplacian $\mathbf{L_X}$ can then be used as a semi-metric to measure the spatial variability. Indeed, for a given vector $\mathbf{u} = (u_1, ...u_n)$, we have:

$$\|\mathbf{u}\|^2_{\mathbf{L_X}} = \mathbf{u}^T \mathbf{L_X} \mathbf{u} = \sum_{i \sim j} X_{ij}(u_i - u_j)^2,$$

which is the squared distance between values of $\mathbf{u}$ weighted by their spatial proximities contained in $\mathbf{X}$. We develop an original regularization procedure aiming to reduce the variation of labels along the structural network. Whereas the vector of parameters is traditionally regularized, our approach considers that the vector of labels can be regularized using the structural network $\mathbf{X}$. Denoting by $\mathbf{Z}^q = (Z_{1q}, \ldots, Z_{nq})$ the vector of individuals for label q, we propose the following penalty:

$$\text{pen}(\mathbf{Z}; \mathbf{L_X}) = \sum_{q=1}^{Q} \|\mathbf{Z}^q\|^2_{\mathbf{L_X}} = \sum_{q=1}^{Q} \sum_{i \sim j} X_{ij}(Z_{iq} - Z_{jq})^2.$$

Let us consider the case where $X_{ij} \in \{0, 1\}$ to interpret the penalty. In this case,

$$\text{pen}(\mathbf{Z}; \mathbf{L}_\mathbf{X}) = \sum_{q=1}^{Q} \sum_{i \sim j} (Z_{iq} - Z_{jq})^2 = \sum_{q=1}^{Q} \sum_{i \sim j} 1_{\{Z_{iq} \neq Z_{jq}\}}$$

with $i \sim j$ standing for entities i and j connected in the structural network, so that the penalty accounts for the number of edges in the structural network that have discordant labels.

Considering the new penalized likelihood, the regularized EM algorithm is based on the conditional expectation of the penalized complete-data likelihood

$$\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{[h]}) - \lambda \times \mathbb{E}_{\boldsymbol{\gamma}^{[h]}} \left\{ \text{pen}(\mathbf{Z}; \mathbf{L}_\mathbf{X}) | \mathbf{Y} \right\},$$

with λ a penalty constant. Since the penalty term does not involve the parameters but only the labels, the maximization step is unchanged [75]. As it is the case in the standard MixNet framework, the posterior distribution of labels given the observations is not tractable and is approximated within the variational framework. Keeping similar notations, $\tau_{iq} \simeq \mathbb{E}_{\boldsymbol{\gamma}} \{ Z_{iq} | \mathbf{Y} \}$ the approximate posterior expectation of labels, the computation of these approximate posterior probabilities is achieved by solving a fixed-point algorithm [5, 75] and we account for spatial constraints through the penalty term, which induces the spatial homogeneity of the labels (we refer to the original article for further details [FP5]).

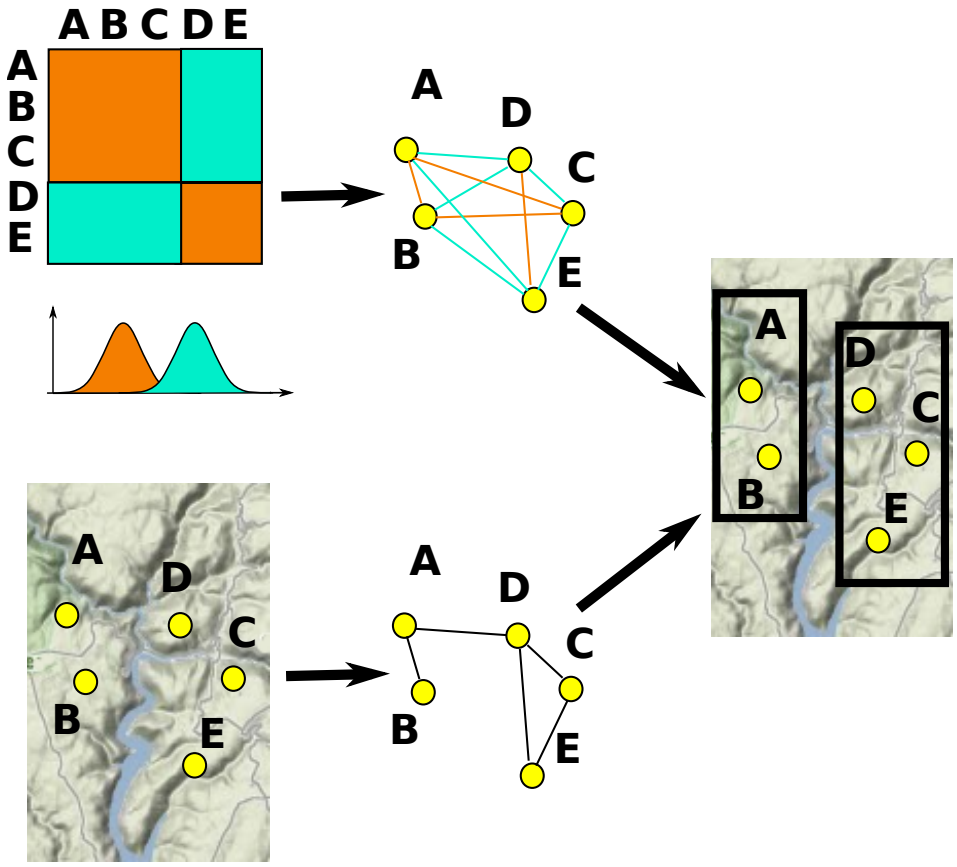


Figure 3.2: Framework of the spatial version of MixNet. The ecological network (top) records the ecological distance between entities. The structural network (bottom) summarizes the proximity between geographical locations. Our method deciphers groups of entities (black squares on the right) using both networks into a model-based strategy associated to a regularization framework. While entities A,B,C and D,E form two groups in the ecological data, the geographical constraints leads to lastly grouping C with D and E.

Chapter 4

The generalized fused lasso

Network data have become so common that they now constitute some *prior* knowledge for downstream statistical analysis: for instance, two proteins of the same biological pathway are likely to share similar effects on the response to a treatment or on disease development. Consequently, statistical methods, like regression and model selection have recently focused on structured sparsity. In addition to the classical sparsity assumption (under which only a small fraction of the variables are relevant), these methods work under the assumption that two connected covariates in the network may share similar effects on the response variable. Consequently the objective of structured sparsity is twofold: improve model selection by using some *prior* knowledge on the structure; increase prediction performance by effective dimensionality reduction based on the *prior* knowledge that several covariates may share the exact same coefficient.

Most methods proposed so far use a penalized version of the log-likelihood based on some structured sparsity-inducing penalty. The fused lasso of Tibshirani [116] is one particular example: in addition to the ℓ_1-norm penalty of the lasso [115], the fused lasso penalizes the ℓ_1-norm of the vector of successive differences. It is therefore especially adapted for smoothing, when covariates are ordered and are likely to share similar effects with their direct neighbor. In particular, it has been applied to the analysis of copy number data [117] and the fused penalty was further developed in the context of patient status prediction using Support Vector Machines based on a fused penalty [98]. Our first motivation in the following developments was to investigate the performance of logistic regression based on the fused lasso for discriminating cancer subtypes based on copy number data (Figure 4.1). However, the fused-lasso has further been generalized to handle more complex structure among feature effects, in particular networks of features [61]. The network is modelled as a graph with vertices standing for the p coefficients of the model, and with a set of edges. It is used in the penalty as *prior* information to penalize the absolute value of the difference of connected coefficients, leading to the generalized fused lasso.

As any *prior* information, the underlying graph can be more or less informative. For instance, the clustered lasso [109] was proposed when only the existence of a structure is assumed but no particular knowledge allows for its precise description. Its main step involves a penalty based on

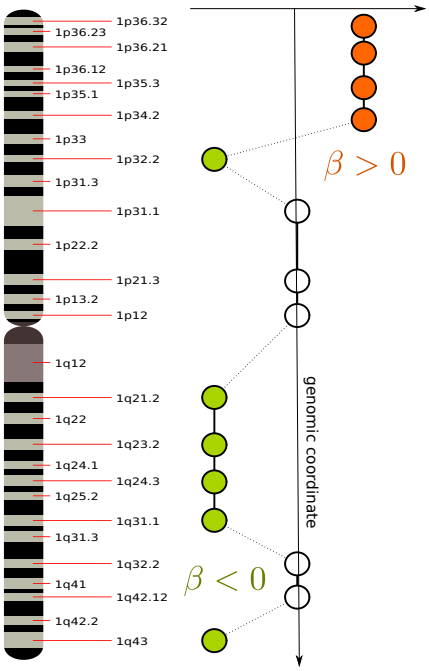


Figure 4.1: Illustration of the use of the fused penalty with logistic regression for array CGH data analysis. If $Y_i = 1$ corresponds to cancer status in a logistic model with parameters $\boldsymbol{\beta}$, positive coefficients are associated with increased risk for instance. As measurements are organized along the genome, the fused penalty may help in the identifying chromosomal regions associated with similar risk.

the ℓ_1-norm of the vector of all the $p(p-1)/2$ differences among the parameter values. This strategy corresponds to the generalized fused lasso with the graph set to a clique that connects all coefficients. When penalizing all differences, it is very likely that some differences are unnecessarily penalized, which raises the question of the method robustness to graph misspecification. Interestingly, any structured-sparsity approach is concerned by this robustness property, but this question has never been thoroughly investigated [12].

In Viallon et al. (2014) [FP15] we focus on the adaptive generalized fused lasso in the context of generalized linear models. In the following, "adaptive" refers to the use of adaptive weights as developed for the lasso [129]. We prove that adaptive generalized fused lasso estimators enjoy asymptotic oracle properties in the fixed p setting. In particular, we observe that only adaptive versions of the generalized fused lasso enjoy asymptotic oracle properties (i.e., such that, as n grows to infinity, the correct support is recovered with probability tending to one and estimates of non-zero coefficients perform as well as if the true underlying model were given in advance). In a further step we investigate the empirical benefits of using an ℓ_1-based fusion penalty on support recovery and prediction as compared with other penalization strategies, under logistic models.

4.1 The adaptive generalized fused lasso in generalized linear models

We consider the generalized linear models framework [77] with Y_i the *response* variable $i = 1, \ldots, n$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ a p-dimensional vector of features. We further set $\mathbf{z}_i = (1, \mathbf{x}_i^T)^T$, and we consider the *fixed design* case with $\sum_{i=1}^n x_{ij} = 0$. For generalized linear models the distribution of the response variable is given by

$$f(y_i, \boldsymbol{\beta}^*, \phi) = \exp\left(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

where ϕ is a dispersion parameter and functions $b(\cdot)$, $a(\cdot)$ and $c(\cdot, \cdot)$ are known. The linear predictor η_i is given by $\mathbf{z}_i^T\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}_{\setminus 0}^*)^T \in \mathbb{R}^{p+1}$ stands for the vector of coefficients, with β_0^* the intercept parameter and $\boldsymbol{\beta}_{\setminus 0}^* = (\beta_1^*, \ldots, \beta_p^*)$. The mean $\mu_i = \mathbb{E}(Y_i)$ is related to the linear predictor via the link function g: $g(\mu_i) = \eta_i$. Here we consider the canonical link function. Estimation of the parameter vector $\boldsymbol{\beta}^*$ is usually performed by the maximum likelihood method. It consists in minimizing J, given by $J(\boldsymbol{\beta}) = -\sum_{i=1}^n \log f(y_i, \boldsymbol{\beta}, \phi)$, with respect to $\boldsymbol{\beta}$. In the simulation studies and the applications below we focus on the logistic model for which $Y_i \in \{0, 1\}$, $a(\phi) = 1$, $b(x) = \log(1 + \exp(x))$ and $c \equiv 0$. Under logistic models, the mean and the linear predictor are related by $\mu_i = 1/(1 + \exp(-\mathbf{z}_i^T\boldsymbol{\beta}^*)) = g^{-1}(\eta_i)$.

As mentioned above, we further focus on the generalized fused lasso [61]. Consider a graph $G = (V, E)$, with node set $V = \{1, ...p\}$ that corresponds to the coefficient indices in $\boldsymbol{\beta}_{\setminus 0}$, and edge set E that corresponds to pairs of connected coefficient indices (j, ℓ) with $j > \ell$. The graph G that is used in the penalty is fixed and represents some $prior$ knowledge, given by an expert. The adaptive generalized fused lasso penalty consists in penalizing all coefficients along with all

coefficient differences for which an edge exists in G:

$$\text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}) = \lambda_n^{(1)} \sum_{j \in V} w_j^{(1)} |\beta_j| + \lambda_n^{(2)} \sum_{(j,\ell) \in E} w_{j\ell}^{(2)} |\beta_j - \beta_\ell|.$$

In the fixed p case considered here, and following the idea of the adaptive lasso [129], adaptive weights $w_j^{(1)}$ and $w_{j\ell}^{(2)}$ are based on initial Maximum-Likelihood estimates $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$. More precisely, for some $\gamma > 0$, we set $w_j^{(1)} = |\widetilde{\beta}_j|^{-\gamma}$ and $w_{j\ell}^{(2)} = |\widetilde{\beta}_j - \widetilde{\beta}_\ell|^{-\gamma}$. The rationale is to penalize more heavily coefficients (or differences of coefficients) when their initial estimates are small. A typical value (that we use) for γ is 1. The adaptive generalized fused lasso criterion Q is then simply defined, for given graph G and weights $\mathbf{w}$, as

$$Q(\boldsymbol{\beta}) = J(\boldsymbol{\beta}) + \text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}). \tag{4.1}$$

4.2 Asymptotic properties of the generalized fused lasso

We study the asymptotic properties of the adaptive generalized fused lasso estimator in generalized linear models for fixed p and growing n. In the following we show that for appropriate choices of $\lambda_n^{(m)} = O(\sqrt{n})$ for $m = 1, 2$, the adaptive generalized fused lasso estimator $\widehat{\boldsymbol{\beta}}^{ad}$, defined as the minimizer of criterion Q (Eq 4.1), enjoys asymptotic oracle properties, contrasting with its non-adaptive counterpart.

Before stating our results some notations are needed (technical assumptions are fully detailed in the article [FP15]). Let $\mathcal{A} = \{1 \leq j \leq p,\, \beta_j^* \neq 0\}$ be the *support* of $\boldsymbol{\beta}^*_{\backslash 0}$ and $p_0 = |\mathcal{A}|$ its cardinality. Further consider the set

$$\mathcal{B} = \{(j, \ell) \in E,\, \beta_j^* \neq 0 \text{ and } \beta_j^* = \beta_\ell^*\} \subset \mathcal{A} \times \mathcal{A}.$$

Then the number s_0 of distinct non-zero values in $\boldsymbol{\beta}^*_{\backslash 0}$ "supported" by G needs to be precisely defined (s_0 can be seen as the theoretical model complexity "supported" by G). To this end, first observe that $\mathcal{A} \subseteq V$ and $\mathcal{B} = \{(j, \ell) \in E : \beta_j^* \beta_\ell^* \neq 0, \beta_j^* = \beta_\ell^*\} \subseteq E$. Then consider the sub-graph $G_{\mathcal{B}} = (\mathcal{A}, \mathcal{B})$ of G that corresponds to the sub-graph made of non-null and equal coefficients. Let us denote by s_0 the number of its connected components (e.g., in the particular case where G is a chain graph, s_0 is the number of segments consisting of non-zero and equal coefficients). Observe that $d_0 \leq s_0 \leq p_0$, where $p_0 = |\mathcal{A}|$ is the number of non-zero coefficients in $\boldsymbol{\beta}^*_{\backslash 0}$ and d_0 is the number of *distinct* non-zero values in $\boldsymbol{\beta}^*_{\backslash 0}$. We actually have $s_0 = p_0$ if and only if $(\beta_j^* = \beta_\ell^* \neq 0 \Rightarrow (j, \ell) \notin E)$. Moreover, two coefficients that are theoretically equal can not be fused together if they do not belong to the same connected component in $G_{\mathcal{B}}$. Furthermore, $s_0 = d_0$ if and only if for all (j, ℓ) such that $\beta_j^* = \beta_\ell^*$, j and ℓ belong to the same connected component of $G_{\mathcal{B}}$. Now denote by $\mathcal{A}_1, \ldots \mathcal{A}_{s_0}$ the sets of vertices of each connected components of $G_{\mathcal{B}}$. Of course, we have $\mathcal{A} = \bigcup_{s=1}^{s_0} \mathcal{A}_s$. Further set $j_s = \min\{\mathcal{A}_s\}$ for $s = 1, \ldots, s_0$.

Now we can define $\boldsymbol{\beta}^*_{\mathcal{B}} = (\beta^*_0, \beta^*_{j_1}, \ldots, \beta^*_{j_{s_0}})^T$, which is composed by the intercept and the s_0 distinct non-zero values of $\boldsymbol{\beta}^*_{\setminus 0}$ supported by G; we further set $\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}}$ its estimate. Now denote by $\mathbf{X}_{\mathcal{B}}$ the matrix of size $n \times s_0$, whose s-th column is $X_{\mathcal{B}_s} = \sum_{j \in \mathcal{A}_s} X_j$, where X_j is the j-th column of $\mathbf{X}$. Further set $\mathbf{Z}_{\mathcal{B}} = (\mathbf{1}_n, \mathbf{X}_{\mathcal{B}})$ and denote by $\mathbf{C}_{\mathcal{B}}$ the $(s_0+1) \times (s_0+1)$ positive definite matrix that is defined as the limit, as $n \to \infty$, of $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}})/n$, where $\mathcal{I}(\boldsymbol{\beta})$ stands for the empirical Fisher's matrix of $\boldsymbol{\beta}$, with $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}}) = \mathbf{Z}^T_{\mathcal{B}} \boldsymbol{D} \mathbf{Z}_{\mathcal{B}}$ and $\boldsymbol{D}$ denotes an $n \times n$ diagonal matrix ($D_{ii} = \mu_i(1-\mu_i)$ under logistic regression models for instance). Finally introduce $\mathcal{A}_n = \{1 \leq j \leq p,\ \widehat{\beta}^{ad}_j \neq 0\}$ and $\mathcal{B}_n = \{(j,\ell) \in E,\ \widehat{\beta}^{ad}_j \neq 0 \text{ and } \widehat{\beta}^{ad}_j = \widehat{\beta}^{ad}_\ell\}$. We have now all the ingredients to state our main result.

Theorem 1 *If $\lambda^{(m)}_n / \sqrt{n} \to 0$ and $\lambda^{(m)}_n n^{(\gamma-1)/2} \to \infty$, $m = 1, 2$, then, under technical assumptions (detailed elsewhere [FP15]), the adaptive generalized fused lasso estimator satisfies the following properties:*

1. *Consistency in variable selection: $\mathbb{P}[\mathcal{A}_n = \mathcal{A}] \to 1$ and $\mathbb{P}[\mathcal{B}_n = \mathcal{B}] \to 1$ as $n \to +\infty$.*
2. *Asymptotic normality: $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}} - \boldsymbol{\beta}^*_{\mathcal{B}}\right) \longrightarrow_d \mathcal{N}\left(\mathbf{0}_{s_0+1}, {\mathbf{C}_{\mathcal{B}}}^{-1}\right)$.*

Interestingly Theorem 1 allows us to compare the asymptotic theoretical performance of various graph-based methods. Observing that $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}})$ is the information matrix of the true submodel as soon as $s_0 = d_0$, Theorem 1 states that in the fixed p scenario, the estimator $\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}}$ is asymptotically efficient as soon as $s_0 = d_0$, which is notably the case for clique-based methods [108, 109]. Moreover, because adding edges in any given graph between coefficients with theoretical different values does not modify the set $\mathcal{B}$, and so leaves the quantity s_0 unchanged, our theoretical results state that, asymptotically, adding edges in the graph can only improve the adaptive generalized fused lasso performance. However, removing edges between coefficients with theoretical equal value may modify the set $\mathcal{B}$ and increase the quantity s_0, leading to poorer asymptotic performance. These results being asymptotic, we evaluate the finite sample properties of the generalized fused lasso in the forthcoming simulation study, with an emphasis on its robustness to graph misspecification.

4.3 Performance of the generalized fused lasso

We performed an extensive simulation study to compare the performance of the generalized fused lasso with other penalized strategies, in the logistic regression setting. Our main objective was to study the impact of a graph misspecification on the generalized fused lasso performance. For this purpose we developed a very detailed simulation framework. To study the robustness of selection method to a graph misspecification, we generated graphs with varying suitabilities such that equal (resp. non-equal) coefficients were connected with probability θ (resp. $1-\theta$) as illustrated in Figure 4.2.

When no *prior* information is available on graph G, a strategy can be to use no graph (with the lasso) or to use a graph that connects all coefficients (clique-graph). This latter option was

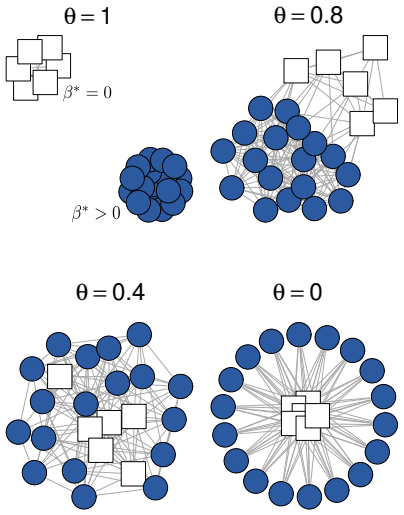


Figure 4.2: Illustration of our procedure that generates graphs with varying suitability. θ stands for the probability for two nodes (two elements of $\boldsymbol{\beta}^*$) of being connected. The p elements of $\boldsymbol{\beta}^*$ are fixed with p_0 non-null and equal elements. When θ increases fewer edges connect distinct coefficients, the most [resp. least] favorable configuration for generalized fused estimates being when $\theta = 1$ [resp. $\theta = 0$].

also considered here to compare the generalized fused lasso with clique-based methods [108, 109]. The adaptive generalized fused lasso was solved with the coordinate-wise optimization algorithm of [61] implemented in the `FusedLasso` `R` package (developed by H. Hofling), that is available from the `CRAN`. We also mention that $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$ were calibrated using the BIC in practice. In the following, we give a brief summary of our simulation results on support recovery. For this purpose we define the accuracy in support recovery by the proportion of correctly assigned (null or non-null) coefficients with respect to the true simulated coefficients (Figure 4.3).

In the best-case scenario (perfectly suited graph, $\theta = 1$), all graph-based methods are more accurate than the lasso for support recovery (and to a lesser extent for prediction), which reflects a cooperative effect that is characteristic of ℓ_1-based fused penalties: we showed that ℓ_1-based fused strategies needed less signal than the lasso to recover the support of vector $\boldsymbol{\beta}$ thanks to the edges provided to the graph connecting null (or non-null) coefficients together. Then we showed that using adaptive weights and/or relaxation increases the robustness to graph misspecifications. Next, focusing on the strategy that consists in penalizing all possible differences using the clique-graph [108, 109], we observe that they are close to those of graph-based methods with low suitability ($\theta = 0; 0.4$). Generalized fused lasso estimates obtained with the clique-graph never significantly improve upon the lasso for support recovery in our experiments. These results complement those of Theorem 1 above: while using a clique-graph is asymptotically optimal, this strategy is clearly sub-optimal on finite samples.

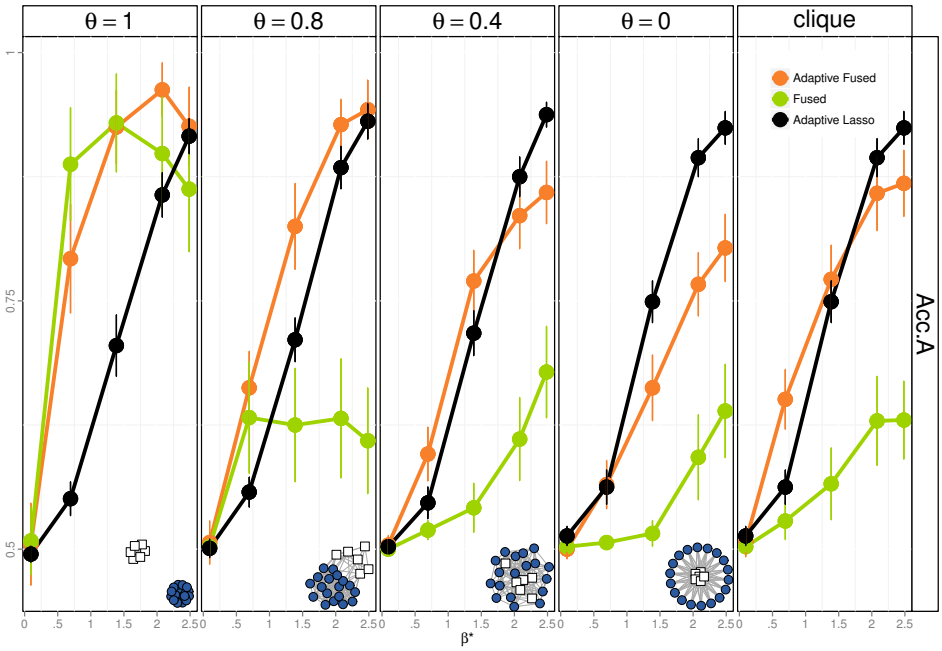


Figure 4.3: Accuracies in support recoveries for different ℓ_1-based strategies, according to the strength of the signal (intensity of β^*) and the suitability of the graph provided to the fused penalties (θ).

4.4 Network-based prediction of cancer status based on expression data

Genomics has faced a flood of network data in the last years, ranging from protein-protein interaction data, pathway data to regulation networks [52, 104]. The molecular characterization of cancers has been at the core of many projects, especially to establish molecular subtypes of histologically similar tumors. In particular finding genomic signatures has been the *graal* for many studies to predict patient outcome, survival or relapse [54]. Such signatures can be determined using a penalized logistic regression model based on gene-expression data as covariates. Here we consider the prediction of the 5-year relapse status of 214 women with breast cancer (80 relapse in the sample) [54]. Covariates correspond to the measurement of the $p = 54,613$ gene expressions reduced to the 248 genes differentially expressed (FDR=0.05), and we use 5-fold Monte Carlo cross validations to assess prediction performance. Interestingly, the expression of different genes is structured according to some unknown regulatory network that can be inferred from the data using Gaussian Graphical models for instance [33]. Our hypothesis here is that using this inferred network can help in the prediction of patients outcome. However, since this regulatory network is not perfectly known, our strategy is based on the hypothesis that there is no "true" regulatory network, and we explore the robustness of the generalized fused lasso to the addition/removal of edges in the penalty, as we did in the simulation study. To proceed we consider the regulatory network that is inferred on the training data by the `SIMoNE` package [33]. This package is based on sparse Gaussian Graphical models, and by varying the amount of shrinkage, we were able to consider networks with increasing number of edges, and then to assess the impact of changes in the network on prediction performance and estimated model dimensions.

The first conclusion is that the gain in using fused-based strategies is massive: the AUC (Area under the Curve) jumps from $\sim$0.7 for the lasso to $\sim$0.95 for generalized fused estimates, and the empirical error rate drops from $\sim$0.3 to $\sim$0.1, which clearly indicates that the network has helped in the correct classification of samples (Figure 4.4). Moreover, as previously mentioned in the simulations, fused-based methods are more performant than the lasso, but there is no significant difference among them for prediction. Very interestingly, the classification performance is maintained with the addition of edges, which suggests that the suitability of these new edges is about $\theta \simeq 0.5$. Then the number of non-null estimated coefficients is higher for all fused-penalties, and the estimated number of distinct non-null coefficients (a crude estimate of quantity s_0) converges towards a set of $\sim$10 distinct values for estimated parameters.

4.5 Discussion

We investigated theoretical and empirical properties of the generalized fused lasso, in various settings. From the theoretical point of view, we especially show that using adaptive weights leads to estimators enjoying asymptotic oracle properties. However, for the true underlying dimension of the problem d_0 (that is the number of distinct non-null values in $\boldsymbol{\beta}^*$) to be equal to the asymptotic

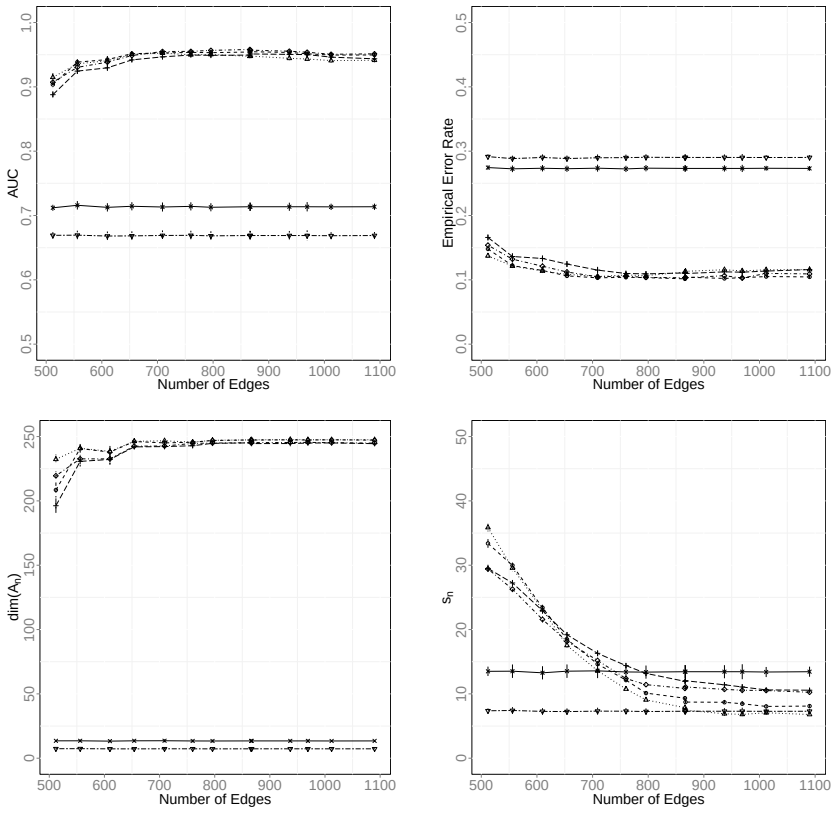


Figure 4.4: Performance of penalization strategies for AUC (Area Under the Curve) and Empirical Error Rate on the breast cancer dataset with an increasing number of edges in the network provided in the penalty. Bottom panels display the estimated model size and complexities (number of non-null estimated coefficients, and number of distinct estimated coefficients) according to an increase in the number of edges in the penalty graph. Fused lasso (dotted), adaptive fused lasso (dashed), relaxed fused lasso (dotdash), relaxed adaptive fused lasso (longdashed), relaxed adaptive lasso (plain), GE-Net (two-dash).

dimension s_0 of the estimator, the graph G used in the penalty has to enjoy the following property: for all (j, ℓ) such that $\beta^*_j = \beta^*_\ell$, j and ℓ should belong to the same connected component of $G_{\mathcal{B}}$, the sub-graph of G such that $G_{\mathcal{B}} = (\mathcal{A}, \mathcal{B})$. In particular, our results indicate that setting G to the clique connecting all coefficients of $\boldsymbol{\beta}^*$ together (in which case all the $p(p-1)/2$ differences are penalized) is asymptotically optimal. In other words, it means that, asymptotically, adding misleading edges in the graph is harmless, while forgetting relevant ones can be harmful.

From the modelling point of view however, we empirically studied the robustness of generalized fused lasso estimates against graph misspecification on finite samples. We demonstrated that the performance of generalized fused lasso estimates on finite samples are deeply related to the suitability of the graph in the penalty, especially for support recovery. In particular, we show that, under the designs considered in our simulations, the clique-based strategy is clearly sub-optimal for support recovery, so that misleading edges are harmful on finite samples. The graph used in the penalty constitutes a formal description of some *prior* knowledge on the problem that is investigated, and has to be determined with caution, especially if support recovery matters.

We established the asymptotic oracle properties of the adaptive generalized fused lasso estimates under generalized linear models, for fixed p. These results are the first established for fused lasso estimates in the setting of generalized linear models and for the generalized fused penalty based on a graph, and they should be extended to cover the high-dimensional case. Most published works on the fused lasso in high-dimension focus on the chain-based fused penalty in the Gaussian sequence model. The extension to generalized linear models would be an interesting lead. From a practical point of view, preliminary simulations (not shown here) showed that classification performance were degradated in the high dimensional setting, which makes the application of the fused lasso for cancer discrimination highly not performant. For now, an interesting direction would be to develop a method for feature pre-selection, in order to focus on a subset of *prior* relevant features.

Part III

Analysis of sequencing data and future projects

Chapter 5

Analysis of replication origins data

The faithful duplication of mammalian genomes at each S phase[1] of a mitosis is under the control of a spatiotemporal program that orchestrates and regulates both the positioning and the timing of firing of replication starting points also called replication origins. The molecular mechanisms involved in coordinating the activation of 50,000 to 100,000 origins in each cell and at each cell cycle are still poorly understood, despite the need for a comprehensive understanding of these processes. Indeed, defects in the normal sequence of events leading to replication initiation may be directly responsible for genomic instability and/or the deregulation of differentiation programs. Consequently, the first and necessary step towards understanding this regulation is to refine our vision of the spatiotemporal replication program. For this reason several laboratories have chosen to map both the spatial and temporal programs of replication, in different systems and cell lines.

The temporal program of replication has been successfully analyzed in many laboratories with no particular controversy. By contrast, attempts to identify replication origins remain a subject of passionate debate in the field, as the intrinsic rarity of replication bubbles makes it difficult to purify the genomic material. The most popular method for mapping replication starting points in mammals is the purification of short nascent strands (SNS). Briefly, replication starts by an opening of the double-stranded DNA molecule to be replicated, which forms a "bubble" that separates the two strands and starts the replication fork. At these precise loci, new DNA is synthetized in order to initiate the replication process, and these newly synthetized fragments are known as Short Nascent Strands (Figure 5.1). These Short Nascent Strands can be isolated and sequenced. Once mapped on the human genome (Figure 5.2), their accumulation indicate an "origin" activity [24, 28, 48].

Agreement on a consensual protocol for SNS enrichment and quantification has also become a critical issue as the scale of investigation of replication origins has changed profoundly in recent years. Beginning with investigations of individual loci, and continuing with the microarray technology, there has recently been another technological shift in this field towards the use of ultra-deep sequencing [51]. *Origin-omics* has now become a way of thinking about replication that incorporates tens of thousands of loci embedded within various genomic landscapes. The emphasis also

[1] The S phase is the part of the cell cycle during which DNA is synthetized.

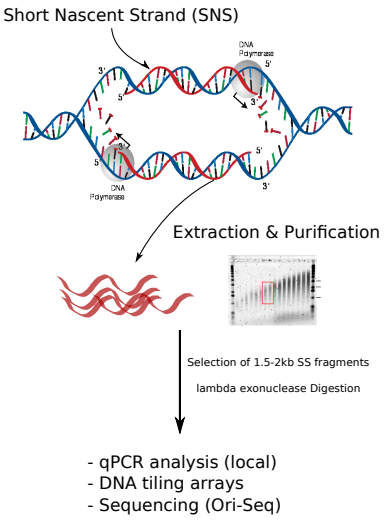


Figure 5.1: An opened replication bubble that separates the two native strands (in blue). New complementary fragments called Short Nascent Strands are synthetized (in red) to initiate the replication process. These fragments are isolated and purified according to their size (on a gel). They are further amplified and sequenced. Note that in one experiment there are $\sim 10^8$ cells that may use the same origins. Consequently our data are cumulated data on a population of cells (we do not habe access to single-molecule bubbles).

needs to shift from protocols to methods used for the analysis of genome-wide replication data. Indeed, despite a spectacular increase in the sensitivity of detection, *Origin-omics* is already subject to the same pitfalls as all other types of *omics*: the difficulty achieving an appropriate balance between the specificity and sensitivity of the analysis method. In a recent study based on the ultra-deep sequencing of SNS, origins were detected using chIP-Seq tools [20] for peak detection. This resulted in 250,000 identified origins in different human cell lines [18]. We noticed however one possible caveat in the use of chIP-Seq tools for the detection of replication origins based on sequenced SNS. Indeed, *prior* to the sequencing, SNS are first selected based on their size (about 1.5-2kb). Hence the resolution of detection of replication origins cannot be less than this size. It is therefore possible that chIP-Seq tools tend to split the signal into multiple peaks and hence tend to overestimate the number of replication origins. In Picard et al. (2014) [FP7] we addressed this issue of resolution of detection. We proposed a peak-detection method that is adapted to the special case of SNS sequencing data, based on the *prior* control of the resolution of detection of exceptional local enrichments of reads. The method relies on sliding windows, the size of which is imposed by the size of the sequenced SNS fragments. We deal with multiple testing by providing a significance threshold that controls for false-positive detections, and that is adaptive to local coverage variations. The consensus on the SNS purification protocol made it possible to apply our method to our samples (K562 cells) and to published data [18] (from four different cell lines), which allows us to compare detection methods on SNS data.

5.1 Sliding windows for the detection of significant read enrichments

OriSeq data analysis based on SNS material consists in detecting significant read enrichments corresponding to accumulations of SNS throughout the human genome. For a given origin, reads accumulate around the initiation starting point with a span determined by the size of the SNS fragments. It is important to notice that SNS are selected based on their size (between 1.5-2kb), and then fragmented and sequenced. Hence, for a given origin the resolution of detection can not be smaller than 1.5-2kb. Tools for the detection of peak-like patterns in ChIP-Seq data, such as SoleSearch [18, 20], have been used for detection purposes, without controlling for the size of the peak, which results in peaks smaller than 1kb on average. In our method we control the resolution of detection by considering sliding windows of size 2kb. We define an appropriate statistical model for discriminating between signal and noise and controlling for false-positive peaks, while accounting for the genome-ordered structure of the data. Read occurrences throughout the genome are supposed to follow a Poisson process $N(t)$ with a heterogeneous intensity $\lambda(t)$ that can be interpreted as the coverage process (Figure 5.2). We also assume that at a given position t along the genome, the number of reads $X(t)$ follows a geometric distribution $\mathcal{G}(p)$. We then consider $R(t) = \sum_{i=1}^{N(t)} X(i)$, which counts the number of reads along the genome and, to detect local exceptional read accumulations, we compute $\Delta R(t,u) = R(t+u) - R(t)$, which quantifies the number of reads within a window of size $u =$ 2kb (Figure 5.3).

To assess the significance threshold of detection we used scanning statistic results for compound distributions, making it possible to calculate the probability of the richest window actually being a false positive [29]. The detection is performed at level α by setting threshold δ_α such that

$$\mathbb{P}\left\{\max_t \left(\Delta R(t,u)\right) > \delta_\alpha\right\} \leq \alpha.$$

Threshold α was calibrated using independent input DNA from public databases since input DNA was not available at the time of the experiment. We applied the detection method to input DNA (ie a dataset with no biological signal) and we assessed the percentage of nucleotides detected as origins in the SNS data that were also detected as peaks in the input DNA data.

To account for coverage heterogeneities, we segment the coverage process (N) into regions of constant intensities (constant λ). We use a segmentation model for this purpose, based on the Poisson distribution adapted from segmentation models we developed for array CGH data analysis [92]. This segmentation step has two main advantages: First, it automatically detects regions of constant coverage (constant λ) and regions with extremely low coverage that are excluded from the study. Second, it allows our significance thresholds to adapt to coverage variations. An example of detection is provided in Figure 5.3. Using the scan method, we detected between 60,000 and 90,000 replication origins (depending on read depth), which cover $\sim$12% of the genome [FP7].

Very interestingly, the field of *OriginOmics* has recently been enriched by another genome-wide map of replication origins obtained by bubble trapping [80], which is based on the sequencing of EcoR1 fragment containing at least one replication bubble. This new map consists of $\sim$125,000 EcoR1 fragments that cover 25% of the human genome. We took this opportunity to confront

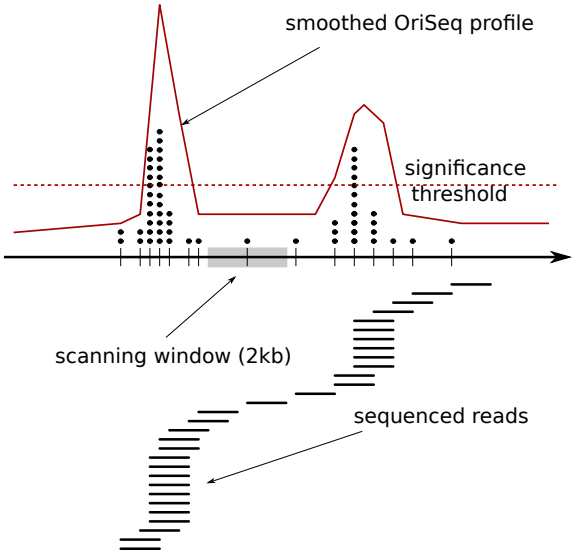


Figure 5.2: Short Nascent Strands are amplified and corresponding reads are aligned along the human genome (bottom).“Ticks” correspond to mapping positions along the genome that constitute the coverage process, modelled by an inhomogeneous Poisson Process. In situations where an origin is shared by many cells, reads will accumulate at the approximate same location along the genome. This is modelled by an accumulation process $(X(t))$. The task of the statistical analysis is to detect significant enrichment along the genome, coming from both coverage and accumulation processes.

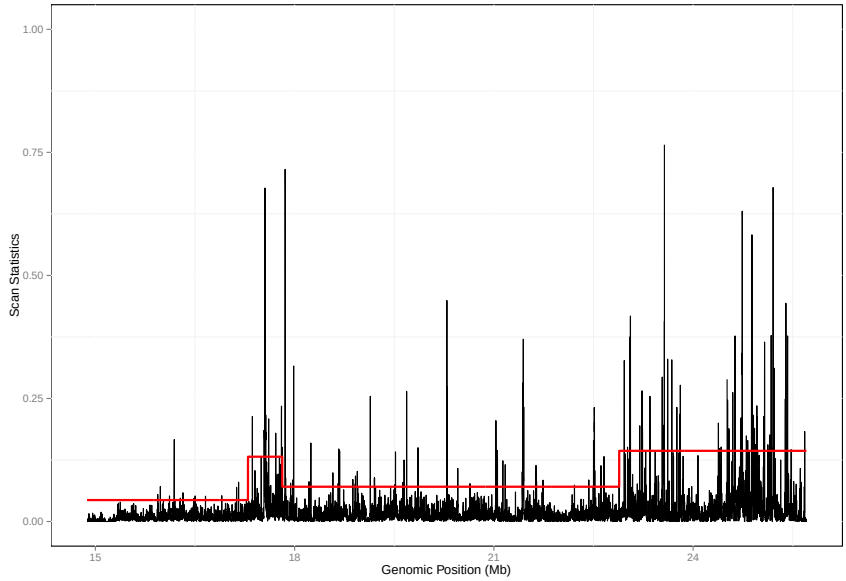


Figure 5.3: Snapshot of peak detection on a region of chromosome 20 from HeLa cells. Scanning windows are represented according to their genomic position. The red line corresponds to the threshold used in the scan methods, which adapts to regional coverage.

different genome-wide detections of replication origins based on different methods and protocols, which had never been done before [FP7]. Comparisons between SNS-based origins and bubble trapping based origins on different cell lines show a good agreement between maps. Furthermore these comparisons (not detailed here) indicate that the sensitivity and specificity of the detection of origins based on SNS data is significantly improved with our dedicated method compared to previously used chIP-Seq tools.

5.2 Poisson functional regression to detect peaks in NGS data

In September 2013 we started a collaboration with V. Rivoirard and our graduate student S. Ivanoff, on Poisson functional regression. The motivation of this new project was to investigate some theoretical questions that have been raised by the analysis of NGS data. Even if the replication origin project was a first motivation, our point was to propose general statistical methodology that could be applied to any aligned-based NGS signal. The first issue we dealt with was the denoising of the read-accumulation data using Poisson functional regression. Our first approach in signal detection relied on a probabilistic modelling of the coverage and the accumulation process, mainly because we wanted to assess a significativity threshold to the accumulation of reads. Here we

propose another model that is inspired from functional data analysis that was developed in Part I. Indeed, OriSeq data (as well as ChIP-Seq) can be modelled as pairs $(X_i, Y_i)_{i=1\dots n}$, where the X_is stand for the genomic location of read i, and $Y_i \in \mathbb{N}$ are independent observations of random Poisson variables that model the number of reads observed at position X_i. We consider the Poisson functional model:

$$Y_i | X_i \sim \mathcal{P}\left(f_0(X_i)\right),$$

with f_0 an unknown function to estimate. Our strategy is to estimate f_0 using the so-called dictionary approach. We start by setting a candidate $\log f$ as a linear combinations of elements of a known finite dictionary $\Upsilon = \{\varphi_j\}_{j\in J}$:

$$\log f(x) = \sum_{j\in J} \beta_j \varphi_j(x),$$

with J a set of cardinal p. The dictionary functions φ_j are can be based on the Haar or Fourier basis. However, the point of the dictionary approach is the ability to choose any combination of these bases according to expected characteristics of the function to estimate. For example, the Haar basis is most efficient if $\log f_0$ is piecewise constant, whereas the Fourier basis is best suited to estimate a regular periodic function. If $\log f_0$ presents both of these aspects then it will be more efficient to combine the Haar and the Fourier basis to catch both aspects of the signal. In the case of SNS data analysis, peaks will be searched by adding Daubechies wavelets, and should the resolution be a matter, focusing on a scale adapted to the length of the replication origins is also feasible.

As estimating f_0 will be equivalent to selecting the vector of regression coefficients $\boldsymbol{\beta} = (\beta_j)_{j\in J}$, f_0 will have a sparse decomposition on this combined dictionary, which limits the accumulation of estimation errors. To proceed we consider the penalized maximum likelihood framework to estimate the coefficients of the model. Denoting by $\mathbf{A}$ the design matrix of size $n \times p$ defined by $A_{ij} = \varphi_j(X_i)$ the log-likelihood associated with this model is, up to a constant,

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{j\in J} \beta_j (\mathbf{A}^T \mathbf{Y})_j - \sum_{i=1}^{n} \exp\Big(\sum_{j\in J} \beta_j A_{ij}\Big),$$

where $\mathbf{A}_j$ is the j-th column of the matrix $\mathbf{A}$. The score function is then

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j} \quad = \quad \mathbf{A}_j^T \big(\mathbf{Y} - \exp\big(\mathbf{A}\boldsymbol{\beta}\big)\big), \tag{5.1}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\exp(\mathbf{A}\boldsymbol{\beta}) = (\exp((\mathbf{A}\boldsymbol{\beta})_1), \dots, \exp((\mathbf{A}\boldsymbol{\beta})_n))^T$. The high-dimensional framework arises when considering rich dictionaries ($p > n$). In this situation there is no unique $\boldsymbol{\beta}$ such that $\mathbf{A}^T\big(\mathbf{Y} - \exp(\mathbf{A}\boldsymbol{\beta})\big) = 0$ and therefore no unique maximizer of the likelihood. The usual way to bypass this issue is to add an ℓ_1-penalty to the log-likelihood which yields the Lasso [118]:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in R^p}{\operatorname{argmin}} \left\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}. \tag{5.2}$$

In many practical applications, parameters β_j can be grouped: if we consider a wavelet dictionary for instance, wavelets coefficients of the same scale can be grouped together so that each scale is selected globally. Therefore we also consider the group-lasso strategy (not detailed here). Then we can derive from the Karush-Kuhn-Tucker conditions [23] that $\widehat{\boldsymbol{\beta}}$ satisfies the so-called Dantzig constraint:

$$\forall j,\ |\mathbf{A}_j^T(\mathbf{Y} - \exp(\mathbf{A}\widehat{\boldsymbol{\beta}}))| \leq \lambda_j. \tag{5.3}$$

Therefore even if we cannot require the score function to be equal to 0, it will be controlled by the λ_j's.

A first aspect of the project is to propose a theoretical derivation of the weights λ_j. This question can be also viewed as the calibration of a thresholding rule for Poisson regression, inspired from the hard and soft thresholding procedures proposed in the Gaussian framework [40]. Early attempts to analyze functional Poissonnian data was to apply variance stabilizing transforms [17] and to consider the transformed-data as Gaussian. However this method is only valid for high intensities (in a regime where the Gaussian approximation is reasonable). The particularity of Poisson regression is that the thresholding weights should adapt to the heteroscedasticity of the model. It would indeed be poorly adapted to choose, say, the universal penalty $\lambda = \sigma\sqrt{2\log p}$, since in the Poisson regression model the errors are necessarily tied to the intensity, as $\mathbb{E}[Y_i|X_i] = \mathbb{V}(Y_i|X_i)$, which makes difficult to consider a direct equivalent to σ^2. Using the penalized likelihood framework and concentration results on continuous Poisson processes, S. Ivanoff proposed theoretical λ_js that are used to control the fluctuations of $\mathbf{A}_j^T\mathbf{Y}$. Consequently a key ingredient to get the weights is to consider the variance $V_j = \mathbb{V}(\mathbf{A}_j^T\mathbf{Y})$ (that equals $\sum_{i=1}^n f_0(X_i)\varphi_j^2(X_i)$ if the model were true). This quantity actually plays an analogous role to σ^2 in our model, and λ_j will be proportional to $\sqrt{\widetilde{V}_j}$ in the main result, with $\widetilde{V}_j$ an overestimation of V_j (see below).

The following theorem gives precise data-driven λ_j's:

Theorem 2 *Let j be fixed and $\gamma > 1$ be a constant. We define $\widehat{V}_j = \sum_{i=1}^n \varphi_j^2(X_i)Y_i$ the natural unbiased estimator of V_j and*

$$\widetilde{V}_j = \widehat{V}_j + \sqrt{2\gamma\log p\widehat{V}_j \max_i \varphi_j^2(X_i)} + 3\gamma\log p \max_i \varphi_j^2(X_i).$$

Let us define

$$\lambda_j = \sqrt{2\gamma\log p\widetilde{V}_j} + \frac{\gamma\log p}{3}\max_i|\varphi_j(X_i)|. \tag{5.4}$$

Then

$$\mathbb{P}\Big(|\mathbf{A}_j^T(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])| > \lambda_j\Big) \leq \frac{3}{p^\gamma}.$$

In his work, S. Ivanoff also provides oracle inequalities to assert the theoretical efficiency of the choice of the weights.

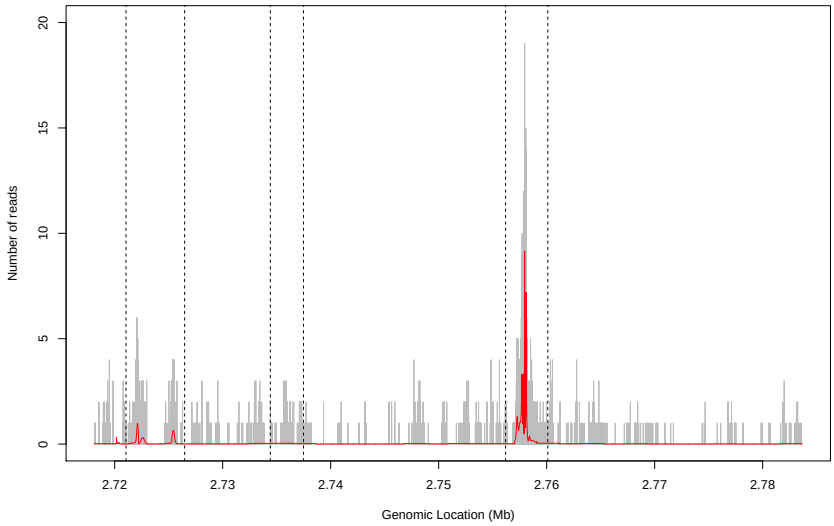


Figure 5.4: OriSeq signal along chromosome X for Hela cells. The grey line represents the number of reads that map each position along the genome, and the red line corresponds to the estimated intensity function using the lasso. The dictionary was chosen by 2-fold cross validation. Vertical dashed lines represent the origins that were detected using scan statistics.

Preliminary application and perspectives. Most effort from now has been put on the theoretical part of this work, and we are currently investigating the empirical performance of the method on simulated data. Our results show that the lasso with theoretical weights is more performant than other methods (like the Haar-Fisz transform [46]) for reconstruction for instance. The great interest in having theoretical derivations for the weights λ_js is that computations do not need heavy ressources to obtain $\widehat{\beta}$ (we use available packages for this purpose [79]). Therefore, available ressources can be dedicated to the choice of the dictionary, that can be performed by cross validation. As a starting application, we considered the OriSeq signal that we analyzed using scan statistics [FP7], and we estimated the intensity function on these data. Figure 5.4 shows that on a portion of chromosome X, the lasso method identifies somes peaks that were detected as replication origins, but can detect smaller peaks as well. One origin is not detected though. These results are very preliminary, but they illustrate how the present method could be used to denoise and detect peaks in NGS data.

5.3 Differential analysis of peak-like data

With the identification of a consensus set of replication origin in the human genome, new questions emerge concerning the regulation of this spatial program. Indeed replication origins must not be fired more than once within the S phase, and forks can frequently stall for instance if they encounter damaged bases [22]. Thus origins of replication are first "licensed" by the formation of a multiprotein complex termed the pre-Replicative Compex (pre-RC) before entry into the S phase. Many licensed replication origins remain dormant and thus constitute a pool of latent origins to survive replication stresses that block replication for movement [50]. Our project is then to study the plasticity in the replication process through the identification of activated dormant origins under replicative stress. We will use ChIP-Seq data to localize the pre-RC in the human genome (and thus to map the dormant origins) and we will confront this map with the map of activated origins to unravel the dynamics of origins licensing.

Very interestingly this project raises methodological questions that may be of very broad interest in the field of NGS data analysis. Indeed, among the application of NGS, ChIP-Seq (chromatin Immuno-precipitation followed by NGS) is probably the most successful to date. Like SNS data, ChIP-Seq data analysis consists in the detection of an exceptional local accumulation in reads with respect to a reference (input data). The result of such procedure is a list of peaks positions along the genome. While the analysis of single-profile data has raised considerable attention [103, 128, 125], the differential analysis of peaks has been marginally studied to date [110, 66]. In the following is presented a new strategy based on continuous Poisson Processes comparison that will constitute a major direction of future research.

Our project is to compare the local densities of peaks between conditions, in order to assess the significance of the difference. In a first step, let us consider two conditions A and B for which we have a collection of peak positions along the genome. We start by modelling these positions by two heterogeneous Poisson Processes denoted by N_A and N_B, with respective intensities λ_A and λ_B.

We will suppose that N_A and N_B lie on $[0,1]$ and that λ_A and λ_B are positive and in $L^1(0,1)$. For some resolution $\eta \in (0,1)$ that will define the resolution of the test (which can be determined by the user for instance), for all $t \in [0,1]$, we aim at finding the intervals $I_\eta(t) = [t-\eta, t+\eta] \cap [0,1]$, such that the intensities of the Poisson Processes are equal. More specifically, let us consider the following testing problem:

$$\mathcal{H}_t : \left\{\lambda_A = \lambda_B = \lambda \text{ on } I_\eta(t)\right\} \text{ against } \mathcal{A}_t : \left\{\lambda_A \neq \lambda_B \text{ on } I_\eta(t)\right\} \tag{5.5}$$

Since N_A and N_B are independent, we can define the joint Poisson Process $N = (N_A, N_B)$ of intensity $\lambda_A + \lambda_B$, along with a Bernoulli process $\boldsymbol{\varepsilon}^0 = \left\{\varepsilon_t^0 \in \{-1,+1\}, t \in N\right\}$. Thus N_A stands for the set of points $t \in N$ such that $\varepsilon_t^0 = 1$ and N_B is the complement of N_A in N.

Kernel-based statistics. To test the local differences between N_A and N_B based on sliding windows, a first possibility can be to compare the number of points coming from A and B within window $I_\eta(t)$ [100]:

$$N_A[I_\eta(t)] = \int_{I_\eta(t)} dN_A(s)ds \text{ and } N_B[I_\eta(t)] = \int_{I_\eta(t)} dN_B(s)ds.$$

However, this strategy does not account for the possible differences in points location that may occur *within* the window. In other words, this strategy is not adaptive to the regularity of intensity functions λ_A and λ_B. Following recent results on Poisson Processes comparison [45] and more generally on the two-sample problem [53] we introduce the kernel-based statistics:

$$\forall t \in [0,1],\ T_{h,\eta}(t, \boldsymbol{\varepsilon}^0) = \sum_{s,s' \in I_\eta(t)} K_h(s,s')\varepsilon_s^0\varepsilon_{s'}^0,$$

with $K_h(s,s')$ a positive symmetric kernel of bandwidth h. For interpretation purposes $T_{h,\eta}$ can be re-written such as :

$$T_{h,\eta}(t, \boldsymbol{\varepsilon}^0) \quad = \quad \sum_{\substack{s,s' \in I_\eta(t) \\ (s,s') \in N_A^2 \cup N_B^2}} K_h(s,s') - \sum_{\substack{s,s' \in I_\eta(t) \\ (s,s') \in N_A \times N_B}} K_h(s,s').$$

In the following, we say that s and s' are h-neighbors when $|s - s'| \leq h$. Then when considering the uniform kernel (i.e. $K_h(s,s') = 1_{|s-s'| \leq h}$), the first term of $T_{h,\eta}(t, \boldsymbol{\varepsilon}^0)$ is the number of h-neighbors coming from the same process (either N_A or N_B), within window $I_\eta(t)$, whereas the second term accounts from h-neighbors from different processes. This motivates the difference of magnitude between resolutions h and η such that $h = o(\eta)$. In the case where $h = \eta$ the statistics resumes to

$$T_{\eta,\eta}(t, \boldsymbol{\varepsilon}^0) \quad = \quad \left(N_A[I_\eta(t)] - N_B[I_\eta(t)]\right)^2.$$

Given N, $(N_A + N_B)[I_\eta(t)]$ is a constant, so that a decision rule based on $T_{\eta,\eta}(t, \boldsymbol{\varepsilon}^0)$ consists in rejecting H_t if $N_A[I_\eta(t)]$ is big/small enough. As already mentioned $T_{\eta,\eta}(t, \boldsymbol{\varepsilon}^0)$ does not consider any information regarding distances between the occurrences of processes N_A and N_B. Consequently introducing smoothing parameter $h = o(\eta)$ is thought to increase the power of the testing procedure.

Distribution of the test statistics under $\mathcal{H}_t$. Considering the joint process along with the associated Bernoulli process $(N, \boldsymbol{\varepsilon}^0)$, given that N_A and N_B are independent we have:

$$\begin{aligned}
\mathbb{P}\{\varepsilon_t^0 = +1 | t \in N\} &= \frac{\lambda_A(t)}{\lambda_A(t) + \lambda_B(t)} \underset{\mathcal{H}_0}{=} \frac{1}{2} \\
\mathbb{P}\{\varepsilon_t^0 = -1 | t \in N\} &= \frac{\lambda_B(t)}{\lambda_A(t) + \lambda_B(t)} \underset{\mathcal{H}_0}{=} \frac{1}{2}
\end{aligned}$$

The key idea of the testing strategy is to work conditionally to the joint process. Indeed, since our objective is to determine whether the two processes are different, by conditionning on N we just have to focus on the labels A and B. As shown in Fromont et al. (2013) [45], conditionally on N, the distribution of $T_{h,\eta}(t, \boldsymbol{\varepsilon}^0)$ under $\mathcal{H}_t$ is equal to the distribution of the bootstrapped version of

$$T_{h,\eta}(t, \boldsymbol{\varepsilon}) = \sum_{s,s' \in I_\eta(t)} K_h(s, s') \varepsilon_s \varepsilon_{s'},$$

with $\boldsymbol{\varepsilon} = \{\varepsilon_t, t \in N\}$ a sequence of Rademacher random variables. This methodology provides a direct way to compute the p-values of the tests. Given the joint process N, generate $\boldsymbol{\varepsilon}^b = \{\varepsilon_t^b; t \in N\}$ as a sequence of *i.i.d.* random variables with $\varepsilon_t^b \sim (\delta_{-1}/2 + \delta_{+1}/2)$. In the following we will use notation $\{\boldsymbol{\varepsilon}^b\}_{B_0} = \{\boldsymbol{\varepsilon}^1, \ldots, \boldsymbol{\varepsilon}^{B_0}\}$ for the sake of simplicity. Thus we consider the resampled p-values defined by

$$\widehat{p}(t, \{\boldsymbol{\varepsilon}^b\}_{B_0}, \boldsymbol{\varepsilon}^0) = \frac{1}{B_0 + 1} \left(1 + \sum_{b=1}^{B_0} \mathbb{1}_{\{T(t,\varepsilon^b) \geq T(t,\varepsilon^0)\}} \right).$$

An illustration of the statistics and the p-value process is provided in Figure 5.5 on simulated data.

Control of the FWER and of the FDR in the continuous setting. An interesting question raised by this testing strategy is the control of the level of the test. One particularity of our approach is that the number of hypothesis tested in not finite as it is usually the case. By considering the continuous Poisson Process framework, we basically suppose that the coordinates of the occurrences (peaks) are intrinsically random over a *continuum* of possible locations. Consequently, the number of tested hypothesis is infinite, and the correction of the multiple testing should be adapted to this "continuous" setting [21]. It generalizes some approaches that we developed recently with hidden

Figure 5.5: Simulated example for the comparison of two Poisson processes. Ticks indicate occurrences of each process with a color for each. Top panel shows the kernel statistics computed with the Gaussian kernel. The statistics increased when the intensities of the processes are likely to be different. Bottom panel shows the p-value process computed on $B_0 = 200,000$ bootstraps. Note that this p-value process is not yet adjusted for continuous multiple testing.

Markov models to account for dependencies among test statistics along the genome [FP6]. Let us denote the (measurable) set of true null hypotheses (the *global* null hypothesis):

$$\mathcal{H}_0 = \{t \in [0,1] \ : \ \lambda_A = \lambda_B \text{ on } I_\eta(t)\}.$$

From the p-values defined above, and for all $\zeta \subseteq [0,1]$ we define the rejection set, for some (potentially data-dependent) threshold u_ζ,

$$R(u_\zeta) = \{t \in [0,1] \ : \ \widehat{p}(t, \{\varepsilon^b\}_{B_0}, \varepsilon_0) \leq u_\zeta\},$$

In our case, $R(u_\zeta)$ corresponds to the intervals on which the global hypothesis is rejected. To evaluate the quality of a threshold u_ζ, we consider the following quantities:

$$\text{FWER}(u_\zeta) = \mathbb{P}\left(\mathcal{H}_0 \cap R(u_\zeta) \neq \emptyset \mid N\right) \tag{5.6}$$

$$\text{FDR}(u_\zeta) = \mathbb{E}\left[\frac{\Lambda\left(\mathcal{H}_0 \cap R(u_\zeta)\right)}{\Lambda\left(R(u_\zeta)\right)} \,\middle|\, N\right], \tag{5.7}$$

where Λ denotes the Lebesgue measure on $[0,1]$ and with the convention $0/0 = 0$. While $\text{FWER}(u_\zeta)$ is the probability to make at least one error in $R(u_\zeta)$, the continuous FDR, $\text{FDR}(u_\zeta)$ is the average proportion of errors in $R(u_\zeta)$. It can be viewed as the cumulated length of intervals corresponding to false rejections, divided by the cumulated length of rejection. From an intuitive point of view, controlling $\text{FWER}(u_\zeta)$ ensures no false positive in $R(u_\zeta)$ with high probability, while controlling $\text{FDR}(u_\zeta)$ allows false positives but in a small proportion w.r.t. the rejection volume $\Lambda\left(R(u_\zeta)\right)$. Hence, in practice, FWER is more stringent than criteria based upon FDR. Our objective is to develop procedures to control these quantities using the Bootstrapped p-values.

Extensions and future projects. We are currently working on the definition of control procedures as detailed above, as well as on an efficient implementation, since the Bootstrap is very computationally intensive. Then a first extension to consider will be to test more than 2 conditions and to consider multiple measurements of the same condition. Since the methodology is based on the joint process, considering multiple treatments and repeats should not be an issue. Another lead will be to adjust the test to the coverage of each condition. Indeed, if peaks of one condition have been called based on an experiment with high coverage, whereas the other condition has lower coverage, the hypothesis to test would rather be: $\lambda_A \propto \lambda_B$.

Then a promising extension will be to consider that when NGS peaks are called, some information remains on their characteristics. For instance, the height of the peak could be an important information to consider. This puts our model in the marked Poisson Process framework, in which occurrences of the peaks N_A, N_B are recorded with some continuous marks Y_A, Y_B such that $Y_A(t)$ is the intensity of the peaks at position $N_A(t)$. The test statistics can be enriched by using a second kernel K^Y that quantifies the distance of the peaks in terms of their intensities:

$$\forall t \in [0,1], \ T_{h,\eta}(t, \varepsilon^0) = \sum_{s,s' \in I_\eta(t)} K_h(s,s') K^Y_{h'}(y_s, y_{s'}) \varepsilon^0_s \varepsilon^0_{s'}.$$

Provided the height of the peak is meaningful for the comparison, this strategy should be more powerful to detect differences between conditions. This mark could be discrete as well to characterize the peaks by some qualitative information. Coming back to copy number analysis, occurrences considered here could be breakpoints positions recorded among individuals, and their characteristics could be the standard copy number status (deleted-normal-amplified). In more general terms, our framework can be adapted to the comparison of any events that can be modelled by ponctual processes like recombination hotspots, breakpoints, motif locations for instance.

Chapter 6

Chromatin landscape of replication origins

Chromatin is the state in which the DNA molecule is packaged within the cell, and the nucleosome is the core of this structure. It is formed by an octamer of four histones (H2A,H2B, H3,H4) around which the DNA molecule is wrapped. This tight association between DNA and histones makes chromatin a very flexible structure, as chemical modifications of histones can modify the physical properties of the chromatin fibers (like its compaction properties and accessibility for instance), which can be associated with different biological processes [13]. Many studies have tried to decipher the roles of histone marks and nucleosomal organization in origin selection. However, our understanding of the complex relationships between chromatin states and replication has been limited by the scale of investigation, as all studies consider replication timing domains of 200 kb to 2 Mb, potentially resulting in a lack of resolution [10, 11, 57, 78]. Nonetheless, it is now well established that genomes are organized into early-, mid- and late-replicating domains[1], and early domains have been shown to be associated with active epigenetic marks such as H3K4me1, 2 and 3, H3K27ac, H3K36me3 and H3K9ac [106]. Moreover, origins of replication have been found embedded within many types of chromatin substrates [24, 47, 96], suggesting that any regulatory effect of chromatin structure would not be homogeneous across replication initiation sites. Now that a consensus set of replication origins has been identified, the time has come to unravel the genomic as well as the epigenetic characteristics that make these particular loci replication origins. In the following we will focus on the connections between the spatial and temporal programs of replication at fine scales.

6.1 Searching for epigenetic characteristics of replication origins

In Picard et al. (2014) [FP7] we started to investigate the connections between replication origins and their chromatinian environment. To proceed we basically computed the fraction of origins that

[1] these domains correspond to genome portions that are replicated in the early-mid or late S phase of the cell cycle

were overlapping with 6 chromatin modifications (H3K4me1, H3K9me3, H3K27me3, H4K20me1, H3K9ac and H2AZ) as provided by public datasets of ChIPseq analysis. These fractions were computed for origins replicated in 6 different timing categories (from early to late-replicated origins), as shown in Figure 6.1. The H4K20 monomethylation mark thought to control origin licensing has been shown to be associated with 50% of origins, this enrichment being significant for origins activated early or in mid-S (Figure 6.1). The dynamic association of replication origins with open chromatin marks, such as H3K9ac, H3K4me3 and H2AZ, was strong (and significant) for origins replicated early in S phase, whereas origins activated in the second part of S phase were less associated with such marks (Figure 6.1). We also found that these marks tended to be absent from late-activated origins. Overall, 64% of origins carried none of these three open chromatin, indicating that most origins may not be directly driven by the presence of open chromatin marks, as previously proposed [96, 24]. The association with heterochromatin marks has been reported to be negatively correlated with replication timing. We, therefore, also investigated two histone marks known to be enriched in facultative and constitutive heterochromatin. Early origins displayed a significant depletion of H3K9me3, whereas late origins were characterized by a significant enrichment in this mark (Figure 6.1). By contrast, we found that origins activated early and in mid-S phase were enriched in H3K27me3, which was thus associated with a large proportion of replication origins (40%) (Figure 6.1).

Once we had elucidated the spatiotemporal interactions between origins and histone modifications, we further investigated whether they were associated with functional effects such as efficiency (Figure 6.1), length and density [FP7]. We first investigated the responses to separate associations, and then studied the effect of combinations of marks. The separate analysis identified H4K20me1 and H3K27me3 as potential regulators of the replication program. When associated with CpG islands (CGIs), origins carrying these marks were characterized by a higher efficiency (Figure 6.1), suggesting that they were associated with a larger number of initiation events. Colocalization with H4K20me1 and H3K27me3 was also associated with a higher density of origins [FP7].

We then characterized the functional responses associated with marks combinations that we identified for early, mid-S phase and late origins. We found that H4K20me1 and open chromatin marks co-localize on 38% of early origins and 16% of for mid S-phase origins, this proportion being increased for CGI origins to 64% of early and 48 % of mid-S phase origins, (Figure 6.2). Moreover, H4K20me1 also colocalized with H3K27me3, particularly in origins activated in mid-S phase (Figure 6.2). These highly frequent colocalizations of marks were associated with different functional responses, as the coupling between H4K20me1 and H3K27me3 was the only combination to be associated with a significant increase in efficiency whatever the timing of replication (Figure 6.2). The presence in $\sim$60% of origins of H4K20me1 or H3K27me3 (or both), and the strong functional responses associated with the colocalization of these marks suggests their potential importance in the control of the human genome replication program.

We further focused on spatial interactions between marks that might characterize the temporal progression of replication. For each origin detected in K562 cells, we considered its linear distance to the closest mark, H2AZ, H4K20me1, H3K27me3, H3K9me3, H3K9ac or H3K4me3. We

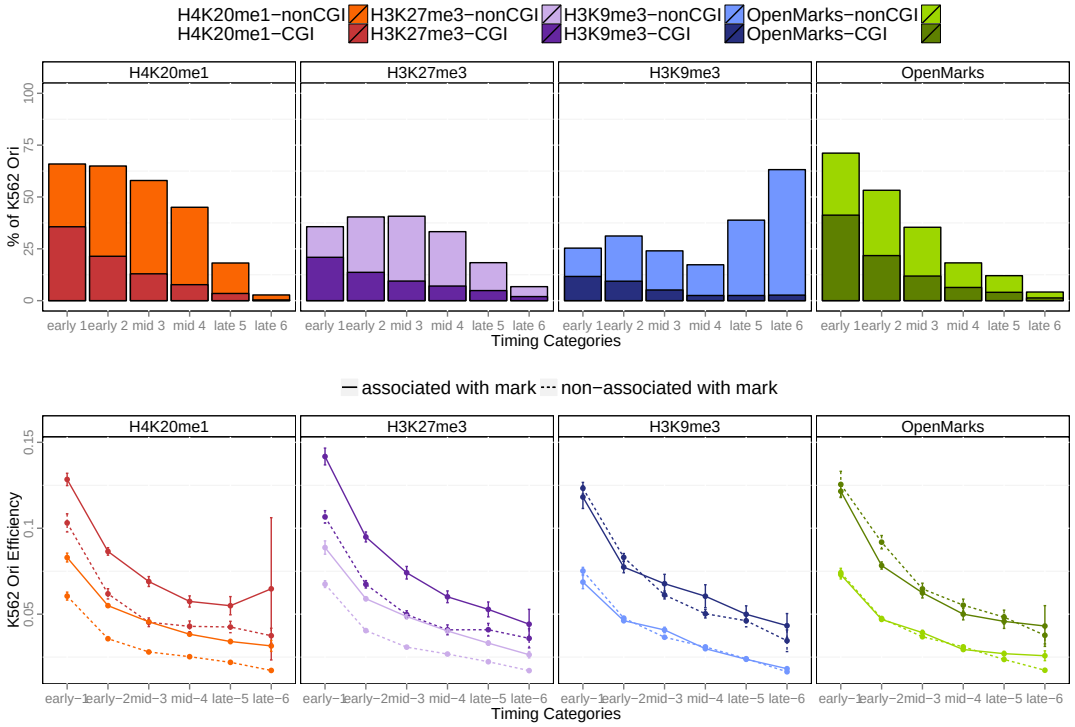


Figure 6.1: Top: Percentage of origins associated with chromatin marks according to timing categories (from early to late origins) and CGI association. Bottom: Variations of origin efficiency (as defined by the number of reads divided by the length of the origin) with replication timing, mark associations, and association with CGIs.

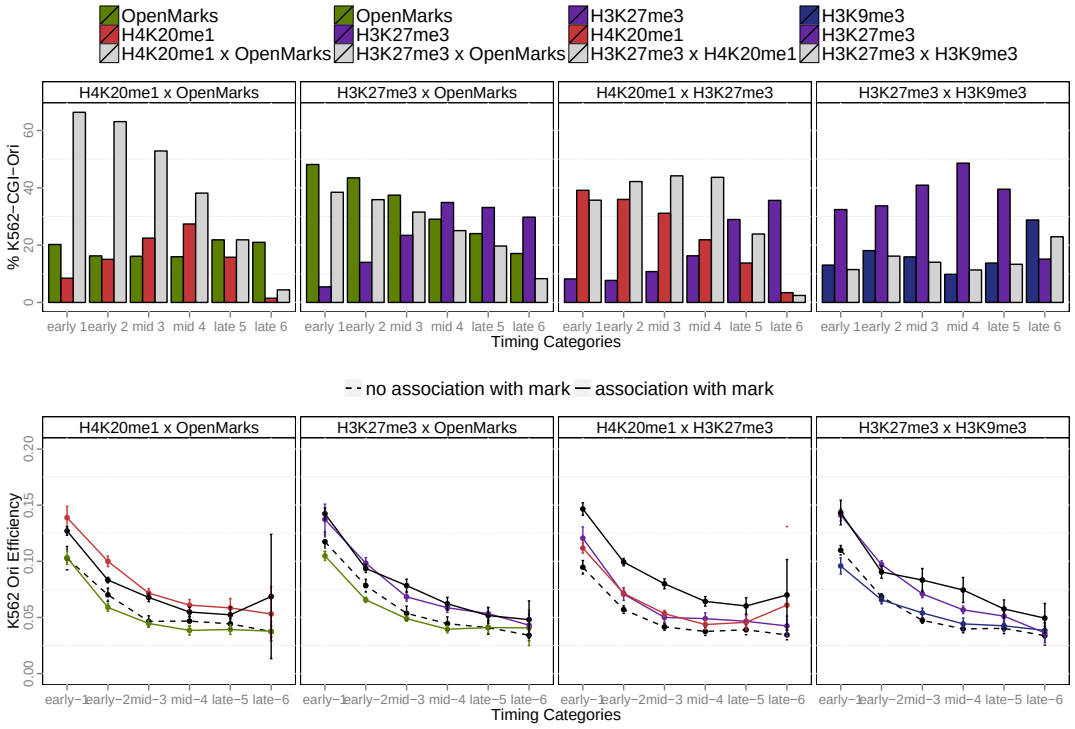


Figure 6.2: Top: Percentage of CGI-origins associated with chromatin marks (K562 cells), according to timing categories (from early to late origins). The colored bars indicate the percentage of CGI-origins associated with one mark only, and grey bars represent the proportion of CGI-origins carrying both marks simultaneously. Bottom: Variations of origin efficiency (as defined by the number of reads divided by the length of the origin) according to replication timing.

then used a discriminant analysis to identify combinations of chromatin marks that could discriminate (and thus characterize) early, mid- and late S-phase origins on the basis of their spatial co-localizations with replication origins. A first combination of marks was characterized by the proximity of early origins to open chromatin marks (H2AZ, H3K9ac and H3K4me3) and H4K20me1. The distance between early origins and open marks increased with the progression of replication, whereas mid-S phase origins remain strongly associated with H4K20me1. Mid-S phase origins were also characterized by a strong association with H3K27me3, and the coupling of H4K20me1 and H3K27me3 with the exclusion of other marks constituted a strong characteristics of this category of origins. Finally the association with H3K9me3 was identified as characteristics of late origins, further from H4K20me1 and H3K27me3.

6.2 Statistical modelling for the integration of epigenomic data

The identification of a "chromatin code" has focused much attention in the community of computational biologists. Several groups have proposed models to integrate many ChIP-Seq signals along the genome, with the extensive use of Hidden Markov Models applied to the study of gene expression regulation [42]. In our project, we will focus on our exceptional data set on replication origins to explore the combination of chromatin marks that characterize the spatiotemporal program of replication. There are currently very few methods to model this interplay between marks, whereas our results indicate that the conjunction of H4K20me1 and H3K27me3 is statistically correlated with increased origin efficiency. In a first step we propose to use multivariate methods to model the interplay between chromatin marks in association with replication origins. The challenge here will be to integrate the spatial information in the model, either the genomic organization (1D) or the 3D nucleus organization using conformation data to unravel the spatiotemporal signatures characterizing replication origins. The link between conformation and replication has long been investigated, as early replicating domains seem to be spatially clustered [106]. Consequently we will pay a particular attention to the integration of conformation data, with the increasing availability of chromosome capture data [73]. A first direction will be to consider mutlivariate methods (like Principal Component Analysis or Partial Least Squares) are particularly suited for this purpose as they provide a compression of the data which allows a synthetic representation / interpretation. Partial Least Squares seems more appropriate in our case since we aim at studying the effect of covariates on a response variable such as the efficiency of replication origins. With S. Lambert-Lacroix we co-supervise a graduate student (G. Durif) on this theme. New developments on functional PLS could be a way to integrate the spatial information into the model [38]. With called ChIP-Seq data being modelled by Poisson point processes functional PLS or PCA for Poisson processed could be a promising direction.

Another modelling direction to study the spatial interactions between replication origins and chromatin marks will be to consider time-varying coefficients models. By studying $Y(t)$ the observed intensity of the replication signal at position t considering p ChIP-Seq signals $X_1, \ldots, X_p$ that all depend on t we will enrich the Poisson functional model described previously to incorporate

genomic covariates such that:

$$\log \mathbb{E}(Y(t)) = \sum_{j=1}^{p} \alpha_j(t) X_j(t).$$

Then function $\alpha_j(t)$ can be decomposed on a dictionary and used to study the interactions between the replication signal and the ChIP-Seq signals. One challenging question will be to investigate the interaction between functional regressors that may explain the spatial variations of the replication signal. This strategy will be based on non-called data (raw ChIP-Seq signals), but we will also consider the modelling of called data. This lead concerns the use of multivariate Hawkes models which constitutes a very powerful framework to model the interactionsbetween point processes. This strategy has already been considered to unvarel interactions between transcription factors [25], and recent theoretical developments [58] will allow us to consider sparse versions of such models.

Concluding remarks

Statistics has this particularity of being a "useful" discipline which may partly explain its extremely open spectra, from (coffee-)talking to other scientists to build an appropriate statistical model, to computational statistics, optimization, software development and to the demonstration of mathematical results. My research is part of the "applied statistics" field. I have been interested in the construction of statistical models based on biological expertise (around a coffee), and I developed computational strategies and software packages for biologists to use them. A constant course of action has been to use modern statistics on modern biological issues. This concern has required the constant collaboration with biologists on one hand, and also with statisticians interested in theoretical aspects on the other. This reminds me a quote my former PhD advisor J.-J. Daudin who once said: "On est toujours l'appliqué de quelqu'un", "One is always somebody's applied guy", which always reminds me that the "applied" term is somehow merely relative.

From the "application" point of view, a first strategic decision to make when working on genomic data is to determine if we should work on raw data or not. Working on raw data is often close to signal processing, and involves the detection of particular patterns in the signal (like peaks or breakpoints) among technical and biological noise. To this respect it can be viewed as a denoising procedure that catches some information and results in called data (peaks position and height, breakpoint positions and status for instance). Given the number of published methods and pipelines it becomes crucial to wonder if it is worth publishing for instance a 100th method to detect peaks in ChIP-Seq data with the risk of being in limbo of Bioconductor packages, or if we should work with the results of published method (or with their aggregation). To this end, the varying coefficient model may offer a powerful framework that may justify its application to ChIP-Seq data analysis to account for genomic covariates. For these reasons I also choose to investigate point process models to work on "called" data (like peak positions). Once again, modern statistical developments such as the use of the lasso for intensity estimation, or the FDR in continuous time are very promising when applied to genomic data, and they allow answering the very difficult question of peak positions comparison and testing. Moreover, Poisson point process models being easily generalized to any dimension, they may be a very powerful framework to handle 3D conformation data that are increasingly available.

Another particularity of applied statistics in multidisciplinary research is also the notion of "results": the proposition of a model associated with an estimation framework and software for instance does not constitute a *result* of the same nature as the demonstration of a theorem for instance. Thus my concern has often been to determine what *works* and what does not, and what *will* work or will not (despite the fact that what people use is sometimes correlated with what is published first and not necessarily with what actually works). I would like to stand back, and to wonder if functional model will work on genomic data. That I can not say for sure, but I think that the functional framework is appropriate to account for the genomic organization of some data, and offers a great flexibility. It is associated with non-parametric statistics that may be less specific compared to parametric models that allow the integration of *prior* expertise. Nevertheless, they allow to model very complex phenomena, such as the inter-variability of genome-ordered data. Moreover, the association with penalized estimation makes the inference framework very powerful from the computational point of view, which has become an important constraint with the size of current genomic data.

From a personal perspective, I find the most challenging task is to choose the direction on which *research* should be done, since in many cases, simple tools can be efficient and our work becomes close to engineering. It is often a way to learn how biologists work, their habits, which is very fruitful at the end when the purpose is to propose them some tools. With the increasing importance of "mapped" data, *ie* data that are analyzed through their genomic organization, my choice has been to look into strategies that account for this structure, through functional models and point process models, along with the use of hidden variable models like mixture models or mixed models. Penalized estimation has been another constant of my work, to get parcimonious models with interpretable parameters. This has raised the recurrent question of penalty calibration for which there is not necessarily a consensus method. To this respect, the theoretical developments on the Poisson functional model have been very powerful in practice.

Given the increasing complexity of the genomic data we are facing, it has become important to develop different modelling strategies in order to be flexible when new data and questions arise. This requires constant discussions with statisticians involved in more theoretical aspects, as Statistics has also evolved a lot during the past years. This is why I thought this document was a "tour", just to visit some possibilities to model and handle genomic data.

$\star$

Franck Picard Bibliography (2006-2014)

[FP1] J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008.

[FP2] M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1):31–40, 2013.

[FP3] O. Mestre, P. Domonkos, F. Picard, I. Auer, S. Robin, E. Lebarbier, R. Boehm, E. Aguilar, J. Guijarro, G. Vertachnik, et al. Homer: a homogenization softwaremethods and applications. *Idojaras, Quarterly journal of the Hungarian Meteorological Service*, 117(1):47–67, 2013.

[FP4] V. Miele, S. Penel, V. Daubin, F. Picard, D. Kahn, and L. Duret. High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics*, 28(8):1078–1085, Apr 2012.

[FP5] V. Miele, F. Picard, and S. Dray. Spatially constrained clustering of ecological networks. *Methods in Ecology and Evolution*, 5:771–779, 2014.

[FP6] L. Modolo, F. Picard, and E. Lerat. A new genome-wide method to track horizontally transferred sequences: application to drosophila. *Genome Biology and Evolution*, 2014.

[FP7] F. Picard, J. C. Cadoret, B. Audit, A. Arneodo, A. Alberti, C. Battail, L. Duret, and M. N. Prioleau. The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. *PLoS Genet.*, 10(5):e1004282, May 2014.

[FP8] F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin. Assessing the exceptionality of network motifs. *J. Comp. Biol*, 15(1):1–20, 2008.

[FP9] F. Picard, E. Lebarbier, E. Budinska, and S. Robin. Joint segmentation of multivariate gaussian processes using mixed linear models. *Computational Statistics & Data Analysis*, 55(2):1160 – 1170, 2011.

[FP10] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428, Jul 2011.

[FP11] F. Picard, V. Miele, J. J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, 10 Suppl 6:S17, 2009.

[FP12] G. Rigaill, V. Miele, and F. Picard. *Computational Intelligence Methods for Bioinformatics and Biostatistics*. 79. Springer, 2014.

[FP13] M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform*, Feb 2010.

[FP14] T. Vernoux, G. Brunoud, E. Farcot, V. Morin, H. Van den Daele, J. Legrand, M. Oliva, P. Das, A. Larrieu, D. Wells, Y. Guedon, L. Armitage, F. Picard, S. Guyomarc'h, C. Cellier, G. Parry, R. Koumproglou, J. H. Doonan, M. Estelle, C. Godin, S. Kepinski, M. Bennett, L. De Veylder, and J. Traas. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Mol. Syst. Biol.*, 7:508, 2011.

[FP15] V. Viallon, S. Lambert-Lacroix, H. Hoefling, and F. Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, pages n/a–n/a, 2014.

[FP16] H. Zanghi, F. Picard, V. Miele, and C. Ambroise. Strategies for online inference of model-based clustering in large and growing networks. *Annals of Applied Statistics*, 4(2):687–714, 2010.

Bibliography

[1] F. Abramovich, T. Sapatinas, and B.W. Silverman. Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society Series B Stat Methodol*, 60:725–749, 1998.

[2] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[3] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *R. Modern Physics*, 74(1):47–97, 2002.

[4] U. Amato and T. Sapatinas. Wavelet Shrinkage Approaches to Baseline Signal Estimation from Repeated Noisy Measurements. *Advances and Applications in Statistics*, 51:21–50, 2005.

[5] C. Ambroise and G. Govaert. Convergence of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters*, 19(10):919–927, 1998.

[6] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–955, 2001.

[7] A. Antoniadis and T. Sapatinas. Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, 51(10):4793–4813, 2007.

[8] Miguel B. Araujo, Alejandro Rozenfeld, Carsten Rahbek, and Pablo A. Marquet. Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, 34(6):897–908, 2011.

[9] M. Assunção, M. Corrêa Neves, G. Câmara, and C. da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.

[10] B. Audit, A. Baker, C. L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes, and A. Arneodo. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nat Protoc*, 8(1):98–110, Jan 2013.

[11] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton Carafa, C. Thermes, and A. Arneodo. Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res.*, 37(18):6064–6075, Oct 2009.

[12] C. A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–179, Jul 2013.

[13] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Res.*, 21(3):381–395, Mar 2011.

[14] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[15] E. Ben-Yaacov and YC. Eldar. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):139–145, 2008.

[16] H. Bensmail, B. Aruna, O. J. Semmes, and A. Haoudi. Functional clustering algorithm for high-dimensional proteomics data. *J. Biomed. Biotechnol.*, 2005:80–86, Jun 2005.

[17] Panagiotis Besbeas, Italia De Feis, and Theofanis Sapatinas. A comparative simulation study of wavelet shrinkage estimators for Poisson counts. *Intern. Statist. Review*, 72(2):209–237, 2004.

[18] E. Besnard, A. Babled, L. Lapasset, O. Milhavet, H. Parrinello, C. Dantec, J. M. Marin, and J. M. Lemaitre. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.*, 19(8):837–844, Aug 2012.

[19] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE PAMI*, 22(7):719–725, 2000.

[20] K. R. Blahnik, L. Dou, H. O'Geen, T. McPhillips, X. Xu, A. R. Cao, S. Iyengar, C. M. Nicolet, B. Ludascher, I. Korf, and P. J. Farnham. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, 38(3):e13, Jan 2010.

[21] Blanchard, Gilles and Delattre, Sylvain and Roquain, Etienne. Testing over a continuum of null hypotheses with false discovery rate control. *Bernoulli*, 20(1):304–333, 02 2014.

[22] J. J. Blow and X. Q. Ge. Replication forks, chromatin loops and dormant replication origins. *Genome Biol.*, 9(12):244, 2008.

[23] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin : Springer-Verlag, 2011.

[24] J. C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, and M. N. Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, 105(41):15837–15842, Oct 2008.

[25] L. Carstensen, A. Sandelin, O. Winther, and N. R. Hansen. Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC Bioinformatics*, 11:456, 2010.

[26] NP. Carter. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genetics*, 39:S16–S21, 2007.

[27] H. Caussinus and O. Mestre. Detection and correction of artificial shifts in climate series. *JRSS-C*, 53(3):405–425, 2004.

[28] C. Cayrou, D. Gregoire, P. Coulombe, E. Danis, and M. Mechali. Genome-scale identification of active DNA replication origins. *Methods*, 57(2):158–164, Jun 2012.

[29] HP. Chan and NR. Zhang. Scan statistics with weighted observations. *Journal of the American Statistical Association*, 102:595–602, 2007.

[30] S.-S. Chen, D.-L. Donoho, and M.-A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, January 2001.

[31] Z. Chen and D.B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59:762–769, 2003.

[32] S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van de Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavare, J. D. Brenton, B. Ylstra, and C. Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, 8:R215, 2007.

[33] J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25(3):417–418, Feb 2009.

[34] A. Cleynen and E. Lebarbier. Segmentation of the poisson and negative binomial rate models: a penalized estimator. *ESAIM PS*, accepted, 2014.

[35] M.R.T. Dale and M.-J. Fortin. From graphs to spatial graphs. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):21–38, 2010.

[36] Wesley Dáttilo, Paulo R. Guimarães, and Thiago J. Izzo. Spatial structure of ant-plant mutualistic networks. *Oikos*, 122(11):1643–1648, November 2013.

[37] J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008.

[38] Aurore Delaigle and Peter Hall. Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352, 02 2012.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[40] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[41] J. E. Eckel-Passow, A. L. Oberg, T. M. Therneau, and H. R. Bergen. An insight into high-resolution mass-spectrometry data. *Biostatistics*, 10:481–500, Jul 2009.

[42] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, Mar 2012.

[43] M. Frazier, B. Jawerth, and G. Weiss. *Littlewood-Paley Theory and the Study of function Spaces*. 79. American Mathematical Society, 1991.

[44] J. Fridlyand, A. M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. M. Ljung, A. N. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. W. Gray, F. Waldman, D. Pinkel, and D. G. Albertson. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96, 2006.

[45] Magalie Fromont, Batrice Laurent, and Patricia Reynaud-Bouret. The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461, 06 2013.

[46] Piotr Fryzlewicz and Guy P. Nason. A Haar-Fisz algorithm for Poisson intensity estimation. *J. Comput. Graph. Statist.*, 13(3):621–638, 2004.

[47] S. Gay, A. M. Lachages, G. A. Millot, S. Courbet, A. Letessier, M. Debatisse, and O. Brison. Nucleotide supply, not local histone acetylation, sets replication origin usage in transcribed regions. *EMBO Rep.*, 11(9):698–704, Sep 2010.

[48] S. A. Gerbi and A. K. Bielinsky. Replication initiation point mapping. *Methods*, 13(3):271–280, Nov 1997.

[49] M. Giacofci. *Classification non supervisée et sélection de variables dans les modèles mixtes fonctionnels. Applications à la biologie moléculaire.* PhD thesis, Université de Grenoble, 2013.

[50] D. M. Gilbert. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat. Rev. Genet.*, 11(10):673–684, Oct 2010.

[51] D. M. Gilbert. Replication origins run (ultra) deep. *Nat. Struct. Mol. Biol.*, 19(8):740–742, Aug 2012.

[52] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826, Jun 2002.

[53] Gretton, A. and Borgwardt, K. M. and Rasch, M. J. and Schölkopf, B. and Smola, A. A kernel method for the two-sample problem. *J. Mach. Learn. Res.*, 1:1–10, 2008.

[54] M. Guedj, L. Marisa, A. de Reynies, B. Orsetti, R. Schiappa, F. Bibeau, G. MacGrogan, F. Lerebours, P. Finetti, M. Longy, P. Bertheau, F. Bertrand, F. Bonnet, A. L. Martin, J. P. Feugeas, I. Bieche, J. Lehmann-Che, R. Lidereau, D. Birnbaum, F. Bertucci, H. de The, and C. Theillet. A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9):1196–1206, Mar 2012.

[55] R. Guimera and L.A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[56] D. Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7):801–823, 2008.

[57] J. L. Hamlin, L. D. Mesner, O. Lar, R. Torres, S. V. Chodaparambil, and L. Wang. A revisionist replicon model for higher eukaryotic genomes. *J. Cell. Biochem.*, 105(2):321–329, Oct 2008.

[58] NR. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, accepted:na, 2014.

[59] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller. Processing and classification of protein mass spectra. *Mass Spectrom Rev*, 25:409–449, 2006.

[60] T. D. Hocking, G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, and J. P. Vert. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14:164, 2013.

[61] H. Höfling, H. Binder, and M. Schumacher. A coordinate-wise optimization algorithm for the Fused Lasso. *Arxiv preprint arXiv:1011.6409*, 2010.

[62] S.-Y. Huang and H. Horng-Shing Lu. Extended gaussmarkov theorem for nonparametric mixed-effects models. *Journal of Multivariate Analysis*, 76(2):249 – 266, 2001.

[63] P. Hupe, N. Stransky, JP. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, 2004.

[64] L. Jacob, P. Neuvial, and S. Dudoit. More power via graph-structured tests for differential expression of gene networks. *The Annals of Applied Statistics*, 6(2):561–600, 2012.

[65] G. James and C. Sugar. Clustering for sparsely sampled functional data. *Journal of the Americal Statistical Association*, 98:397–408, 2003.

[66] H. Ji, X. Li, Q. F. Wang, and Y. Ning. Differential principal component analysis of ChIP-seq. *Proc. Natl. Acad. Sci. U.S.A.*, 110(17):6789–6794, Apr 2013.

[67] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

[68] Steven T Kalinowski. Evolutionary and statistical properties of three genetic distances. *Molecular Ecology*, 11:1263–1273, August 2002.

[69] D. Keller, R. Holderegger, and M. J. van Strien. Spatial scale affects landscape genetic analysis of a wetland grasshopper. *Molecular Ecology*, 22(9):2467–2482, May 2013.

[70] Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[71] Ann E Krause, Kenneth A Frank, Doran M Mason, Robert E Ulanowicz, and William W Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, November 2003.

[72] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P. J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770, 2005.

[73] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[74] J. Liu, J. Mohammed, J.and Carter, S. Ranka, T. Kahveci, and M. Baudis. Distance-based clustering of CGH data. *Bioinformatics*, 22(16):1971–1978, 2006.

[75] M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, 4:715–742, 2010.

[76] J.C. Marioni, N.P. Thorne, and S. Tavare. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, 2006.

[77] P. McCullagh and J. A. Nelder. *Generalized Linear Models. 2nd ed.* New-York : Chapman & Hall, 1989.

[78] A. J. McNairn and D. M. Gilbert. Epigenomic replication: linking epigenetics to DNA replication. *Bioessays*, 25(7):647–656, Jul 2003.

[79] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008.

[80] L. D. Mesner, V. Valsakumar, M. Cieslik, R. Pickin, J. L. Hamlin, and S. Bekiranov. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.*, 23(11):1774–1788, Nov 2013.

[81] Yann Moalic, Daniel Desbruyères, Carlos M Duarte, Alejandro F Rozenfeld, Charleyne Bachraty, and Sophie Arnaud-Haond. Biogeography revisited with network theory: Retracing the history of hydrothermal vent communities. *Systematic Biology*, 61(1):127–137, 2012.

[82] J. S. Morris, K. A. Baggerly, H. B. Gutstein, and K. R. Coombes. Statistical contributions to proteomic research. *Methods Mol. Biol.*, 641:143–166, 2010.

[83] J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B Stat Methodol*, 68:179–199, 2006.

[84] Guy Nason. *Wavelet methods in statistics with R*. Springer, 2010.

[85] M.E. Newman and E.A. Leicht. Mixture models and exploratory analysis in networks. *PNAS*, 104(23):9564–9569, 2007.

[86] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *PNAS*, 99:2566–2572, 2002.

[87] K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *JASA*, 96(455):1077–1087, 2001.

[88] AB. Olshen, ES. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[89] P. J. Park. Experimental design and data analysis for array comparative genomic hybridization. *Cancer Invest.*, 26:923–928, 2008.

[90] D. M. Pavlovic, P. E. Vertes, E. T. Bullmore, W. R. Schafer, and T. E. Nichols. Stochastic Blockmodeling of the Modules and Core of the Caenorhabditis elegans Connectome. *PLoS ONE*, 9(7):e97584, 2014.

[91] Miguel Pereira, Pedro Segurado, and Nuno Neves. Using spatial network structure in landscape management and planning: A case study with pond turtles. *Landscape and Urban Planning*, 100(1-2):67–76, 2011.

[92] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27, 2005.

[93] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*, 6:27, 2005.

[94] F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation/clustering model of the analysis of array CGH data. *Biometrics*, 63(3):758–766, 2007.

[95] R. Pique-Regi, A. Ortega, and S. Asgharzadeh. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics*, 25:1223–1230, 2009.

[96] M. N. Prioleau, M. C. Gendron, and O. Hyrien. Replication of the chicken beta-globin locus: early-firing origins at the 5' HS4 insulator and the rho- and betaA-globin genes show opposite epigenetic modifications. *Mol. Cell. Biol.*, 23(10):3536–3549, May 2003.

[97] PM. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. Bayesian DNA copy number analysis. *BMC Bioinformatics*, 10(10):1–19, 2009.

[98] F. Rapaport, E. Barillot, and J. P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–382, Jul 2008.

[99] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, April 2010.

[100] S. Robin, S. Schbath, and V. Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics*, 8:84, 2007.

[101] S. Robin and V. T. Stefanov. Simultaneous occurrences of runs in independent Markov chains. *Methodology and Computing in Applied Probability*, 11(2):267–275, 2009.

[102] C. Rouveirol, N. Stransky, Ph. Hupé, Ph. La Rosa, E. Viara, E. Barillot, and F. Radvanyi. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, 22(7):849–856, 2006.

[103] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27(1):66–75, Jan 2009.

[104] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick,

D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.

[105] O. M. Rueda and R. Diaz-Uriarte. Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics*, 10:308, 2009.

[106] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, 20(6):761–770, Jun 2010.

[107] S. P. Shah. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet. Genome Res.*, 123:343–351, 2008.

[108] D.B. Sharma, H.D. Bondell, and H.H. Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22:319–340, 2013.

[109] Y. She. Sparse regression with exact clustering. *Electron. J. Statist.*, 4:1055–1096, 2010.

[110] L. Shen, N. Y. Shao, X. Liu, I. Maze, J. Feng, and E. J. Nestler. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE*, 8(6):e65598, 2013.

[111] Thomas Smaltschinski, Ute Seeling, and Gero Becker. Clustering forest harvest stands on spatial networks for optimised harvest scheduling. *Annals of Forest Science*, 69(5):651–657, 2012.

[112] A. M. Snijders, N. Nowak, R. Segraves, S. Blakwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A.N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29:263–264, 2001.

[113] S. Stjernqvist, T. Ryden, M. Skold, and J. Staaf. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23(8):1006–1014, 2007.

[114] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[115] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[116] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B.*, 67:91–108, 2005.

[117] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, Jan 2008.

[118] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[119] Dean Urban and Timothy Keitt. Landscape connectivity: a graph-theoretic perspective. *Ecology*, 82(5):1205–1218, 2001.

[120] M. Van de Wiel, R. Brosens, P.H.C. Eilers, C. Kumps, G.A. Meijer, B. Menten, E. Sistermans, F. Speleman, M. E. Timmerman, and B. Ylstra. Smoothing waves in array CGH tumor profiles. *Bioinformatics*, 25:1099–1104, 2009.

[121] M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform*, Feb 2010.

[122] MA. Van de Wiel, KI. Kim, SJ. Vosse, WN. van Wieringen, SM. Wilting, and B. Ylstra. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23(7):892–894, 2007.

[123] D.A. van Dyk. Fitting mixed-effects models using efficient EM-type algorithms. *Jour. Comp. and Graph. Statistics*, 9:78–98, 2000.

[124] W. N. Van Wieringen, M. A. Van De Wiel, and B. Ylstra. Weighted clustering of called array CGH data. *Biostatistics*, 9:484–500, Jul 2008.

[125] E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7):e11471, 2010.

[126] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

[127] N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

[128] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.

[129] H. Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.