Adaptive sparse Poisson functional regression for the analysis of NGS Data.

S. Ivanoff[‡], F. Picard[⋆], V. Rivoirard[‡]

[‡]CEREMADE, Univ. Paris Dauphine,F-75775 Paris, France
[⋆]LBBE, Univ. Lyon 1, F-69622 Villeurbanne, France

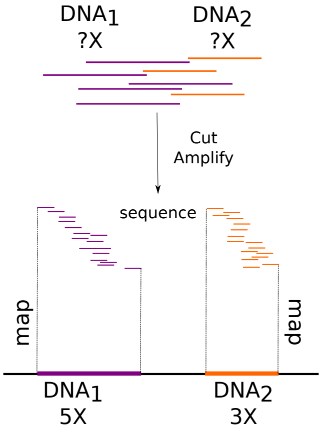
April 2016

Outline

1. Studying replication by high throughput sequencing
2. Poisson functional regression
3. Calibration of the Lasso weights
4. Simulation study
5. Application on OriSeq data

Next Generation Sequencing Data

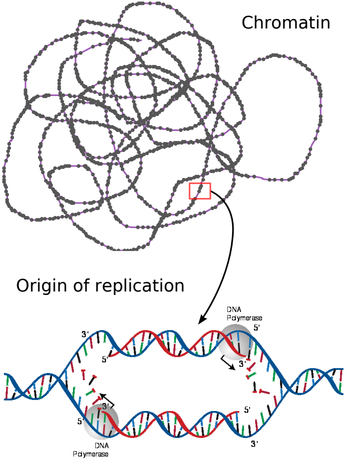
- Massive parallel sequencing of DNA molecules
- Can be used to quantify DNA in a sample
- Expression, copy numbers, DNA-prot. interactions
- Focus on mapped data



Outline of the OriSeq project

- DNA replication: duplication of 1 molecule into 2 daughter molecules
- The exact duplication of mammalian genomes is strongly controlled
- Spatial control (loci choice)
- Temporal control (firing timing)

⇒ What are the (epi)genetic determinants of these controls ?

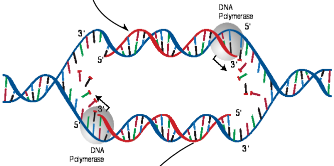


Mapping human replication origins: a technical challenge

- "bubbles" are small and instable (last only minutes by cycle)
- no clear consensus sequence (like in S. Cerevisiae)
- their specification is associated with both DNA sequence and chromatin structure

⇒ *Origin*-Omics: SNS Sequencing

Short Nascent Strand (SNS)



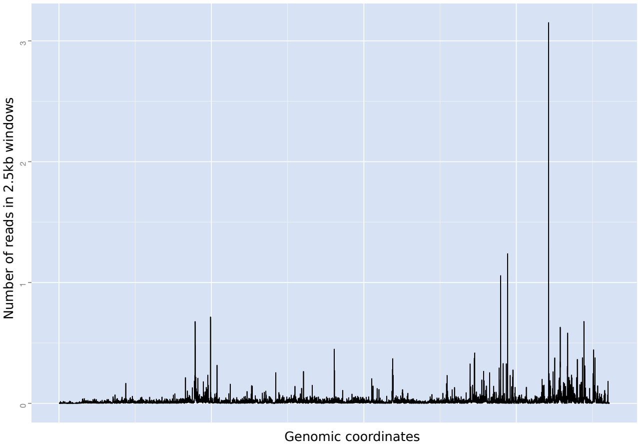
Extraction & Purification

Selection of 1.5-2kb SS fragments

lambda exonuclease Digestion

- qPCR analysis (local)
- DNA tiling arrays
- Sequencing (Ori-Seq)

Example of OriSeq data



Outline

Introduction to Poisson functional Regression

- Y_t the observed number of reads at position X_t on the genome, with $t = 1, \ldots, n$.
- Model sequencing data by functional Poisson regression

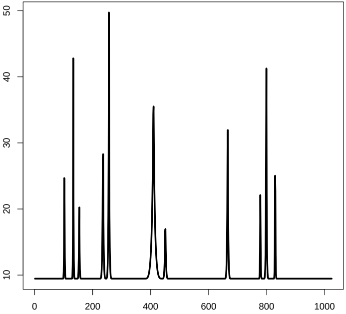
$$Y_t | X_t \sim \mathcal{P}\left(f_0(X_t)\right),$$

- **Goal**: estimate f_0.
- f a candidate estimator of f_0, decomposed on a functional dictionary with p elements $\{\varphi_1, \ldots, \varphi_p\}$:

$$\log f(x) = \sum_{j=1}^{p} \beta_j \varphi_j(x)$$

Dictionaries vs. basis approaches

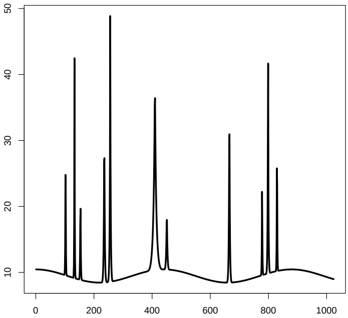
- The basis approach is designed to catch specific features of the signal ($p = n$)
- If many features are present simultaneously ?
- Consider overcomplete dictionaries ($p > n$)
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier



How to select the dictionary elements ?

Dictionaries vs. basis approaches

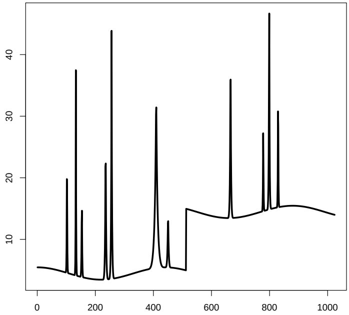
- The basis approach is designed to catch specific features of the signal ($p = n$)
- If many features are present simultaneously ?
- Consider overcomplete dictionaries ($p > n$)
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier



How to select the dictionary elements ?

Dictionaries vs. basis approaches

- The basis approach is designed to catch specific features of the signal ($p = n$)
- If many features are present simultaneously ?
- Consider overcomplete dictionaries ($p > n$)
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier



How to select the dictionary elements ?

A Penalized likelihood framework

- We consider a likelihood-based penalized criterion to select $\boldsymbol{\beta}$,
- We denote by $\mathbf{A}$ the $n \times p$-design matrix with $A_{ij} = \varphi_j(X_i)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$
- The log-likelihood of the model is:

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

- Selection can be performed by the lasso such that:

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

Outline

1. Studying replication by high throughput sequencing
2. Poisson functional regression
3. Calibration of the Lasso weights
4. Simulation study
5. Application on OriSeq data

Weights calibration using concentration inequalities

- Gaussian framework with noise variance σ^2, weights for the Lasso $\propto \sigma\sqrt{\log p}$
- λ_j is used to control the fluctuations of $\mathbf{A}_j^T \mathbf{Y}$ around its mean,
- key role of V_j, a variance term (the analog of σ^2) defined by

$$V_j = \mathbb{V}(\mathbf{A}_j^T \mathbf{Y}) = \sum_{i=1}^{n} f_0(X_i)\varphi_j^2(X_i).$$

- For any j, we choose a data-driven value for λ_j as small as possible so that with high probability, for any $j \in \{1, ...p\}$,

$$\mathbf{A}_j^T(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])| \leq \lambda_j.$$

Form of the data-driven weights for the Lasso

- Let j be fixed and $\gamma > 0$ be a constant. 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

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

- then

$$\Pr\left(|\mathbf{A}_j^T(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])| \geq \lambda_j\right) \leq \frac{3}{p^\gamma}.$$

Using the group Lasso for Poisson Functional Regression

- In some situations, coefficients can be grouped:

$$\{1,\ldots,p\} = G_1 \cup \ldots \cup G_K$$

- For wavelets: group by scales for instance (or adjacent positions)
- In this case selection can be performed by the group-Lasso such that:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{-\log\mathcal{L}(\boldsymbol{\beta}) + \sum_{k=1}^{K}\lambda_k\|\beta_{G_k}\|_2\right\}.$$

- The block ℓ_1-norm on $\mathbb{R}^p$ is defined by:

$$\|\boldsymbol{\beta}\|_{1,2} = \sum_{k=1}^{K}\|\beta_{G_k}\|_2 = \sum_{k=1}^{K}\sqrt{\sum_{j\in G_k}|\beta_j|^2}$$

Form of the data-driven weights for the group-Lasso

- λ_k^g should depend on sharp estimates of the variance parameters $(V_j)_{j \in G_k}$
- Let $k \in \{1, \ldots, K\}$ be fixed and $\gamma > 0$ be a constant. Assume that there exists $M > 0$ such that for any x, $|f_0(x)| \leq M$.
- Let

$$c_k = \sup_{\mathbf{x} \in \mathbb{R}^n} \frac{\|\mathbf{A}_{G_k} \mathbf{A}_{G_k}^T \mathbf{x}\|_2}{\|\mathbf{A}_{G_k}^T \mathbf{x}\|_2}.$$

- For all $j \in G_k$, still with $\widehat{V}_j = \sum_{i=1}^n \varphi_j^2(X_i) Y_i$, define

$$\begin{aligned} \widetilde{V}_j^g &= \widehat{V}_j + \sqrt{2(\gamma \log p + \log |G_k|) \widehat{V}_j \max_i \varphi_j^2(X_i)} \\ &+ 3(\gamma \log p + \log |G_k|) \max_i \varphi_j^2(X_i). \end{aligned}$$

Form of the data-driven weights for the group-Lasso

- Let $\gamma > 0$ be fixed. Define $b_k^i = \sqrt{\sum_{j \in G_k} \varphi_j^2(X_i)}$ and $b_k = \max_i b_k^i$. Finally, we set

$$\lambda_k^g = \left(1 + \frac{1}{2\sqrt{2\gamma \log p}}\right) \sqrt{\sum_{j \in G_k} \widetilde{V}_j^g} + 2\sqrt{\gamma \log p \, D_k},$$

- $D_k = 8Mc_k^2 + 16b_k^2 \gamma \log p$.

$$\Pr\left(\|\mathbf{A}_{G_k}^T(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])\|_2 \geq \lambda_k^g \right) \leq \frac{2}{p^\gamma}.$$

- The form of the weights is analog to the weights in the Gaussian setting
- We show that the associated Lasso / Group Lasso procedure are theoretically optimal (oracle inequalities).
- With a theoretical form for the weights, much computing power is spared !

Outline

Simulation settings

- We considered the classical Donoho & Johnstone functions (Blocks, bumps, doppler, heavisine)
- The intensity function f_0 is set such that (with $\alpha \in \{1, \dots, 7\}$)

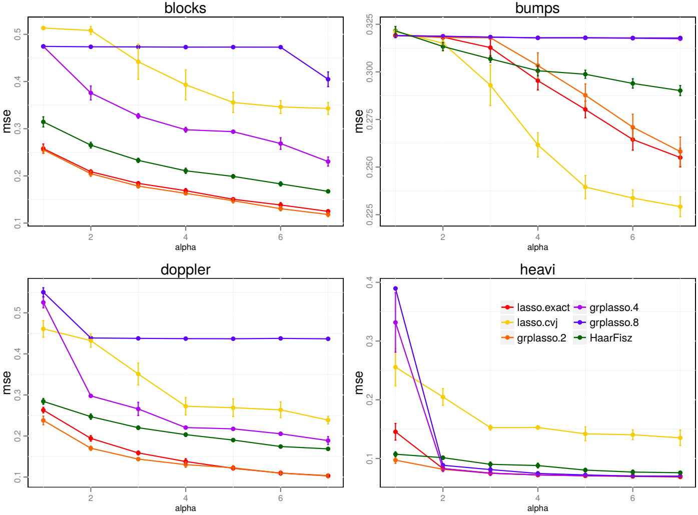
$$f_0 = \alpha \exp g_0$$

- Observations are sampled on a fixed regular grid ($n = 2^{10}$) with $Y_t \sim \mathcal{P}(f_0(X_t))$.
- Use Daubechie, Haar and Fourier as elements of the dictionary
- Check the normalized reconstruction error:

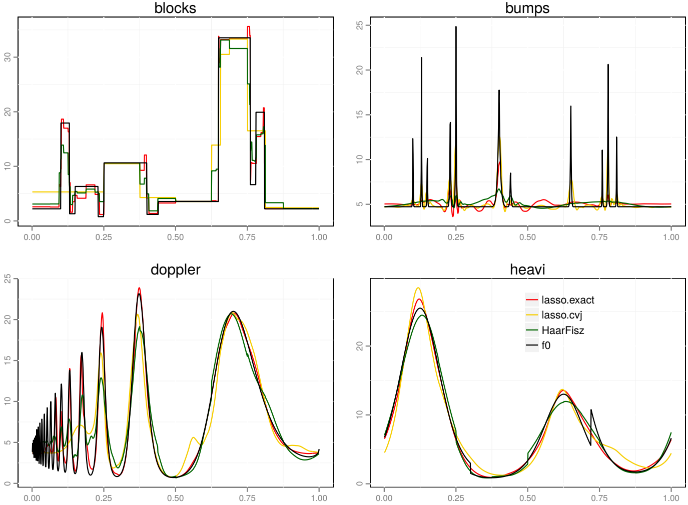
$$MSE = \frac{\|\widehat{f} - f_0\|_2^2}{\|f_0\|_2^2}$$

- Compete with the Haar-Fisz transform and cross-validation

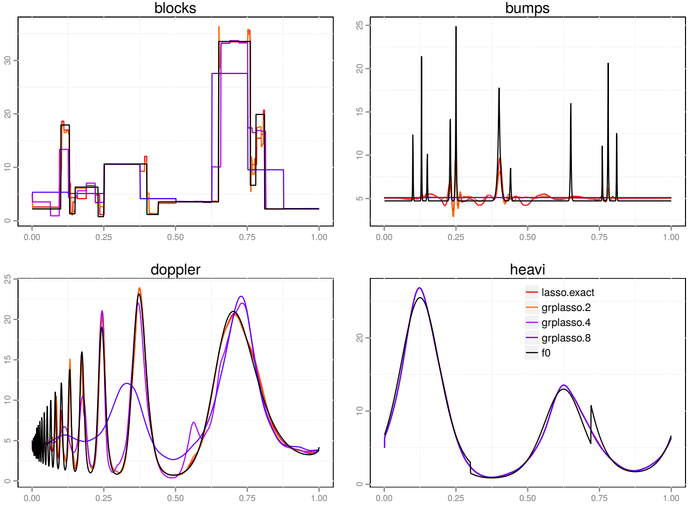
Reconstruction errors



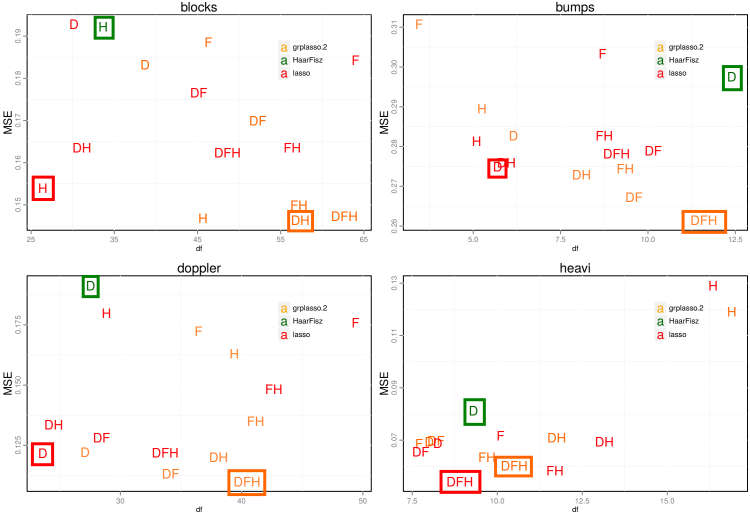
Estimated intensity functions (Lasso)



Estimated intensity functions (group-Lasso)

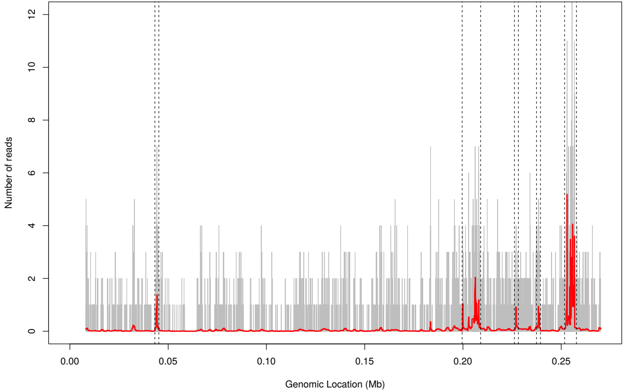


Choosing the best dictionary by cross-validation



Outline

Promising results on OriSeq



Perspective for the integration of multiple datasets

- This model-based framework offers many perspectives for the analysis of peak-like data
- Comparison of two conditions (A vs B) and/or normalization

$$
\begin{aligned}
\log f_A(x) &= \sum_{j \in \mathcal{J}} \beta_j \varphi_j(x) \\
\beta_j &= \alpha_j^{\mathsf{B}} + \gamma_j
\end{aligned}
$$

- Extend to varying coefficients models ($\beta_j(x)$).
- Consider the Negative Binomial distribution
- Preprint: `http://arxiv.org/abs/1412.6966`