ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

A dictionary learning based method for aCGH segmentation

Salvatore Masecchia[1] and Saverio Salzo[1] and Annalisa Barla[1] and Alessandro Verri[1]

1 - DIBRIS
via Dodecaneso 35 Genova - Italy

Abstract. The starting point of our work is to devise a model for segmentation of aCGH data. We propose an optimization method based on dictionary learning and regularization and we compare it with a state-of-the-art approach, presenting our experimental results on synthetic data.

1 Introduction

Copy number variations (CNVs) are alterations of the DNA that result in the cell having an abnormal number of copies of one or more sections of the DNA. Recurrent aberrations across samples may indicate an oncogene or a tumor suppressor gene, but the functional mechanisms that link altered copy numbers to pathogenesis are still to be explained. Array-based Comparative Genomic Hybridization (aCGH) is a modern whole-genome measuring technique that evaluates the occurrence of copy variants across the genome of samples (patients) versus references (controls) on the entire genome, extending the original CGH technology [1].

A signal measured with an aCGH is made of a piecewise linear (and constant) component plus some noise. The typical analysis on such data is *segmentation*, that is the automatic detection of altered copy numbers (amplifications or deletions). Differently from other molecular data, as gene expression, with aCGH it is possible to exploit the intrinsic data structure to improve the downstream analysis.

Many methods have been proposed for the extraction of CNVs based on different principles like filtering (or smoothing), segmentation, breakpoint-detection and calling [2, 3, 4], taking into account either one sample at a time or all samples together. Some interesting results in literature exploit the possibility to adopt regularization methods for a joint segmentation of many aCGH profiles with the simultaneous detection of shared change-points across samples. The works proposed by [4, 5, 6] follow this stream, and are based on total variation (TV) or fused lasso signal approximation. There, the TV is equally applied between each pair of consecutive points on the signal (*probes*), usually ordered by their physical location along chromosomes (*loci*). A simple observation is that there is no biological meaning to force such *continuity* between the last probe of a chromosome and the first probe on the next one. For this reason, in [4] the algorithm is run chromosome-by-chromosome. However, this solution does not allow to directly identify recurrent alterations, occurring on two different chromosomes (*e.g.,* due to an unbalanced translocation). Moreover, since the coefficients may assume either positive and negative values, it is difficult to

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

understand the role of the corresponding patterns in the representation of the signal. In Dictionary Learning such patterns are usually called *atoms*.

In this work, we present a novel model for aCGH segmentation based on the minimization of a functional combining several penalties. Our method is an extension of the model proposed by [4] addressing the following improvements: *(i)* the segmentation is performed on a signal possibly composed of multiple chromosomes, still preserving independency among chromosomes; *(ii)* the coefficients are constrained to be positive, hence simplifying the interpretability of the coefficients matrix in favor of selecting more representative atoms, especially when co-occurent alterations take place.

In the remainder of the paper we discuss such model, motivating the choice of each penalty. We also present a set of experiments on four types of synthetic data, comparing the results and highlighting the advantages over the model proposed in [4]. The employment of synthetic data offers a more controlled environment. This is the first attempt to validate the effectiveness of such model.

2 A new model for aCGH segmentation

In this work we propose an extension of the model proposed by [4] improving several important aspects that will increase the interpretability of the results. Both approaches are based on regularization combining different penalties simultaneously. We now present the problem more formally, recalling, first, the FLLat (Fused Lasso Latent feature) model [4] and then describing the proposed CGHDL (CGH analysis with Dictionary Learning) model.

We are given $S \in \mathbb{N}$ samples $(\mathbf{y}_s)_{1\leq s\leq S}$, with $\mathbf{y}_s \in \mathbb{R}^L$. Then, one seeks J *simple* atoms $(\boldsymbol{\beta}_j)_{1\leq j\leq J}$ which possibly give complete representation of all samples, in the sense that: $\mathbf{y}_s \approxeq \sum_{j=1}^{J} \theta_{js}\boldsymbol{\beta}_j \quad \forall s = 1,\dots,S$ for some vectors of coefficients $\boldsymbol{\theta_s} = (\theta_{js})_{i\leq j\leq J} \in \mathbb{R}^J$. To achieve this, in [4], the following model is proposed:

$$\min_{\boldsymbol{\theta}_s,\boldsymbol{\beta}_j} \sum_{s=1}^{S} \left\|\mathbf{y}_s - \sum_{i=1}^{J} \theta_{js}\boldsymbol{\beta}_j\right\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_1 + \mu \sum_{j=1}^{J} TV(\boldsymbol{\beta}_j) \text{ s.t. } \|\theta_{\cdot j}\|_2^2 \leq 1 \quad \forall j\,. \quad (1)$$

The ℓ^1 penalization term forces each atom $\boldsymbol{\beta}_j$ to be sparse and the total variation term $TV(\boldsymbol{\beta}_j) = \sum_{s=2}^{S} |\beta_{js} - \beta_{j,s-1}|$, induces small variations in the atoms. The hard constraints on the coefficients $\theta_{\cdot j}$ are imposed for consistency and identifiability of the model. Indeed, multiplying a particular feature $\boldsymbol{\beta}_j$ by a constant, and dividing the corresponding coefficients by the same constant leaves the fit unchanged, but reduces the penalty.

Our model is an extension of (1), driven by the following optimization problem depending on the three regularization parameters $\lambda, \mu, \tau > 0$:

$$\begin{aligned} \min_{\boldsymbol{\theta}_s,\boldsymbol{\beta}_j} & \sum_{s=1}^{S} \left\|\mathbf{y}_s - \sum_{i=1}^{J} \theta_{js}\boldsymbol{\beta}_j\right\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_1^2 + \mu \sum_{j=1}^{J} TV_{\mathbf{w}}(\boldsymbol{\beta}_j) + \tau \sum_{s=1}^{S} \|\boldsymbol{\theta}_s\|_1^2 \\ & \text{s.t. } 0 \leq \theta_{js} \leq \theta_{\max}, \quad \forall j = 1,\dots,J \quad \forall s = 1,\dots,L. \end{aligned} \quad (2)$$

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

This model improves (1), in several aspects.

First, we use a *weighted total variation*: $TV_{\mathbf{w}}(\boldsymbol{\beta}_j) = \sum_{l=1}^{L-1} w_l |\beta_{l+1,j} - \beta_{l,j}|$, where $\mathbf{w} = (w_l)_{1 \leq l \leq L-1} \in \mathbb{R}^{L-1}$ are properly chosen weights. $TV_{\mathbf{w}}$ is a generalized total variation due to the presence of the weights $\mathbf{w}$. This modification is introduced in order to relax at some points the constraint of *small jumps* on the atoms. Actually we will use weights that are always 1 with some sparse exceptions where w_l is close to zero. When dealing with aCGH, this allows to treat signals composed by several chromosomes as a whole, but still guaranteeing an independent analysis for each chromosome. This ensures the capability of identifying concomitant alterations occurring on different chromosomes.

Second, we constrain the coefficients to be positive. This avoids a cancellation effect in the representation of the signal leading to a simpler matrix of coefficients and a matrix of atoms which more clearly reveals the latent patterns in the data. In this way interpretability of the results is improved. For instance, when losses and gains occur within data at the same locus, the model selects different atoms to describe them as different phenomena. We further penalize the coefficients by the term $\tau \sum_{s=1}^{S} \|\boldsymbol{\theta}_s\|_1^2$, which induces sparsity along the set of weights associated to each sample separately. This permits to regulate how much different atoms each sample can combine in order to reconstruct the original signal.

Third, instead of $\lambda \sum_{i=1}^{J} \|\boldsymbol{\beta}_j\|_1$, which forces a general sparsity in (1), we use the term $\lambda \sum_{i=1}^{J} \|\boldsymbol{\beta}_j\|_1^2$, that gives a structured sparsity along the columns of the matrix of atoms $[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J]$.

To solve (2), we use a proximal alternating algorithm, as studied in its generality in [7]. We set $\mathbf{Y}$, $\mathbf{B}$ and $\boldsymbol{\Theta}$ as the matrices of data, atoms and coefficients respectively, and introduce the partial functions:

$$\begin{aligned} \varphi_{\mathbf{B}}(\boldsymbol{\Theta}) &= \frac{1}{2} \|\mathbf{Y} - \mathbf{B}\boldsymbol{\Theta}\|_F^2 + \delta_{\Delta_S^J}(\boldsymbol{\Theta}) + \tau \sum_{s=1}^{S} \|\boldsymbol{\Theta}(:,s)\|_1^2 \\ \psi_{\boldsymbol{\Theta}}(\mathbf{B}) &= \frac{1}{2} \|\mathbf{Y} - \mathbf{B}\boldsymbol{\Theta}\|_F^2 + \lambda \sum_{j=1}^{J} \|\mathbf{B}(:,j)\|_1^2 + \mu \sum_{j=1}^{J} TV_{\mathbf{w}}(\mathbf{B}(:,j)), \end{aligned} \tag{3}$$

where $\delta_{\Delta_{S\times J}}$ is the indicator function of the box set $\Delta_{S\times J} = [0, \theta_{\max}]^{S\times J}$. Then, the *alternating proximal algorithm*, with $\eta_k, \zeta_k > 0$, is as follows:

$$\left| \begin{aligned} \boldsymbol{\Theta}_{k+1} &= \mathrm{prox}_{\eta_k \varphi_{\mathbf{B}_k}}(\boldsymbol{\Theta}_k) := \mathrm{argmin}_{\boldsymbol{\Theta}} \big(\varphi_{\mathbf{B}_k}(\boldsymbol{\Theta}) + (2\eta_k)^{-1} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}_k\|_F^2 \big), \\ \mathbf{B}_{k+1} &= \mathrm{prox}_{\zeta_k \psi_{\boldsymbol{\Theta}_{k+1}}}(\mathbf{B}_k) := \mathrm{argmin}_{\mathbf{B}} \big(\psi_{\boldsymbol{\Theta}_{k+1}}(\mathbf{B}) + (2\zeta_k)^{-1} \|\mathbf{B} - \mathbf{B}_k\|_F^2 \big). \end{aligned} \right. \tag{4}$$

In (4) $\mathrm{prox}_{\zeta\psi_{\boldsymbol{\Theta}}}$, $\mathrm{prox}_{\eta\varphi_{\mathbf{B}}}$ denote the proximity operators with respect to the partial functions (3). They can be computed approximately, by a duality based (inner) algorithm, with a given and controlled precision [8].

3 Experiments

In this section we describe the generation of synthetic data and the experimental results obtained by employing CGHDL. The model of the signal follows [4] and

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

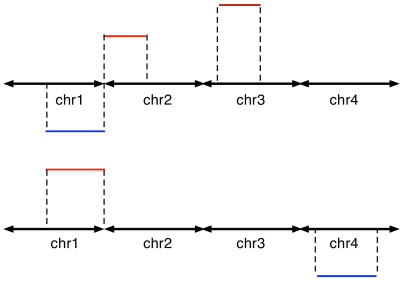

Fig. 1: Mean signals of the two patterns used for *Dataset 4* generation.

the additive noise model follows [2].

3.1 Synthetic data generation

The signal is defined as:

$$y_{ls} = \mu_{ls} + \epsilon_{ls}, \quad \mu_{ls} = \sum_{m=1}^{M_s} c_{ms} I_{\{l_{ms} \leq l \leq l_{ms}+k_{ms}\}}, \quad \epsilon_{ls} \sim N(0, \sigma^2), \qquad (5)$$

where $l = 1, \ldots, L$, $s = 1, \ldots, S$, μ_{ls} is the mean, and σ is the standard deviation of the noise ϵ_{ls}. The mean signal $\mu_{\cdot s}$ is a step function where M_s is the number of segments (regions of CNVs) generated for sample s and c_{ms}, l_{ms} and k_{ms} are the height, starting position and length respectively for each segment. We chose $M_s \in \{1, 2, 3, 4, 5\}$, $c_{ms} \in \{\pm 1, \pm 2, \pm 3, \pm 4, \pm 5\}$, $l_{ms} \in \{1, \ldots, L - 100\}$ and $k_{ms} \in \{5, 10, 20, 50, 100\}$, $L = 1000$, $S = 20$. According to this general schema, we generated four types of datasets:
Dataset 1: The samples are generated in order to minimize the probability of sharing segments, following the same schema as in [4, Sec. 4.1, Dataset 1].
Dataset 2: Following [4, Sec. 4.1, Dataset 2], the samples are designed to have common segments of CNVs. Each shared segment appears in the samples according to a fixed proportion randomly picked between $(0.25, 0.75)$. Starting points, lengths are shared among the selected samples, whereas the amplitudes c_{ms} still may vary within samples. The unshared segments are built as in *Dataset 1* for a maximum of 5 segments per sample.
Dataset 3: The atoms $\boldsymbol{\beta}_j$ are generated according the same schema of (5). The coefficients θ_{js} are randomly sampled in $[0, 1]$ and the signal is built as $\mathbf{Y} = \mathbf{B}\boldsymbol{\Theta}$.
Dataset 4: This dataset is explicitly designed to mimic a real signal composed of different chormosomes. We build three classes of samples. One third of the samples has mean signal as in the upper panel of Fig.1, one third has mean signal as in the lower panel of Fig.1 and the remaining third is built as *Dataset 1*.

3.2 Parameter selection

The choice of the parameters (J, λ, μ, τ) is done according to the Bayesian information criterion (BIC) [9]. The BIC mitigates the problem of overfitting by

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

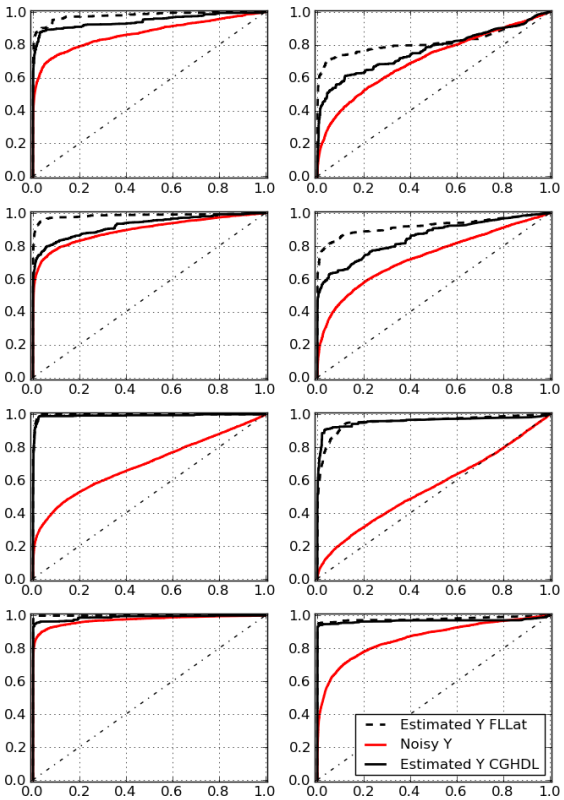

Fig. 2: ROC curves for varying levels of noise and different dataset type. The left column shows $\sigma = 1.0$, while the right column shows $\sigma = 2.0$. Red lines refer to the performances on the noisy $\mathbf{Y}$, dashed and solid lines refer to FLLat and CGHDL performances, respectively.

introducing a penalty term for the complexity of the model. In our case the BIC is written as: $(SL) \cdot \log\left(\frac{\|\mathbf{Y}-\mathbf{B\Theta}\|_F^2}{SL}\right) + k(\mathbf{B}) \log(SL)$ and $k(\mathbf{B})$ is computed as the number of jumps in $\mathbf{B}$ and ultimately depends on the parameters (J, λ, μ, τ). Note that, differently from [4], we also use BIC to select the number of atoms J.

3.3 Results

For the experiments we used Python scripts, implementing *ex novo* our approach and wrapping the available R code (FLLat) for the method (1). The number of atoms J varied in $\{5, 10, 15, 20\}$. Fig. 2 shows the performances of the two approaches. Following [4], ROC curves are built by evaluating the correct detection of alterations based on the denoised signal: the results are comparable. Performances on the raw noisy signal are also plotted for reference. Fig. 3 shows a plot of the solutions obtained by the two approaches on *Dataset 4*. The algorithm

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

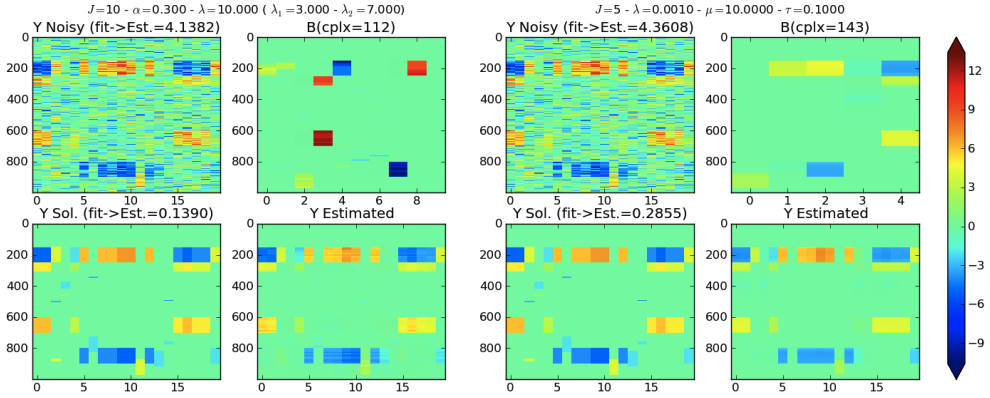

Fig. 3: *Dataset 4* analyzed by FLLat (left panel) and CGHDL (right panel). Each panel shows 4 subplots: top left plot represents the noisy data matrix, top right plot shows the atom matrix with atoms as columns, bottom left subplot is the *true* data matrix and bottom right is the estimated signal.

implementing (1) achieves good results in denoising, selecting $J = 10$ atoms, but fails in detecting the underlying patterns of Fig 1. The selected atoms represent single alterations. Conversely, our approach (right panel in Fig. 3) selects 5 atoms which clearly comprise the two patterns. Summarizing, the constraint on the positive coefficient seems very effective in selecting more informative atoms.

References

[1] A Kallioniemi, O-P. Kallioniemi, D Sudar, D Rutovitz, J W. Gray, F Waldman, and D Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, 258(5083):818–21, October 1992.

[2] 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):557–72, 2004.

[3] H Willenbrock and J Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–91, November 2005.

[4] G Nowak, T Hastie, J R. Pollack, and R Tibshirani. A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics*, June 2011.

[5] J-P Vert and K Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neural Information Processing Systems 23*, 1:1–9, 2010.

[6] HJ Wang and J Hu. Identification of Differential Aberrations in Multiple-Sample Array CGH Studies. *Biometrics*, 67(2):353–62, July 2010.

[7] H Attouch, J Bolte, P Redont, and A Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

[8] S Villa, S Salzo, L Baldassarre, and A Verri. Accelerated and inexact forward-backward algorithms. *Optimization Online*, 2012.

[9] G Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 1978.