Solving Generalized FLSA with ADMM Algorithm for Copy Number Variation Detection in Human Genomes

Jacob Biesinger
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
jake.biesinger@uci.edu

Yifei Chen
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
yifeic@uci.edu

1 Introduction

Structural variations (SVs) account for most of the bases that vary among human genomes [3] and are believed to contribute significantly to variation between individuals, possibly as large of an effect as Single Nucleotide Polymorphisms (SNPs) [9, 6]. Although some types of SV (such as copy number variation (CNV)) have cost-effective methods available for their discovery (SNP and CGH arrays [3]), SVs are still relatively under-ascertained [3]. Novel methods for detecting CNVs, including clone-based sequencing [7], paired-end mapping [8], and read-depth analysis [1], may offer an advantage over array-based methods, detecting different variant classes and sizes.

Recent work by Abyzov et al. [1] showed that these methodologies complement each other, detecting different types of variations and sizes. In particular, they developed a novel program, CNVnator, to detect and genotype CNVs by analysing read-depth (RD), or statistical fluctuations in read coverage across the genome. Their method was effective in detecting CNVs from read-depth data. Inspired by their work, we have here sought to use RD data to detect CNVs but whereas Abyzov et al. formulated the problem in terms of edge detection, we here formulate the problem as fitting a piecewise linear function directly to the RD data. To determine copy number across the genome, we solve the fused lasso problem, borrowing an efficient implementation from Ye et al [12]. The fused lasso has been applied to the CNV detection problem before by Tibshirani et al [10] and others, though only for array data. As far as we know, this is the first application of the fused lasso for detecting CNV from RD data.

1.1 Approach

The fused lasso is an optimization problem that trades off a data-fitting term with two reglarization terms: 1) an L_1 norm that encourages sparsity in the resulting coefficients and 2) an L_1 norm that discourages differences between adjacent coefficients. Intuitively, the first term corresponds to our prior belief that CNVs are rare (in all but a few places, the copy number for a diploid organism should be 2). The second term seeks to smooth the stochasticity in read placement, so that rapid variations in read depth are averaged out to be roughly piecewise constant. The optimization framework gives us explicit control over the strength of our prior beliefs about CNV abundance. Further, we can easily combine several RD datasets to determine variants common to a population. The framework should work efficiently when n, the number of samples, is much smaller than p, the number of dimensions. In the case of genomic data, n may only be one or a few samples while p may be in the millions. In contrast to array CGH data where it is relatively cheap to increase n, high throughput sequencing data is still expensive per sample (though prices are falling fast). We anticipate that as the cost of sequencing decreases, RD data will benefit from higher resolutions per sample and increased sample sizes, giving a further advantage over array-based methods.

In this paper we reformulate the biological problem as a convex optimization problem, by introducing the model of Generalized Fussed Lasso Signal Approximator (G-FLSA), which is shown below:

$$\min_{\beta} \Phi(\beta) = \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|\beta - c_i\|_1 + \lambda_2 \|L\beta\|_1 \tag{1}$$

where c_i represents the set of constant signals, and p_i controls the weights among them. The motivation under G-FLSA is to force the fitting curve to a set of constant values, not just to force sparsity. The model favors small changes between adjacent signals, due to the fused term $\|L\beta\|$.

We expect that the framework will be able to recapture the underlying copy number from raw sequence data without the need for manual partitioning other than equidistant binning. Our method relies on a correct alignment of short reads, though if the coverage is high enough, we should be able to safely make calls within repeat regions of the genome. Since our method is so reliant on a high-quality alignment, we do not expect to recover complicated rearrangements with signatures outside of an apparent change in copy number.

One potential problem is abundant free parameters exist, namely $\{c_i\}$,$\{p_i\}$,λ_1, and λ_2. In particular, to tune them by searching the entire parameter space is computationally intractable if the dimension of p is very large. In reality however, $\{c_i\}$ is often known from some theoretical clue, e.g., CNV problem in our topic. Thus for each sample we can decide the closest c_i, and $\{p_i\}$ can be set to the ratio of number of samples for each c_i. We will verify this heuristics in Section 4.

2 Methodology

2.1 ADMM for Generalized Fused Lasso

Model 1 is a positive weighted sum of three convex terms, and is thus still convex. However, both the L_1 and fused lasso terms are non-differentiable, and are coupled together by the variable β. We first introduce auxiliary variables to transform the problem into an equality constraint problem:

$$\begin{aligned} \min_{\beta,\{a_i\},b} &= \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|a_i\|_1 + \lambda_2 \|b\|_1 \\ s.t. a_1 &= \beta - c_1 \\ a_2 &= \beta - c_2 \\ &\vdots \\ a_K &= \beta - c_K \\ b &= L\beta \end{aligned} \tag{2}$$

The augmented Lagrangian of Problem 2 is given below:

$$\begin{aligned} L\left(\beta,\{a_i\},b,\{u_i\},v\right) = \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|a_i\|_1 + \lambda_2 \|b\|_1 + \sum_{i=1}^{K}\langle u_i, \beta - a_i - c_i\rangle \\ + \langle v, L\beta - b\rangle + \frac{\mu_1}{2}\sum_{i=1}^{K}\|\beta - a_i - c_i\|_2^2 + \frac{\mu_2}{2}\|L\beta - b\|_2^2 \end{aligned} \tag{3}$$

Problem 3 can be solved with Alternating Direction Method of Multipliers (ADMM) [2]. ADMM is a distributed optimization algorithm, which is very suitable for large scale problems. It has been successfully applied in solving Standard Fussed Lasso [12], and Double Regularized SVM [11].

Here for G-FLSA the algorithm details are shown below:

$$\begin{aligned}
\beta^{k+1} &= \arg\min_{\beta} L_{\mu_1,\mu_2}(\beta, \{a_i^k\}, b^k, \{u_i^k\}, v^k) \\
a_i^{k+1} &= \arg\min_{a} L_{\mu_1,\mu_2}(\beta^{k+1}, \{a_i\}, b^k, \{u_i^k\}, v^k), \forall i = 1, 2, ..., K \\
b^{k+1} &= \arg\min_{b} L_{\mu_1,\mu_2}(\beta^{k+1}, \{a_i^{k+1}\}, b, \{u_i^k\}, v^k) \\
u_i^{k+1} &= u^k + \mu_1(\beta^{k+1} - a_i^{k+1} - c_i), \forall i = 1, 2, ..., , K \\
v^{k+1} &= v^k + \mu_2(L\beta^{k+1} - b^{k+1})
\end{aligned} \tag{4}$$

Taking gradient or sub-gradient of primal variables, and setting them to or contains 0 leads to the following:

$$\begin{aligned}
(K\mu_1 + 1)I + \mu_2 L^T L)\beta^{k+1} &= y + \sum_{i=1}^{K}(\mu_1(a_i^k + c_i) - u_i^k) + L^T(\mu_2 b^k - v^k) \\
a_i^{k+1} &= \Gamma_{\frac{\lambda_1 p_i}{\mu_1}}\left(\beta^{k+1} - c_i + \frac{u_i^k}{\mu_1}\right), i = 1, 2, ...K \\
b^{k+1} &= \Gamma_{\frac{\lambda_2}{\mu_2}}\left(L\beta^{k+1} + \frac{v^k}{\mu_2}\right) \\
u_i^{k+1} &= u_i^k + \mu_1(\beta^{k+1} - a_i^{k+1} - c_i), i = 1, 2, ..., K \\
v^{k+1} &= v_k + \mu_2(L\beta^{k+1} - b^{k+1})
\end{aligned} \tag{5}$$

in which Γ is the soft threshold operator.
So to solve the Generalized FLSA problem with ADMM, we repeat Equation 5 until convergence. The first step is to solve a tridiagonal positive definite linear system. With Cholesky factorization, it can be easily solved with two triangular linear systems, and the complexity is bounded by $O(n)$, where n is the size of the signal. The second and third step is to perform soft threshold operation on auxiliary variable, while the fourth and fifth step is to update dual variables with gradient ascend. These steps can be easily implemented as well. We check the change rate of the original objective function, (Equation 1), as convergence criterion, i.e.,

$$\frac{|\Phi(\beta^{k+1}) - \Phi(\beta^k)|}{\Phi(\beta^k)} \leq \varepsilon \tag{6}$$

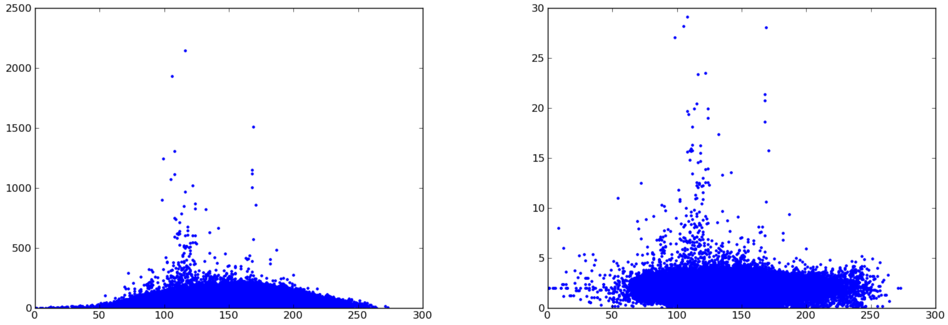
We implemented all the five sub-steps of Equation 5 in Matlab. To further improve performance, it would be easy to reimplement the algorithm in C, utilizing a linear algebra package like BLAS [4].

2.2 Read-depth data normalization and binning

Thanks to recent sequencing efforts by the 1000 genomes project [5], there is an abundance of low-coverage (10x-40x) samples available to the research community. To test our methodology, we examined chromosome 20 of a Yoruban female (individual NA18505, sample 20101123). We counted the number of reads in equidistant bins across the chromosome. We counted both ends of the read pair only if the complete pair was a proper mapping. We explored the effect of varying the bin size on our results. The bin size took on the values 50, 200, 300, 500, 1000, and 5000. Since read-depth has previously been shown to be correlated with GC content, we normalized the read count in each bin by the avarage read count for bins of a similar GC content. Specifically, we followed a method similar to Abyzov et al. [1], using the equation:

$$RD^i_{corrected} = \frac{\overline{RD}_{global}}{\overline{RD}_{gc_i}} \cdot RD^i_{raw} \cdot P - P \tag{7}$$

where RD is the read-depth, and $\overline{RD}$ indicates the average read depth (globally across the chromosome or in other bins with the similar GC content). P is the ploidy of the organism (set to 2 for our study). Scaling by the ploidy and then removing the ploidy effectively removes the bias from the lasso portion of the optimization framework. This effectively removed the correlation between RD and GC (see Figure 1).



(a) GC vs. RD, *raw*

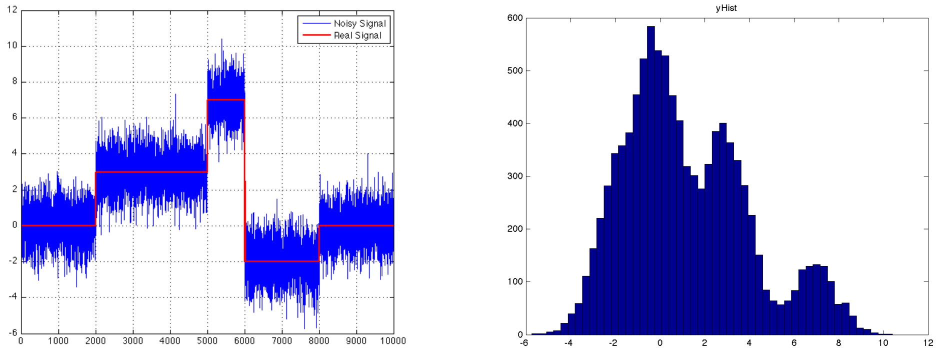
(b) GC vs. RD, *normalized*

Figure 1: Read-depth correlates with GC content.

3 Results

3.1 Simulation Data

To test our method, we generated RD data with a known underlying copy number. We used 10,000 bins and sampled data from a piece-wise constant function with values 0, 3, 7, -2, the ratio between which are 0.4:0.3:0.1:0.2. We add Gaussian noise ($\sigma = 1$) to it. The underlying pattern, noise signal, and histogram of noise signal are shown in Figure 2.



(a) Artificial Data

(b) Histogram of artificial data

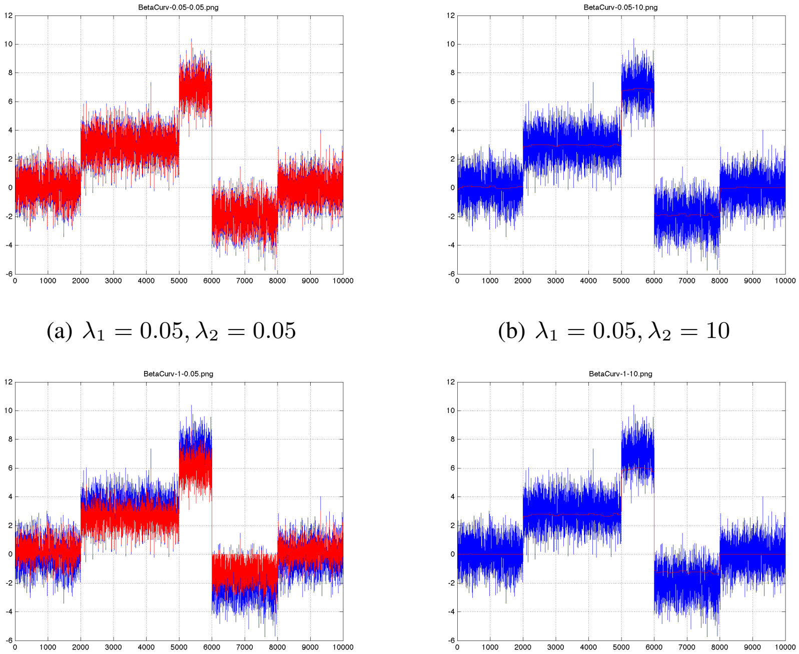
Figure 2: Simulation data and its histogram

For running G-FLSA, we set $\{c_i\} = \{0, 3, 7, -2\}$, and $\{p_i\} = \{0.4, 0.3, 0.1, 0.2\}$, according to the heuristic we mentioned in Section 1.1. The regularization parameter λ_1, λ_2 are chosen on a 2-dimensional grid of {0.05, 0.1, 0.5, 1, 5, 10}. Some sample results, at $\{\lambda_1 = 0.05, \lambda_2 = 0.05\}$, $\{\lambda_1 = 0.05, \lambda_2 = 10\}$, $\{\lambda_1 = 1, \lambda_2 = 0.05\}$, $\{\lambda_1 = 1, \lambda_2 = 10\}$ are shown in Figure. 3, 4.

From Figure 3(d), the tradeoff between a high λ_1 and a high λ_2 is clear. Here, the strength of the L_1 lasso term is evident, pulling the β close to 0 even when the data pulls strongly away from 0. Rather than fitting β to the true mean of the gaussian signal, a "middle road" is taken, about $\frac{\sigma}{2}$ from the mean towards 0.

3.2 CNV Data

Applying our method to a grid of varying bin size, λ_1 and λ_2, we determined the underlying copy number across chromosome 20. A histogram of the learned β is shown in figure 5. Unfortunately,



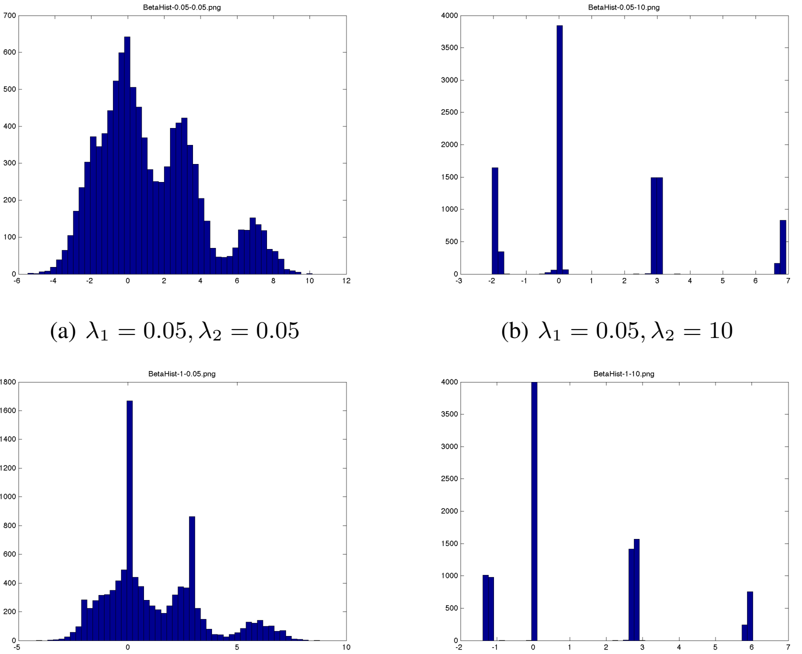
(a) $\lambda_1 = 0.05, \lambda_2 = 0.05$

(b) $\lambda_1 = 0.05, \lambda_2 = 10$

(c) $\lambda_1 = 1, \lambda_2 = 0.05$

(d) $\lambda_1 = 1, \lambda_2 = 10$

Figure 3: Curves of generalized FLSA at different regulation parameters



(a) $\lambda_1 = 0.05, \lambda_2 = 0.05$

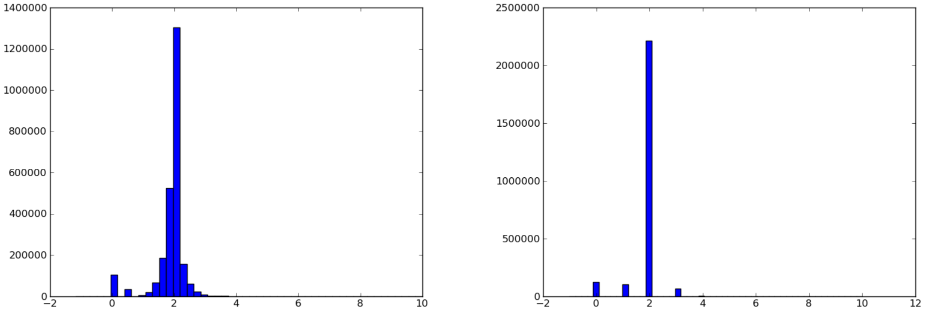
(b) $\lambda_1 = 0.05, \lambda_2 = 10$

(c) $\lambda_1 = 1, \lambda_2 = 0.05$

(d) $\lambda_1 = 1, \lambda_2 = 10$

Figure 4: Histograms of generalized FLSA at different regulation parameters

although we were able to generate predicted copy number variants for all of the bin widths and lambda parameters, because of issues with the genome build, we weren't able to compare our results with known copy number variations. There are several post-processing steps in other variant callers that we did not have time to implement including merging similar regions across the chromosome, performing significance tests and filtering variations that did not pass, filtering low confidence sites and sites with abnormally high mapped reads.



(a) Learned $beta$ for 500bp bins, λ_1=.5, λ_2=1

(b) β rounded to nearest integer

Figure 5: Results for chromosome 20

4 Discussion

The simulation case seems a very "idealized" situation: we know $\{c_i\}$ before hand, i.e., $\{$0, 3, 7, -2$\}$, and we know the real proportion of samples for each c_i, i.e., $\{$0.4, 0.3, 0.1,0.2$\}$, and set them to $\{p_i\}$. Figure.3 suggests that no matter how we set regularization parameter, the reconstructed signal obeys the proportion well. This means the proportion of data associated to each constant signal can be used to set relative regularization strength in between them, i.e., $\{p_i\}$. In this way, we can avoid "dimension disaster" of tuning $\{p_i\}$ when the number of constant signals are growing large.

On the other hand, however, we can see the reconstruction performance still largely depends on the set of regularization parameters λ_1, λ_2 from Figure.3, 4: In Figure.3(a), 4(a), when both regularizations are very small, the G-FLSA model essentially has little effect, as the reconstructed signal shows little difference to original (Figure.2). In Figure.3(d), 4(d), when both regularizations are very strong, G-FLSA forces very smooth piece-wise constant pattern. However, the constant values are biased and tends to "shift away" from the real value and all towards zero. Only in Figure.3(b), 4(b), when we set regularization to a moderate set do we reconstruct the signal well, in terms of both smoothness and little bias.

For simulation part in general, we find G-FLSA, together with its ADMM solver can be applied in modeling & estimating piece-wise constant signals. And sample ratio for each cluster is a good heuristic to avoid dimension problem of tuning parameter $\{p_i\}$. However, the performance still heavily depends on fussed Lasso regularization parameters λ_1, λ_2.

5 Conclusion

Overall, the approach we took to detect copy number variation seems promising. It allows us to explicitly define our prior belief in the abundance of these variations and how rapidly copy number changes across the genome. If we were to pursue this method, we would next want to perform validation and do a more principled parameter search to determine the optimal parameter set. Also, we would certainly need a post-processing step to filter low-confidence fluctuations from the final results. Currently, the only post-processing we do is to round the β values to the nearest integer. As Abyshov et al. found, the post-processing steps can filter many false positives from the final results.

References

[1] A. Abyzov, A.E. Urban, M. Snyder, and M. Gerstein. CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 2011.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. 2010.

[3] D.F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T.D. Andrews, C. Barnes, P. Campbell, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2009.

[4] J. Dongarra. Basic Linear Algebra Subprograms Technical Forum Standard. *Intl. Journal of High Performance Applications and Supercomputing.*, 16:1–111, 2002.

[5] R.M. Durbin, D.L. Altshuler, G.R. Abecasis, D.R. Bentley, A. Chakravarti, A.G. Clark, F.S. Collins, F.M. De La Vega, P. Donnelly, M. Egholm, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[6] L. Feuk, A.R. Carson, and S.W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[7] J.M. Kidd, G.M. Cooper, W.F. Donahue, H.S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

[8] J.O. Korbel, A. Abyzov, X.J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M.B. Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009.

[9] A.J. Sharp, Z. Cheng, and E.E. Eichler. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:407–442, 2006.

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

[11] G.B. Ye, Y. Chen, and X. Xie. Efficient variable selection in support vector machines via the alternating direction method of multipliers. *AI & Statistics 2011*, 2011.

[12] G.B. Ye and X. Xie. Split Bregman method for large scale fused Lasso. *Computational Statistics & Data Analysis*, 2010.