SPARSE SIGNAL RECOVERY METHODS FOR VARIANT DETECTION IN NEXT-GENERATION SEQUENCING DATA

Mario Banuelos, Rubi Almanza, Lasith Adhikari, Suzanne Sindi, and Roummel F. Marcia

Applied Mathematics
University of California, Merced
5200 North Lake Road, Merced, CA, USA.

ABSTRACT

Recent advances in high-throughput sequencing technologies have led to the collection of vast quantities of genomic data. Structural variants (SVs) – rearrangements of the genome larger than one letter such as inversions, insertions, deletions, and duplications – are an important source of genetic variation and have been implicated in some genetic diseases. However, inferring SVs from sequencing data has proven to be challenging because true SVs are rare and are prone to low-coverage noise. In this paper, we attempt to mitigate the deleterious effects of low-coverage sequences by following a maximum likelihood approach to SV prediction. Specifically, we model the noise using Poisson statistics and constrain the solution with a sparsity-promoting ℓ_1 penalty since SV instances should be rare. In addition, because offspring SVs inherit SVs from their parents, we incorporate familial relationships in the optimization problem formulation to increase the likelihood of detecting true SV occurrences. Numerical results are presented to validate our proposed approach.

***Index Terms*—** Sparse signal recovery, convex optimization, next-generation sequencing data, structural variants, computational genomics

1. INTRODUCTION

Recent advances in high-throughput sequencing technologies have led to the collection of vast quantities of genomic data. The 1000 Genomes Project [1], which catalogues human genomic variation in comprehensive detail, and the 3000 Rice Genomes Project [2, 3], which reports an international re-sequencing effort of 3,000 rice genomes, are two successful examples of such large-scale sequencing studies. These massive repositories of data offer the potential to increase our understanding of the complex evolutionary history of different species, identify genetic basis of important phenotypes including disease and – for humans – usher in the era of personalized medicine [4, 5]. A promising class of genetic variant emerging from such studies are structural variants (SVs) – rearrangements of the genome larger than one letter such as inversions, insertions, deletions, and duplications (see Fig. 1).

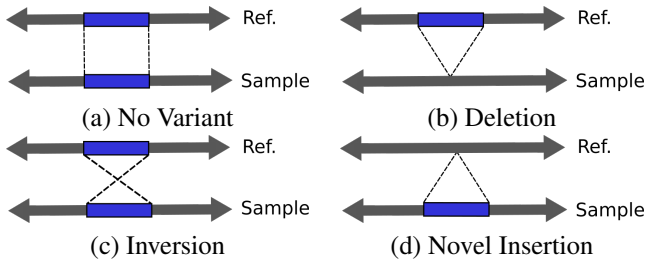

Fig. 1. Example of different structural variations in a sample genome in comparison to the reference genome.

SVs are typically predicted by sequencing fragments from an unknown individual genome and mapping those fragments to a previously identified reference genome [6, 7]. If the starting points of the genomic fragments are chosen uniformly and randomly from the genome, then the expected number of fragments covering any position in the genome is given by a Poisson distribution [8]. The mean of this Poisson distribution is referred to as the *coverage* of the genome. Since, in most large sequencing studies, many individuals will be sequenced at low coverage, even if an individual carries a genetic variant, we may not sample a fragment from that particular region of the genome. Similarly, if we observe a single fragment supporting a variant, it may represent an erroneous mapping rather than a true observation.

There have been many published methods to identify SVs from sequencing data (see, e.g., [9, 10, 11, 12, 13]). However, these approaches almost universally rely on high-coverage of a single individual genome and not on the scenario emerging from many large-scale sequencing efforts where there is low-coverage of many individuals. In addition, prior approaches when applied to populations typically consider each individual in isolation when – in fact – common variants would be shared by many individuals. Finally, most methods utilize a threshold – minimum number of supporting fragments – to prioritize predicted variants rather than a likelihood based statistic. Indeed, inferring SV information from sequencing

This work was supported by NSF Grant CMMI 1333326.

data has proven to be challenging because true SVs are rare and are prone to low-coverage noise. In this paper, we attempt to mitigate the deleterious effects of low-coverage sequences by following a maximum likelihood approach to SV prediction. Specifically, we model the noise using Poisson statistics and constrain the solution to promote sparsity, i.e., SV instances should be rare. Further, we consider multiple individuals and use relatedness among individuals as a constraint on the solution space – to our knowledge, this is the first SV detection algorithm to do so. Specifically, in our work below we use the assumption that a parent and child are sequenced and require that any SVs predicted in the child be present in the parent. Numerical analysis of both simulated and real sequencing data suggest that our approach has the promise to improve SV detection in studies of many low-coverage individuals.

2. SPARSE POISSON LOG-LIKELIHOOD OPTIMIZATION

Let $\vec{f}_i^* \in \{0,1\}^n$ be the vector of genetic variants for an individual i, i.e., $\vec{f}^*_{i,j} = 1$ if individual i has genetic variant j and is 0 otherwise. Let $\vec{y}_i \in \mathbb{Z}_+^n$ be the vector of observations for individual i. The variables $\vec{y}_{i,j}$ obey a Poisson distribution [14] whose mean, c_i, is equal to the sequencing coverage of individual i. In this paper, we specifically consider the structural variants for two individuals who are related, namely a parent and child. Let $\vec{f}_p^*$ and $\vec{f}_c^*$ be the true genomic variants for a parent and child, respectively. Then the corresponding observations, denoted by $\vec{y}_p$ and $\vec{y}_c$, are given by

$$\begin{aligned} \vec{y}_p &\sim \text{Poisson}(A_p \vec{f}_p^*) \\ \vec{y}_c &\sim \text{Poisson}(A_c \vec{f}_c^*), \end{aligned} \tag{1}$$

where $A_p = (c_p - \epsilon)\,\mathbb{I}, A_c = (c_c - \epsilon)\,\mathbb{I} \in \mathbb{R}^{n\times n}$ linearly transforms $\vec{f}_p^*,\ \vec{f}_c^*$ onto an n-dimensional set of observations $\vec{y}_p,\ \vec{y}_c \in \mathbb{Z}_+^n$. The constants c_p and c_c represent the sequencing coverage of the parent and child genome, respectively. It is assumed that ϵ, the error term in the measurement of the true signals, is the same for both observations.

We consider a general framework for the recovery of variant detection given sequencing data from one parent and one child. Our observation $\vec{y}$ will be considered a stacked signal in the form $\begin{bmatrix}\vec{y}_p^T & \vec{y}_c^T\end{bmatrix}^T$, where $\vec{y}_p$ and $\vec{y}_c$ represent observations of parent and child, respectively. Since the true signal $\vec{f}^*$ is also stacked, our observation model is given by

$$\vec{y} \sim \text{Poisson}(\hat{A}\vec{f}^*), \tag{2}$$

where $\hat{A} \in \mathbb{R}^{2n\times 2n}$ is a block-diagonal matrix with upper-left block A_p and lower-left block A_c.

2.1. Problem formulation

Under this Poisson model (2), the probability of observing $\vec{y}$ is given by

$$p(\vec{y}|\hat{A}\vec{f}^*) = \prod_{i=1}^{2n} \frac{(\vec{e}_i^T \hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T \hat{A}\vec{f}^*\right), \tag{3}$$

where $\vec{e}_i$ is the ith canonical basis vector. Under a similar framework in [15] and ignoring constant terms $\log(\vec{y}_i!)$, we minimize the negative Poisson log-likelihood given by

$$F(\vec{f}) = \mathbf{1}^T \hat{A}\vec{f} - \sum_{i=1}^{2n} \vec{y}_i \log\left(\vec{e}_i^T \hat{A}\vec{f} + \epsilon\right), \tag{4}$$

with gradient

$$\nabla F(\vec{f}) = \hat{A}^T\mathbf{1} - \sum_{i=1}^{2n} \frac{y_i}{\vec{e}_i^T \hat{A}\vec{f} + \epsilon} \hat{A}^T \vec{e}_i, \tag{5}$$

where $\mathbf{1}$ is a vector of ones. Hence, we focus on solving the following constrained optimization problem:

$$\begin{aligned} \underset{\vec{f}\in\mathbb{R}^{2n}}{\text{minimize}} \quad & \phi(\vec{f}) \equiv F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & 0 \le \vec{f}_c \le \vec{f}_p \le 1, \end{aligned} \tag{6}$$

where $\vec{f} = \begin{bmatrix}\vec{f}_p \\ \vec{f}_c\end{bmatrix}$, $\tau > 0$ is a regularization parameter, and pen is usually a non-differentiable penalty functional. Here, we impose the constraint $0 \le \vec{f}_c \le \vec{f}_p \le 1$ element-wise to enforce the continuous variables $\vec{f}_c$ and $\vec{f}_p$ to lie between 0 and 1 (i.e., SVs are either present or not), but in addition, to require that a variant in the child genome can be present only when the parent genome also has that variant.

2.2. Sparsity penalty

Our approach to solving (6) is based on SPIRAL [15, 16, 17], which is an iterative method whose iterates are defined from minimizing a sequence of quadratic subproblems. This approach utilizes the second-order Taylor expansion of the Poisson log-likelihood, $F(\vec{f})$, around the current iterate $\vec{f}^k$ and approximates the second derivative matrix by a scalar multiple of the identity matrix $\alpha_k I$, $\alpha_k > 0$ [15, 18, 19]. Thus, the next iterate is given by

$$\begin{aligned} \vec{f}^{k+1} = \begin{bmatrix}\vec{f}_p^{k+1} \\ \vec{f}_c^{k+1}\end{bmatrix} = \underset{\vec{f}\in\mathbb{R}^{2n}}{\arg\min} \quad & F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & 0 \le \vec{f}_c \le \vec{f}_p \le 1, \end{aligned} \tag{7}$$

where

$$F^k(\vec{f}) = \nabla F(\vec{f}^k)^T(\vec{f} - \vec{f}^k) + \frac{\alpha_k}{2}\|\vec{f} - \vec{f}^k\|_2^2.$$

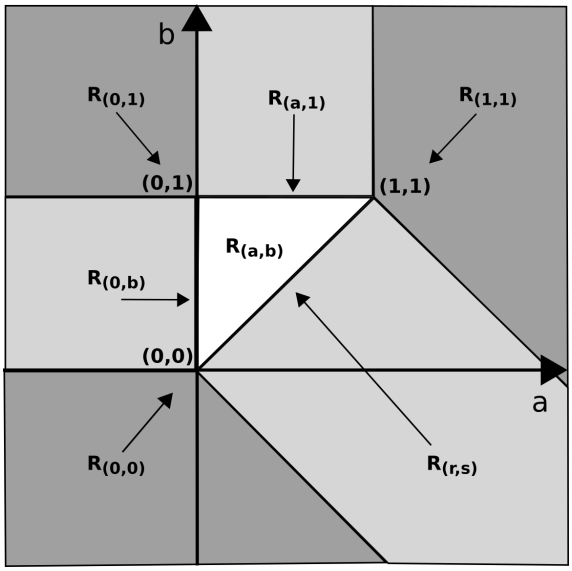

Fig. 2. Plot of a-b plane, where regions are defined in Table 1 and $R_{(a,b)}$ represents the feasible region for the solution of the separable subproblem (11).

Manipulating $F^k(\vec{f})$ leads to the following equivalent optimization formulation:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f}\in\mathbb{R}^{2n}}{\arg\min} \quad & \frac{1}{2}\|\vec{f}-\vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k}\mathrm{pen}(\vec{f}) \\ \text{subject to } \; & 0 \le \vec{f}_c \le \vec{f}_p \le 1, \end{aligned} \tag{8}$$

where

$$\vec{s}^k = \begin{bmatrix} \vec{s}_p^k \\ \vec{s}_c^k \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k). \tag{9}$$

When $\mathrm{pen}(\vec{f}) = \|\vec{f}\|_1 = \sum_{i=1}^n |\vec{f}_i|$, the objective function in (8) decouples in each variable and can be optimized separately, which results in the following *scalar* optimization:

$$\begin{aligned} \underset{f_p,\, f_c\in\mathbb{R}}{\text{minimize}} \; & \frac{1}{2}(f_p - s_p)^2 + \lambda|f_p| + \frac{1}{2}(f_c - s_c)^2 + \lambda|f_c| \\ \text{subject to } \; & 0 \le f_c \le f_p \le 1, \end{aligned} \tag{10}$$

where f_p and f_c correspond to each scalar element of $\vec{f}_p$ and $\vec{f}_c$, respectively. Since both f_p and f_c are non-negative, the absolute values in (10) can be dropped. Completing the squares in (10) and ignoring constant terms yield

$$\begin{aligned} \underset{f_p,\, f_c\in\mathbb{R}}{\text{minimize}} \; & \psi(f_p, f_c) = \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ \text{subject to } \; & 0 \le f_c \le f_p \le 1, \end{aligned} \tag{11}$$

where $a = s_c - \lambda,\ b = s_p - \lambda$. The solution to (11) can be obtained by partitioning the a-b plane into different regions (see Fig. 2). Then the minimizer of (11) depends on the region in which the point (a, b) lies. For example, if $(a, b) \in R_{(a,b)}$, i.e., $0 \le a \le b \le 1$, then the minimizer, (f_c^*, f_p^*), of (11) is the point (a, b). The complete set of minimizers is listed in Table 1.

Region	Condition a	Condition b	(f_c^*, f_p^*)
$R_{(a,b)}$	$0 < a < b$	$0 < b < 1$	(a, b)
$R_{(0,b)}$	$a < 0$	$0 \le b \le 1$	$(0, b)$
$R_{(a,1)}$	$0 \le a \le 1$	$b > 1$	$(a, 1)$
$R_{(0,1)}$	$a < 0$	$b > 1$	$(0, 1)$
$R_{(0,0)}$	$a \le -b$	$b < 0$	$(0, 0)$
$R_{(1,1)}$	$a > 1$	$b \ge -a + 2$	$(1, 1)$
$R_{(r,s)}$	$a > \lvert b\rvert$	$b < -a + 2$	(r, s)

Table 1. Table representing the solution to (11) as a function of a and b. Here, $r = s = (a + b)/2$.

3. RESULTS

The solution to the problem proposed in the previous section was implemented using the SPIRAL-ℓ_1 algorithm in [15] with the appropriate modifications to accommodate for the different constraints (see (6)). The results obtained are compared to those of the original SPIRAL-ℓ_1 approach in order to evaluate the validity of the proposed approach on both simulated and real genomic data.

3.1. Simulated Data

Two simulated test signals, $\vec{f}_p$ and $\vec{f}_c$, of length $n = 10^5$ were used to examine the effectiveness of the proposed approach. We varied the coverage of both between 2 and 10, and the child is chosen to have between 70% to 90% of the variants in the parent. The true signal for the parent $\vec{f}_p$, is set to be 0.5% sparse, so that only 500 variants are present. Furthermore, consistent with the assumption of similar error term in the measurement of the true signals, a single value of $\epsilon = 0.01$ was selected. On the simulated data, we are able to select the optimum value for τ and found on this data the optimal τ occurred between 0.5 and 3. Further, we observed limited sensitivity to τ as the model with and without familial constraints had a similar τ range.

We first examined the parent signal reconstruction. Fig. 3 illustrates a small segment ($n = 2.5 \times 10^4$) of the parent signal with $c_p = 2$, $c_c = 2$, and 90% similarity of variants, the reconstructed signal obtained by the regular SPIRAL constraints, and the reconstructed signal obtained by the familial SPIRAL constraints both at a threshold value of 0.5308. The improvement in variant predictions is visually clear from this figure.

We observed that an increase in the coverage of either child or parent helps improve the quality of the predictions. Moreover, the greater the similarity between parent and child,

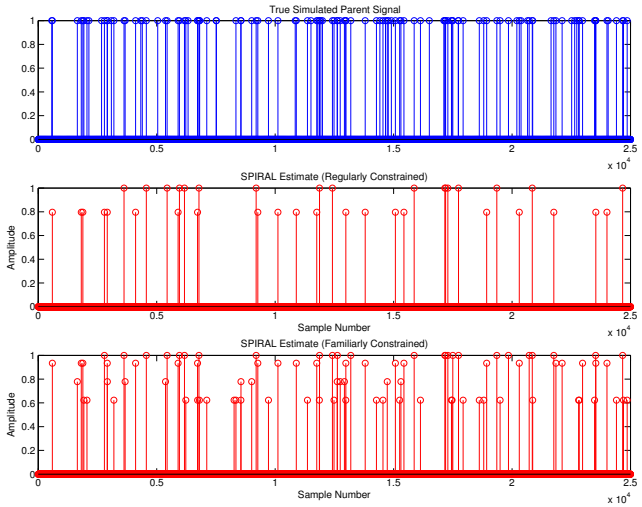

Fig. 3. From top to bottom: A small segment of the parent signal with $c_p = 2$, $c_c = 2$, and 90% similarity of variants; reconstruction using the regular SPIRAL constraints with $\tau = 1.779$ yielded 152 correctly identified out of 500; and reconstruction using the familial constraints with $\tau = 1.221$ yielded 349 correctly identified out of 500.

the more helpful adding the familial constraints results. Fig. 4 further illustrates how the familiarly constrained model ranks all true predictions above all false predictions.

3.2. 1000 Genomes Project Trio Data

We apply our method to the previously sequenced genomes of the father-mother-daughter CEU trio (NA12891, NA12892, NA12878) from the 1000 Genomes Project [1]. These genomes were sequenced to low coverage ($\approx 4\times$) in Pilot 1 of the study and high coverage ($\approx 40\times$) in Pilot 2. Both were aligned to NCBI36. We compared our reconstructions against the reported validated set of low coverage Chromosome 1 deletions longer than 250bp. In addition, we filtered the set of validated deletions by removing cases that overlapped the centromeres or telomeres and removed cases where a reported deletion was marked *LowQual* for all three individuals.

We used the GASV [9] method on this dataset as observations to predict the set of possible SVs. We filtered out SVs predicted to lie in the centromere or telomeres. We took the filtered set of predictions as the observed signals, and the true signals for each individual were constructed by determining if the validated deletions lie in the region predicted by GASV.

In the reconstruction of the parent signals, we separately use the child (NA12878) observed signal to constrain the parent signals. As shown in Figure 5, the reconstructions of both parent signals improve with the added familial constraints proposed by our method. Since NA12878 shares 90% and 92.5% of deletions with NA12891 (father) and NA12892

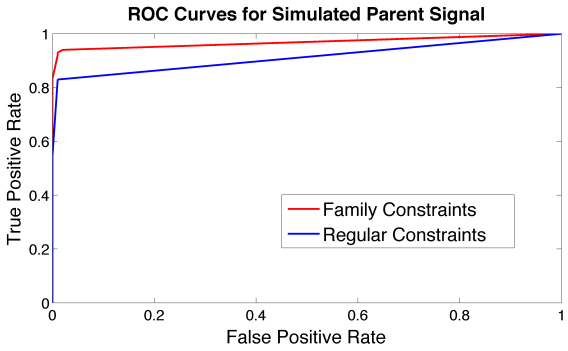

Fig. 4. ROC curves depicting the False Positive Rate vs True Positive Rate for the reconstruction of the parent signal with $c_p = 2$, $c_c = 5$, and 70% similarity of variants using both methods with $\tau = 1.553$ for regular constraints and $\tau = 1.474$ for family constraints.

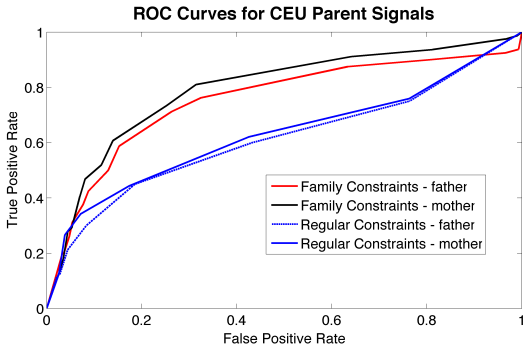

Fig. 5. ROC curves depicting the False Positive Rate vs True Positive Rate for the reconstruction of both CEU parent Chromosome 1 signals using both methods with $\tau = 2.65$.

(mother), respectively, we observe higher true positive rates for false positive rates > 0.1 in the reconstructions with added child data than the other method.

4. CONCLUSIONS

This paper presents a novel approach for inferring structural variants (SVs) from noise-corrupted data sets. We exploit the rare occurrence of SVs by incorporating a sparsity-promoting ℓ_1 penalty regularization term. Furthermore, we mitigate the deleterious effects of low-coverage sequences by following a maximum likelihood approach to SV prediction, and, in particular, using Poisson statistics to model the noise. Finally, we incorporate the relatedness of individuals as a constraint on the solution space. Specifically, we use the assumption that a parent and child are sequenced and require that any SVs predicted in the child be present in the parent. To our knowledge, our proposed approach is the first SV detection algorithm to do so. We demonstrated the effectiveness of our approach on both synthetic data and data from the 1000 Genomes Project.

5. REFERENCES

[1] D. M. Altshuler, E. S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, et al., "A map of human genome variation from population scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[2] X. Huang, T. Lu, and B. Han, "Resequencing rice genomes: an emerging new era of rice genomics," *Trends in Genetics*, vol. 29, no. 4, pp. 225–232, 2013.

[3] J.-Y. Li, J. Wang, and R. S. Zeigler, "The 3,000 rice genomes project: new opportunities and challenges for future rice research," *GigaScience*, vol. 3, no. 1, pp. 1–3, 2014.

[4] L. R. Pal and J. Moult, "Genetic basis of common human disease: Insight into the role of missense snps from genome-wide association studies," *Journal of molecular biology*, 2015.

[5] J. Weischenfeldt, F. Symmons, O.and Spitz, and J.O. Korbel, "Phenotypic impact of genomic structural variation: insights from and for human disease," *Nature Reviews Genetics*, vol. 14, no. 2, pp. 125–138, 2013.

[6] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature methods*, vol. 6, pp. S13–S20, 2009.

[7] S. S. Sindi and B. J. Raphael, "Identification of structural variation," *Genome Analysis: Current Procedures and Applications*, p. 1, 2014.

[8] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[9] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael, "A geometric approach for classification and comparison of structural variants," *Bioinformatics*, vol. 25, no. 12, pp. i222–i230, 2009.

[10] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome research*, vol. 19, no. 7, pp. 1270–1278, 2009.

[11] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, et al., "Breakdancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature methods*, vol. 6, no. 9, pp. 677–681, 2009.

[12] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "Delly: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[13] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. .C. Mell, and I. M. Hall, "Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome," *Genome research*, vol. 20, no. 5, pp. 623–635, 2010.

[14] D. Snyder, *Random Point Processes*, Wiley-Interscience, New York, NY, 1975.

[15] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—theory and practice," *IEEE Trans. on Image Processsing*, vol. 21, pp. 1084 – 1096, 2011.

[16] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "Sparse Poisson intensity reconstruction algorithms," in *Proceedings of IEEE Statistical Signal Processing Workshop*, Cardiff, Wales, UK, September 2009.

[17] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "Sparsity-regularized photon-limited imaging," in *Proceedings of IEEE International Symposium on Biomedical Imaging*, Rotterdam, The Netherlands, April 2010.

[18] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[19] S. J. Wright, R. D. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.