Sparse Maximin Aggregation of Neuronal Activity

Mogensen, Søren Wengel; Lund, Adam; Hansen, Niels Richard

Publication date:
2017

Document version
Peer reviewed version

Citation for published version (APA):
Mogensen, S. W., Lund, A., & Hansen, N. R. (2017). *Sparse Maximin Aggregation of Neuronal Activity*. Paper presented at Signal Processing with Adaptive Sparse Structured Representations, Lisbon, Portugal.

Sparse Maximin Aggregation of Neuronal Activity

Søren Wengel Mogensen
Department of Mathematical Sciences
University of Copenhagen
Email: swengel@math.ku.dk

Adam Lund
Department of Mathematical Sciences
University of Copenhagen

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen

***Abstract*—When analyzing large and inhomogeneous data sets it is of interest to obtain a robust estimate of an underlying signal. We consider a large data set describing neuronal activity in which systematic noise components are present. We propose the use of maximin aggregation and L_1 penalization to obtain a robust and sparse signal from this noisy data. An approximative computational method and an exact LARS-type method giving the entire solution path are presented.**

I. Introduction

Let X and B be random vectors taking values in $\mathbb{R}^p$ and ε be a zero-mean real random variable. Assume

$$Y = X^t B + \varepsilon.$$

We think of X as a vector of predictor values and of B as a vector of coefficients. Say now we have observations $Y_1, \ldots, Y_n$. If B has a degenerate distribution such that with probability one $B = \beta$ for a $\beta \in \mathbb{R}^p$, we have a standard linear regression setting. If the distribution of B is not degenerate we could still ask for a single $\beta \in \mathbb{R}^p$ to capture some feature of the data. For this purpose, define the *maximin effects* [1]

$$\arg\max_{\beta \in \mathbb{R}^p} \min_{b \in F} (2\beta^t \Sigma b - \beta^t \Sigma \beta)$$

where Σ is the population Gram matrix of X and F is the support of the distribution of B. The maximin effects maximize minimal (over F) explained variance when compared to the constant prediction. We will use the term *maximin aggregation* to refer to the process of aggregating effects across F to obtain the estimated maximin effects.

One can show that maximin aggregation enjoys a certain robustness property. Adding new vectors to F will only bring the maximin effects closer to the origin which corresponds to the constant prediction. This feature makes them attractive to use on noisy and inhomogeneous data sets as false positive results are unlikely.

We will from now on only consider the case of *known groups*, meaning that a known partition of the set of observations is available such that the regression coefficient is constant within these groups. We enumerate these groups by the natural numbers $1, \ldots, G$ such that

$$Y_i = X_i^t B_{g(i)} + \varepsilon_i$$

for a labelling function $g : \{1, \ldots, n\} \to \{1, \ldots, G\}$.

To obtain a sparse result we add an L_1 penalization on the parameter vector,

$$\arg\min_{\beta \in \mathbb{R}^p} \max_{g} (-\hat{V}_{\beta,b}) + \lambda \|\beta\|_1 \tag{1}$$

where $\hat{V}_{\beta,b}$ is an empirical version of the explained variance and λ is a non-negative tuning parameter.

II. Computations

We solved (1) for a fixed value of λ using a proximal gradient algorithm that iteratively applies a proximal operator to an initial point in the solution space. As fixed points for the proximal operator are solutions to the original optimization problem, the algorithm finds a solution as long as the proximal operator is firmly non-expansive. By making a softmax approximation to the maximin loss one obtains a locally Lipschitz loss function and hence a locally firmly non-expansive proximal operator. Using warm start, one can efficiently solve (1) for a finite sequence of λ's.

For our data example the design matrix is the same for all g in (1) and one can exploit this to obtain the complete solution path $\beta(\lambda)$ as this will be piecewise linear in $\mathbb{R}^p$ as a function of λ [2], see Figure 4. This is analogous to the LARS algorithm in the standard regression setting [3], [4]. Constructing the entire solution path is an alternative to e.g. the proximal algorithm, but it does not scale as well with the size of the data.

III. Data Example

We applied the methods to a data set obtained using *voltage-sensitive dye imaging* on the visual cortices of ferrets under a stimulus. The observations are spatio-temporal measurements of light intensity (two spatial dimensions and time) and stem from a total of 275 recordings of 13 ferrets. We treat the recordings as the known groups in (1).

Due to the delicate nature of the method a lot of the observations are highly irregular (see Figures 1 and 3) prompting the original authors to discard some of the data [5]. The measurements suffer from both a low signal-to-noise ratio and large, systematic noise components. A further complication of the analysis of this data is its sheer size. A personal computer is not capable of holding the full design matrix in memory and thus design-matrix free methods come in handy [6].

IV. Conclusion

Maximin aggregation combined with penalization offer an attractive way of obtaining a sparse signal from extremely noisy and inhomogeneous data. In our data example this set of methods allows the analyst to obtain meaningful results from the entire data set instead of hand-picking subsets of observations (Figure 2).

The estimation of penalized maximin effects is a computational challenge but is feasible for a sequence of penalty parameter values using e.g. a proximal gradient algorithm. It is also possible to even obtain the full solution path.

Acknowledgments

This work was supported by a research grant (13358) from VILLUM FONDEN.

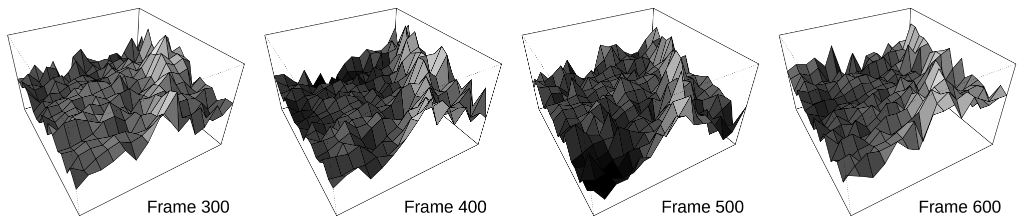

Fig. 1. Snapshots of a single recording (raw data).

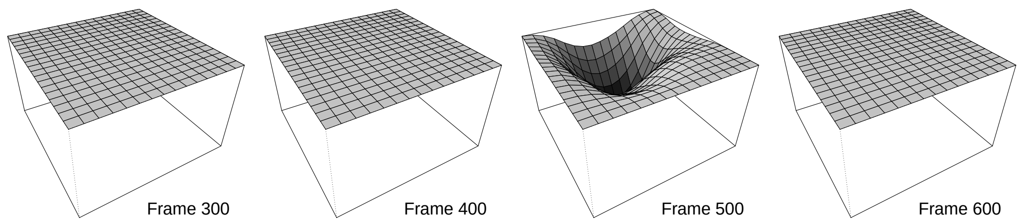

Fig. 2. Snapshots of the fitted maximin effects prediction.

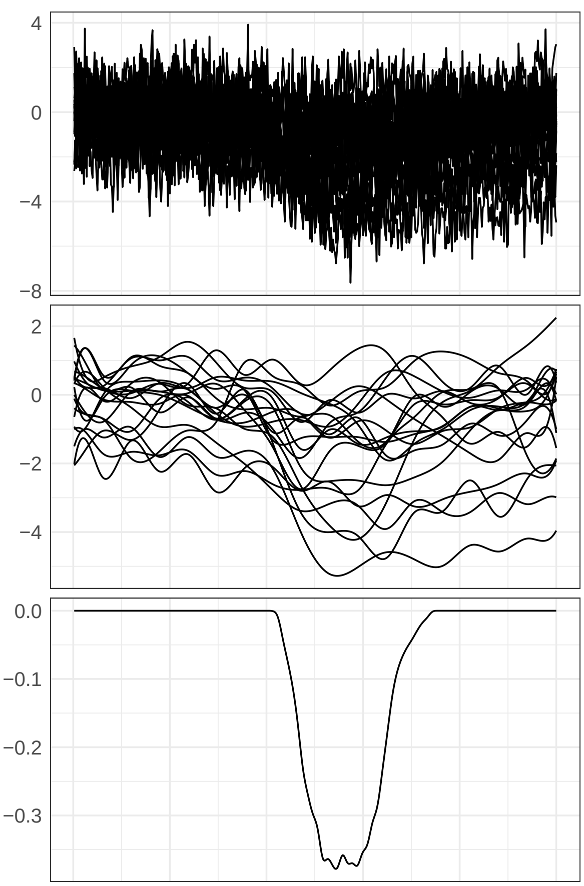

Fig. 3. Measurements from a single recording device during several recordings. Top: 20 randomly selected tracks as they evolve over time. Middle: smoothed version of the 20 tracks above. Bottom: prediction by the maximin effects estimated from the full data set. Note the different scaling of the y-axes.

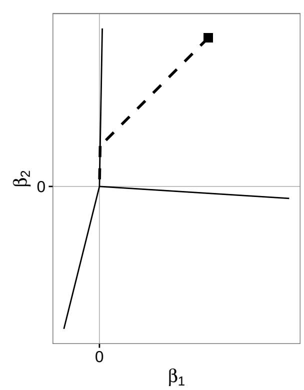

Fig. 4. Simple example of a solution path in $\mathbb{R}^2$. The three black line segments indicate the sets of points in which the loss is not differentiable. The square is the unpenalized maximin effects and the dashed line is the solution path. For large enough values of λ the solution is the zero vector.

References

[1] N. Meinshausen and P. Bühlmann, "Maximin effects in inhomogeneous large-scale data," *The Annals of Statistics*, vol. 43, no. 4, pp. 1801–1830, 2015.

[2] J. Roll, "Piecewise linear solution paths with application to direct weight optimization," *Automatica*, vol. 44, no. 11, pp. 2732–2737, 2008.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[4] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *The Annals of Statistics*, vol. 35, no. 3, pp. 1012–1030, 2007.

[5] P. E. Roland, A. Hanazawa, C. Undeman, D. Eriksson, T. Tompa, H. Nakamura, S. Valentiniene, and B. Ahmed, "Cortical feedback depolarization waves: A mechanism of top-down influence on early visual areas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 12 586–12 591, 2006.

[6] A. Lund, M. Vincent, and N. R. Hansen, "Penalized estimation in large-scale linear array models," *arXiv*, 2015. [Online]. Available: arxiv.org/pdf/1510.03298