Comment to "ℓ_1-Penalization for Mixture Regression Models" by Nicolas Städler, Peter Bühlmann, and Sara van de Geer

Gábor Lugosi
ICREA and Pompeu Fabra University *

May 20, 2010

I would like to congratulate the authors for this very interesting contribution. The generalization of ℓ_1-penalized linear regression to the "mixture-of-Gaussian-regressions" model raises some very interesting questions both from theoretical and algorithmic points of view and the paper offers a variety of powerful tools to attack both problems. In this comment I would like to mention another direction in which algorithmic issues become a relevant and non-trivial challenge.

The basic underlying assumption behind LASSO and various related methods of linear regression is sparsity. In the simplest fixed design regression model, one observes a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ generated by

$$Y_i = \beta_i + \sigma X_i$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector of independent standard normal random variables, $\sigma > 0$ is a parameter, and $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_n) \in \mathbb{R}^n$ is the mean vector to be estimated. The sparsity assumption is that $\boldsymbol{\beta}$ has a small number of non-zero components. Then the LASSO estimate is

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{v} \in \mathbb{R}^n} \left(\|\boldsymbol{Y} - \boldsymbol{v}\|_2^2 + \gamma \|\boldsymbol{v}\|_1 \right)$$

where $\gamma > 0$ is a regularization parameter. The nice thing about this estimate is that it is computationally feasible even when n is quite large and various available *sparsity oracle inequalities* guarantee that the method performs remarkably well under the sparsity assumption. (The references of the discussed paper contain many relevant pointers to the literature.)

In some situations, however, one may know more than just sparsity of the vector to be estimated. For example, it may be natural to assume that the set of indices of the non-zero

*The author acknowledges support by the Spanish Ministry of Science and Technology grant MTM2009-09063 and by the PASCAL Network of Excellence under EC grant no. 506778.

coefficients belongs to a given family of subsets of $\{1, \ldots, n\}$. As an example, one may think about a noisy measurement of an image where components of the observed vector represent pixels. Then it is not unnatural to assume that the underlying mean vector $\boldsymbol{\beta}$ (i.e., the noiseless image) is non-zero on a connected set of pixels, or perhaps $\boldsymbol{\beta}$ is a linear combination of a few such vectors. In fact, such situations abound and we list a few examples below.

Formally, let $\mathcal{C} = \{S_1, \ldots, S_N\}$ be a class of N subsets of $\{1, \ldots, n\}$ and for each $S_i \in \mathcal{C}$, denote by

$$\boldsymbol{s}_i = \left(\mathbb{1}_{\{1 \in S_i\}}, \ldots, \mathbb{1}_{\{n \in S_i\}}\right)$$

the incidence vector of S_i. An assumption that we may call *combinatorial sparsity* is that β can be expressed as

$$\boldsymbol{\beta} \approx \sum_{i=1}^{N} c_i \boldsymbol{s}_i$$

where $\boldsymbol{c} = (c_1, \ldots, c_N) \subset \mathbb{R}^N$ is a vector with a small number of non-zero coefficients. The approximate inequality can be interpreted in the ℓ_2 sense, for example. Similar ideas have lead to the *Group* LASSO (see Yuan and Lin [7]) in which the sets $S_1, \ldots, S_N$ are disjoint "blocks" of indices. Jacob, Obozinski, and Vert [5] consider a more general scenario in which the sets in $\mathcal{C}$ may overlap though $\mathcal{C}$ is still considered to be a relatively small set. However, in many interesting cases (such as some of the examples described below), N is very large, possibly even exponential in n, which poses important additional challenges. We also mention the closely related work of Huang, Zhang, and Metaxas [4] who introduce a general framework for such "structured" sparsity.

The corresponding hypothesis testing problem has received quite a bit of attention recently. In this problem, upon observing Y, one wants to test whether the mean of $\boldsymbol{Y}$ is the all-zero vector or $\mu \boldsymbol{s}_i$ for some $i = 1, \ldots, N$ where $\mu > 0$ is a known parameter. Arias Castro, Candès, Helgason, and Zeitouni [2], Arias Castro, Candès, Durand [3], Addario-Berry, Broutin, Devroye, and Lugosi [1] derive various upper and lower bounds in many interesting cases.

Next we mention some examples considered in these papers.

- **Paths.** Suppose we are given a graph of n vertices with two special nodes u and v. Each component of the vector $\boldsymbol{Y}$ corresponds to a noisy measurement on a vertex. Often (for example, when measuring traffic in certain networks) it makes sense to assume that the mean vector is close to a sparse linear combination of indicators of paths between u and v. Then $\mathcal{C}$ contains all loop-free paths between u and v. This is one of the main examples originally considered in [2]. Typically, each path contains a small number $k \ll n$ of vertices but the number of paths is exponentially large compared to n.

- **Clusters.** Suppose again that one takes a noisy measurement on each vertex of a network (such as a square grid, for example) with n vertices. Then it is often natural to assume that the mean vector $\boldsymbol{\beta}$ is close to a linear combination of a small number

of incidence vectors of connected regions. For example, if the underlying graph is a square grid, one may take $\mathcal{C}$ as the class of all rectangles or the class of all convex polygons, etc. The corresponding testing problem is investigated in depth in [3] where numerous practical applications of problems of this type are also described.

- **Spanning trees.** In one of the examples considered in [1], one is given a complete graph with m vertices and $n = \binom{m}{2}$ edges. The components of $\boldsymbol{Y}$ correspond to a measurement corresponding to every edge. Here $\mathcal{C}$ may be taken as the class of all $N = m^{m-2}$ spanning trees.

- **Perfect matchings.** In another example considered in [1], given a complete bipartite graph $K_{m,m}$ with $n = m^2$ edges, $\mathcal{C}$ represents the set of all $N = m!$ perfect matchings.

The common feature in all these examples is that $\mathcal{C}$ is a very large class with a certain combinatorial or geometric structure and every set $S_i \subset \mathcal{C}$ has a very small cardinality compared to n.

A natural way to approach such problems is to use the incidence vectors $\boldsymbol{s}_1, \dots, \boldsymbol{s}_N$ as a *dictionary* and define the corresponding LASSO estimate by

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{c} \in \mathbb{R}^N} \left(\left\| \boldsymbol{Y} - \sum_{i=1}^N c_i \boldsymbol{s}_i \right\|_2^2 + \gamma \sum_{i=1}^N |c_i| \right)$$

for some $\gamma > 0$. This raises two non-trivial issues. One is bounding the performance of this estimate, while the other, algorithmic problem is whether one can compute $\widehat{\boldsymbol{\beta}}$ efficiently. This is a non-trivial question because N is so big that even algorithms that run in time linear in N are infeasible.

To bound the performance of this estimate, we may use a recent elegant result of Massart and Meynet [6]. Their result is especially useful in our setting because it does not require any condition on the dictionary $\boldsymbol{s}_1, \dots, \boldsymbol{s}_N$. Specialized to our setting, the "ℓ_1 oracle inequality" of Massart and Meynet implies the following. Suppose every set $S_i \in \mathcal{C}$ has the same cardinality $|V_1| = k$ and assume $\gamma \geq 4\sigma\sqrt{1/(kn)}(1 + \sqrt{\log N})$. Then there exists a universal constant $C > 1$ such that

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq C \left(\inf_{\boldsymbol{c} \in \mathbb{R}^N} \left(\left\| \sum_{i=1}^N c_i \boldsymbol{s}_i - \boldsymbol{\beta} \right\|_2^2 + \gamma \sum_{i=1}^N |c_i| \right) + \gamma\sigma\sqrt{\frac{k}{n}} \right)$$

If σ is known, γ may be chosen to be equal to the smallest value for which the bound holds and we get

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq C \left(\inf_{\boldsymbol{c} \in \mathbb{R}^N} \left(\left\| \sum_{i=1}^N c_i \boldsymbol{s}_i - \boldsymbol{\beta} \right\|_2^2 + \sqrt{\frac{\log N}{kn}} \sum_{i=1}^N |c_i| \right) + \frac{\sigma^2}{n}\sqrt{\log N} \right)$$

(for a possibly different value of C). In fact, the term $\sqrt{\log N}$ can be replaced by the expected maximum

$$\mathbb{E} \max_{i=1,\dots,N} \frac{1}{\sqrt{k}} \sum_{j \in S_i} X_i$$

of a Gaussian process indexed by the class $\mathcal{C}$. (Note that the expectation is always bounded by $\sqrt{2 \log N}$.)

In order to make the estimate useful, one must find efficient ways of computing $\widehat{\boldsymbol{\beta}}$. This needs to be done separately for each case and it is a non-trivial challenge to find such efficient algorithms. We believe that in many important cases (such as most examples mentioned above) it should be possible to establish such algorithms. Here we describe one simple case.

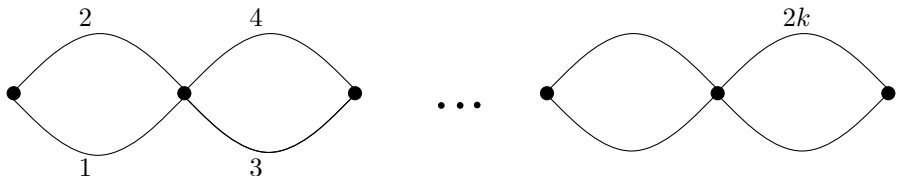

Figure 1: $N = 2^k$ paths of length k in a graph with n edges.

Suppose that the $n = 2k$ components of the vector $\boldsymbol{Y}$ correspond to edges in a (multi-) graph shown on Figure 1. Let $\mathcal{C}$ contain all $N = 2^k$ paths from the leftmost vertex to the rightmost vertex. To compute the LASSO estimate, one may equivalently solve the dual problem of

$$\min_{\boldsymbol{c} \in \mathbb{R}^N} \left\| \boldsymbol{Y} - \sum_{i=1}^{N} c_i \boldsymbol{s}_i \right\|_2^2 \quad \text{subject to} \quad \sum_{i=1}^{N} |c_i| \le B$$

where B is an appropriate constant. Label the edges as indicated on the figure. The key observation is that for every $j = 1, \dots, k$,

$$\sum_{i : 2j \in S_i} |c_i| + \sum_{i : 2j-1 \in S_i} |c_i| = \sum_{i=1}^{N} |c_i|$$

and therefore, denoting the two terms on the left-hand side by a_{2j-1} and a_{2j}, the problem decomposes to solving k 2-dimensional problems of the form

$$\min_{(a_{2j-1}, a_{2j}) \in \mathbb{R}^2} (Y_{2j-1} - a_{2j-1})^2 + (Y_{2j} - a_{2j})^2 \quad \text{subject to} \quad a_{2j-1} + a_{2j} \le B \ ,$$

which, of course, can be solved easily.

Of course, this example is trivial algorithmically while others may require more sophisticated ideas and methods. We believe that a systematical study of such problems with "combinatorial sparsity" is an interesting challenge and may lead to useful estimates in a number of applications.

Acknowledgements

I would like to thank Ery Arias-Castro and Pascal Massart for interesting discussions.

References

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. Combinatorial testing problems. *Annals of Statistics*, 2010, to appear.

[2] E. Arias-Castro, E.J. Candès, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36:1726–1757, 2008.

[3] E. Arias-Castro, E. Candès, and A. Durand. Detection of an anomalous cluster in a network. *arXiv:1001.3209*, 2010.

[4] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Tech Report arXiv:0903.3002*, 2009.

[5] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

[6] P. Massart and C. Meynet. An ℓ_1 oracle inequality for the Lasso. *preprint*, 2010.

[7] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B.*, 68:49–67, 2006.