Electronic Journal of Statistics
Vol. 9 (2015) 1723–1726
ISSN: 1935-7524
DOI: 10.1214/15-EJS998

Comment on "Hypothesis testing by convex optimization"*

Philippe Rigollet†

Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307, USA
e-mail: rigollet@math.mit.edu

Received February 2015.

With the growing size of problems at hand, convexity has become preponderant in modern statistics. Indeed, convex relaxations of NP-hard problems have been successfully employed in a variety of statistical problems such as classification [2, 16], linear regression [7, 5], matrix estimation [8, 12], graphical models [15, 9] or sparse principal component analysis (PCA) [10, 4]. The paper "Hypothesis testing by convex optimization" by Alexander Goldenshluger, Anatoli Juditsky and Arkadi Nemirovski, hereafter denoted by GJN, brings a new perspective on the role of convexity in a fundamental statistical problem: composite hypothesis testing. The role of this problem is illustrated in the light of several interesting applications in Section 4 of GJN.

One of the key insights in GJN is that there exists a pair of distributions, one in each of the composite hypotheses and on which the statistician should focus her efforts. Indeed, Theorem 2.1(ii) guarantees that any test that is optimal for this simple hypothesis problem is also near optimal for the composite hypothesis problem. Moreover, this pair can be found by solving a convex optimization problem. While convexity does not necessarily imply tractability, the convex problem considered here may become simple to the point that closed solutions exist even though no succinct description of the hypothesis sets may be known. This point is illustrated below.

Unlike the papers cited above, where the original problem to be solved is non-convex, GJN assumes given convex hypotheses (or finite unions of convex hypotheses). Hereafter, we investigate the performance of the proposed test when convexity is artificial and arises as a relaxation of a non-convex problem.

Let us consider two examples that fall under the umbrella of *combinatorial testing problems* [1]. Such problems are defined as follows. Assume that one observes a Gaussian random vector $X \sim \mathcal{N}(\mu, I_n)$ for some $\mu \in \mathbb{R}^n$. Let $\mathsf{p} \in \{0,1\}^n$ be a *sparsity pattern* [17]. Given a class $\mathcal{P} \subset \{0,1\}^n$ of sparsity patterns,

*Main article 10.1214/15-EJS1054.
†Supported in part by NSF grants DMS-1317308, CAREER-DMS-1053987.

we are interested in the following hypothesis testing problem:

$$H_0 : \mu = 0 \in \mathbb{R}^d \quad \text{vs} \quad H_1 : \mu \in \lambda\mathcal{P},$$

where $\lambda\mathcal{P} = \{\lambda \mathsf{p} : \mathsf{p} \in \mathcal{P}\}$ for some $\lambda > 0$. The question is: "How large should λ be in order to test with a pre-specified small risk?". Here, the risk of a test is defined as in GJN.

Several classes of sparsity can be considered [1] but perhaps two of them have more direct statistical relevance. The first one is the class of k-sparse vectors defined as $\mathcal{P}_1^n = \{\mathsf{p} \in \{0,1\}^n, \sum_j p_j = k\}$. The problem becomes detection of sparse means, which has applications in various problems including signal processing and steganography. To describe the second problem, assume that $n = d^2$ and fix an arbitrary bijection $T : \mathbb{R}^n \to \mathbb{R}^{d\times d}$, onto the space of $d \times d$ real matrices. The class $\mathcal{P}_2$ of k-clusters (or cliques) is defined as the set $\mathcal{P}_2^n = \{\mathsf{p} \in \{0,1\}^n : T(\mathsf{p}) = \mathsf{q}\mathsf{q}^\top, \mathsf{q} \in \mathcal{P}_1^d\}$. In other words, these are the sparsity patterns p such that $T(\mathsf{p})$ is the adjacency matrix of a clique of size k in an otherwise empty graph of size d. The class $\mathcal{P}_2$ of sparsity patterns has applications in clustering [6, 14] and sparse PCA [3, 4].

These combinatorial testing problems do not fall in the category of *good observation schemes* as defined in GJN because the class $Y = \lambda\mathcal{P}$ is not convex. Moreover, these two sets are of size that is exponential in k and performing the simple hypothesis tests

$$H_0 : \mu = 0 \in \mathbb{R}^d \quad \text{vs} \quad H_1 : \mu = \lambda\mathsf{p},$$

for all $\mathsf{p} \in \mathcal{P}_i^n, \in \{1,2\}$ as recommended in Section 3.1 of GJN is simply intractable. To circumvent this limitation, let us explore a *convexification* of the problem and study instead

$$H_0 : \mu = 0 \in \mathbb{R}^d \quad \text{vs} \quad H_1 : \mu \in \lambda\,\mathsf{conv}(\mathcal{P}),$$

where $\mathsf{conv}(\mathcal{P})$ denotes the convex hull of $\mathcal{P}$ and is defined as the smallest convex set that contains $\mathcal{P}$. In the case of $\mathcal{P}_1^n$ and $\mathcal{P}_2^n$, these convex sets are polytopes. Even so, optimization over polytopes may not be tractable. For example, some polytopes are known to not have a description involving a small number of linear constraints [18] and are therefore not amenable to linear programming. Fortunately, the optimization problems that are required by GJN admit an explicit solution in these two specific cases. Indeed, it follows from equation (7) in GJN that a near optimal test can be found by testing $H_0 : \mu = 0$ against $H_1 : \mu = \lambda\bar{\mu}$ where $\bar{\mu}$ is the point in the polytope $\mathsf{conv}(\mathcal{P})$ with the smallest Euclidean norm. For both polytopes $\mathsf{conv}(\mathcal{P}_1^n)$ and $\mathsf{conv}(\mathcal{P}_1^n)$, such a vector can be easily computed analytically.

We begin with $\mathcal{P}_1^n$. In this case, it is simply the vector $\bar{\mu}_1 = (k/n)\mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the all-ones vector. Moreover, the optimal test of 0 versus $\bar{\mu}_1$ has small risk as soon as $\lambda \geq C\sqrt{n}/k$ for some positive constant C. This rate is known to be optimal when $k \gg \sqrt{n}$ but is suboptimal for smaller values of k [1].

In the case of $\mathcal{P}_2^n$, it can be shown that

$$T(\bar{\mu}_2) = \frac{k(k-1)}{d(d-1)} \mathbf{1}_d \mathbf{1}_d^\top + \frac{k(d-k)}{d(d-1)} I_d,$$

so that the optimal test of 0 versus $\bar{\mu}_2$ has small risk as soon as $\lambda \geq Cd/k^2$ for some positive constant C. As before this rate is optimal if $k \gg \sqrt{d}$ but a better rate can be achieved if $k \ll \sqrt{d}$ [6, 14]. As a result, it seems that convexifying the problems in that way is too coarse for very sparse cases.

While in appearance the two problems seem to have the same computational limitations, they are in reality quite different from this point of view. Indeed, on the one hand, detecting a sparse mean $\mu \in \mathcal{P}_1^n$ can be solved in a very efficient way by simply looking at the k largest entries of X [1]. On the other hand, a recent line of work has shown that optimal detection for k-clusters may not be solvable efficiently if one wishes to use a computationally efficient procedure such as the one employed in GJN. Indeed, sparse PCA [3] and sub-matrix detection [14] both have the k-cluster structure and are known to be intrinsically computationally hard to solve optimally if one believes in the *planted clique conjecture* [11, 13].

References

[1] Addario-Berry, L., Broutin, N., Devroye, L. and Lugosi, G. (2010). On combinatorial testing problems. *Ann. Statist.*, **38** 3063–3092. URL http://dx.doi.org/10.1214/10-AOS817. MR2722464

[2] Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, **101** 138–156. URL http://dx.doi.org/10.1198/016214505000000907. MR2268032

[3] Berthet, Q. and Rigollet, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013 – The 26th Conference on Learning Theory*, Princeton, NJ, June 12–14, 2013 (S. Shalev-Shwartz and I. Steinwart, eds.), vol. 30 of *JMLR W&CP*, 1046–1066.

[4] Berthet, Q. and Rigollet, P. (2013b). Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41** 1780–1815. MR3127849

[5] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37** 1705–1732. MR2533469

[6] Butucea, C. and Ingster, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, **19** 2652–2688. MR3160567

[7] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35** 2313–2351. MR2382644

[8] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9** 717–772. URL http://dx.doi.org/10.1007/s10208-009-9045-5. MR2565240

[9] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.*, **40** 1935–1967. URL http://dx.doi.org/10.1214/11-AOS949. MR3059067

[10] d'Aspremont, A., Ghaoui, L. E., Jordan, M. I. and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, **49** 434–448. URL http://arxiv.org/abs/cs/0406021v3. MR2353806

[11] Jerrum, M. (1992). Large cliques elude the Metropolis process. *Random Structures Algorithms*, **3** 347–359. URL http://dx.doi.org/10.1002/rsa.3240030402. MR1179827

[12] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, **39** 2302–2329. URL http://dx.doi.org/10.1214/11-AOS894. MR2906869

[13] Kučera, L. (1995). Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, **57** 193–212. Combinatorial optimization 1992 (CO92) (Oxford), URL http://dx.doi.org/10.1016/0166-218X(94)00103-K. MR1327775

[14] Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* To appear.

[15] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, **38** 1287–1319. URL http://dx.doi.org/10.1214/09-AOS691. MR2662343

[16] Rigollet, P. and Tong, X. (2011). Neyman-Pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.*, **12** 2831–2855. MR2854349

[17] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, **39** 731–771. URL http://dx.doi.org/10.1214/10-AOS854. MR2816337

[18] Yannakakis, M. (1991). Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, **43** 441–466. URL http://www.sciencedirect.com/science/article/pii/002200009190024Y. MR1135472