Statistica Sinica Preprint No: SS-2016-0220	
Title	Methods for Sparse and Low-Rank Recovery under Simplex Constraints
Manuscript ID	SS-2016-0220
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0220
Complete List of Authors	Ping Li Syama Sundar Rangapuram and Martin Slawski
Corresponding Author	Ping Li
E-mail	pingli@stat.rutgers.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Statistica Sinica

Methods for Sparse and Low-Rank Recovery under Simplex Constraints

Ping Li, Syama Sundar Rangapuram, and Martin Slawski

Baidu Research; Amazon Research; George Mason University

Abstract: The de facto standard approach of promoting sparsity by means of ℓ_1-regularization becomes ineffective in the presence of simplex constraints, that is, when the target is known to have non-negative entries summing to a given constant. The situation is analogous for the use of nuclear norm regularization for the low-rank recovery of Hermitian positive semidefinite matrices with a given trace. In the present paper, we discuss several strategies to deal with this situation, from simple to more complex. First, we consider empirical risk minimization (ERM), which has similar theoretical properties w.r.t. prediction and ℓ_2-estimation error as ℓ_1-regularization. In light of this, we argue that ERM combined with a subsequent sparsification step (e.g., thresholding) represents a sound alternative to the heuristic of using ℓ_1-regularization after dropping the sum constraint and the subsequent normalization. Next, we show that any sparsity-promoting regularizer under simplex constraints cannot be convex. A novel sparsity-promoting regularization scheme based on the inverse or negative of the squared ℓ_2-norm is proposed, which avoids the shortcomings of various alternative methods from the literature. Our approach naturally extends to Hermitian positive semidefinite

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

matrices with a given trace.

Key words and phrases: simplex constraints, estimation of mixture proportions, sparsity, density matrices of quantum systems, D.C. programming

1. Introduction

In this paper, we study the case in which the parameter of interest β^* is sparse and non-negative with a known sum, i.e., $\beta^* \in c\Delta^p \cap \mathbb{B}_0^p(s)$, where, for $c > 0$ and $1 \leq s \leq p$, $c\Delta^p = \{\beta \in \mathbb{R}_+^p : \mathbf{1}^\top \beta = c\}$ is the (scaled) canonical simplex in $\mathbb{R}^p$, $\mathbb{B}_0^p(s) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s\}$, and $\|\beta\|_0 = |S(\beta)| = |\{j : \beta_j \neq 0\}|$ is referred to as the ℓ_0-norm (the cardinality of the support $S(\beta)$). Unlike the constant c, the sparsity level s is regarded as unknown. The specific value of c is not essential; in the sequel, we shall work with $c = 1$, as for all problem instances studied herein, the data can be re-scaled accordingly. The elements of $\Delta^p = \{\beta \in \mathbb{R}_+^p : \mathbf{1}^\top \beta = 1\}$ can represent probability distributions over p items, proportions, or normalized weights. The following are examples of quantities that arise frequently in contemporary data analyses:

- *Estimation of proportions.* Specific examples include determining the proportions of chemical constituents in a given sample and endmember composition of pixels in hyperspectral imaging (Keshava, 2003).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

- *Probability density estimation*, cf. Bunea et al. (2010). Let $(\mathcal{Z}, \mathcal{A}, P)$ be a probability space, with P having a density f w.r.t. some dominating measure ν. Given a sample $\{Z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and a dictionary $\{\phi_j\}_{j=1}^p$ of densities (w.r.t. ν), the goal is to find a mixture density $\phi_\beta = \sum_{j=1}^p \beta_j \phi_j$ that well approximates f, where $\beta \in \Delta^p$.

- *Convex aggregation/ensemble learning.* The following problem has attracted much interest in the field of non-parametric estimation; see Nemirovski (2000). Let f be the target in a regression/classification problem, and let $\{\phi_j\}_{j=1}^p$ be an ensemble of regressors/classifiers. The goal is to approximate f by a convex combination of $\{\phi_j\}_{j=1}^p$.

- *Markowitz portfolios (Markowitz, 1952) without short positions.* Given assets with expected returns $r = (r_j)_{j=1}^p$ and covariance Σ, the goal is to invest according to proportions $\beta \in \Delta^p$ s.t. the variance $\beta^\top \Sigma \beta$ is minimized, subject to a lower bound on the expected return $\beta^\top r$.

Sparsity is often prevalent or desired in these applications.

- In hyperspectral imaging, a single pixel usually contains few endmembers.

- In density estimation, the underlying density may be concentrated in certain regions of the sample space.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

- In aggregation, it is common to work with a large ensemble to improve the approximation capacity, although specific functions may be well approximated by just a few members of the ensemble.

- Portfolios involving only few assets incur less transaction costs and are easier to manage.

At the same time, promoting sparsity in the presence of the constraint $\beta \in \Delta^p$ appears to be more difficult, as ℓ_1-regularization no longer serves this purpose. As clarified in §2, the naive approach of employing ℓ_1-regularization and dropping the sum constraint results in discarded information. The situation is similar for nuclear norm regularization and low-rank matrices that are Hermitian positive semidefinite, with a fixed trace. For example, this arises in quantum state tomography (Gross et al., 2010) when the constraint set results as $\mathbf{\Delta}^m = \{B \in \mathbb{C}^{m\times m} : B = B^H, B \succeq 0, \mathrm{tr}(B) = 1\}$, with ^{H} denoting conjugate transposition. Thus, the presence of simplex constraints and their matrix counterparts require that we use different strategies to deal with sparsity and low-rankedness. Here, we propose strategies that are statistically sound, straightforward to implement, adaptive, in the sense that the sparsity level s (resp., the rank in the matrix case) is not required to be known, and work with a minimum amount of hyperparameter tuning.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Related work. The problem outlined above is discussed well by Kyrillidis et al. (2013). They consider the sparsity level s to be known, and suggest dealing with the constraint set $\Delta_0^p(s) = \Delta^p \cap \mathbb{B}_0^p(s)$ by projected gradient descent based on a near-linear time algorithm used to compute the projection. This approach can be viewed as a natural extension of iterative hard thresholding (IHT, Blumensath and Davies (2009); Shen and Li (2018)).

Pilanci, Ghaoui, and Chandrasekaran (2012) suggest using the regularizer $\beta \mapsto 1/\left\|\beta\right\|_\infty$ to promote sparsity on Δ^p. In addition, they show that in the case of squared loss, the resulting nonconvex optimization problem can be reduced to p second-order cone programs. In practice, however, the computational cost quickly becomes prohibitive, particularly when combined with the tuning of the regularization parameter.

Relevant prior studies include the works of Larsson and Ugander (2011) and Shashanka, Raj, and Smaragdis (2008), who discuss the aforementioned problem in the context of latent variable models for image and bag-of-words data. Larsson and Ugander (2011) propose a so-called pseudo-Dirichlet prior, akin to the log-penalty in Candes, Wakin, and Boyd (2007). Shashanka, Raj, and Smaragdis (2008) suggest using Shannon entropy as a regularizer. A conceptually different approach is pursued in Jojic, Saria, and Koller (2011). Instead of the usual loss + ℓ_1-norm formulation with the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

ℓ_1-norm arising as the convex envelope of the ℓ_0-norm on the unit ℓ_∞-ball, the authors work with the convex envelope of the loss + ℓ_0-norm.

Finally, it is worth mentioning a line of research on sparse regression under linear inequality or equality constraints. Here, relevant works include those of James, Paulson, and Rusmevichientong (2015), Lin et al. (2014), and Shi, Zhang, and Li (2016). Meinshausen (2013) and Slawski and Hein (2013) study the case in which the constraint set is the non-negative orthant. It is shown that, under specific conditions, this constraint has similar effects to those of ℓ_1-regularization. With simplex constraints, this effect applies more broadly, as discussed in §2.

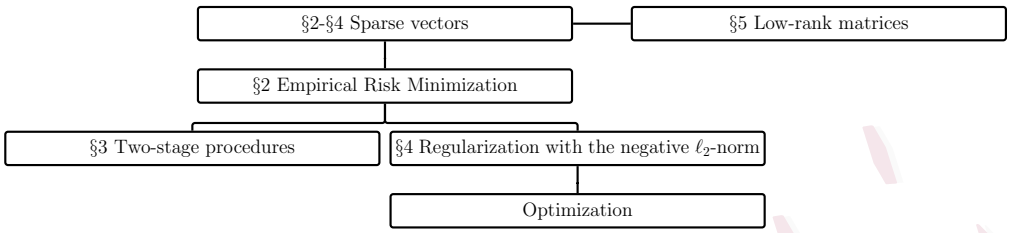
Outline and contributions. As a preliminary step, we provide a brief analysis of high-dimensional estimations under simplex constraints in §2. Such analyses provide valuable insights when designing sparsity-promoting schemes. Note that empirical risk minimization (ERM) and the elements of Δ^p contained in a "high confidence set" for β^* (a construction inspired by the Dantzig selector of Candes and Tao (2007)) already enjoy nice statistical guarantees, including adaptation to sparsity under a restricted strong convexity condition that is weaker than that in Negahban et al. (2012). Next, we discuss strategies to improve on ERM, particularly with respect

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

to the sparsity of the solution and support recovery. As a basic strategy, we consider simple two-stage procedures, thresholding and reweighted ℓ_1-regularization on top of ERM (see §3).

As an alternative, we propose a novel regularization-based scheme in §4, in which $\beta \mapsto 1/\left\|\beta\right\|_2^2$ serves as a relaxation of the ℓ_0-norm on Δ^p. This regularizer naturally extends to the case of positive semidefinite Hermitian matrices of unit trace, as discussed in §5. On the optimization side, the approach can be implemented using difference-of-convex (DC) programming (Pham Dinh and Le Thi, 1997). Unlike other forms of concave regularization, such as the SCAD, capped ℓ_1, or MCP penalties (Zhang and Zhang, 2013) no parameter other than the regularization parameter needs to be specified. For this purpose, we employ a generic BIC-type criterion (Schwarz (1978); Kim, Kwon, and Choi (2012)) with the aim of selecting the correct model (resp., rank, in the matrix case). The Supplementary Material (Li, Rangapuram, and Slawski, 2018) contains all proofs, as well as numerical experiments on compressed sensing, density estimation, portfolio optimization and quantum state tomography that demonstrate the effectiveness of both the two-stage procedures and the regularization-based approach. The following orgchart provides a quick overview of the organization of the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220



Notation. For the convenience of readers, we first present the essential notation. We denote $\|\cdot\|_q$ for $q \in [0, \infty]$, as the usual ℓ_q-norm or the Schatten ℓ_q-norm, depending on the context, and $\langle \cdot, \cdot \rangle$ as the usual Euclidean inner product. We use $|\cdot|$ for the cardinality of a set. The support of $v \in \mathbb{R}^d$ is denoted by $S(v) = \{j : v_j \neq 0\}$. For $J \subseteq \{1, \ldots, d\}$, we let $v_J = (v_j)_{j \in J}$. We write $\mathbf{I}(\cdot)$ for the indicator function. We denote $\{e_1, \ldots, e_d\}$ as the canonical basis of $\mathbb{R}^d$. For $A \subseteq \mathbb{R}^d$, $\Pi_A : \mathbb{R}^d \to A$ denotes the Euclidean projection on A. For the functions $f(n)$ and $g(n)$, we write $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ if $f(n) \geq Cg(n)$ and $f(n) \leq Cg(n)$, respectively, for some constant $C > 0$. We write $f(n) \asymp g(n)$ if both $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$. We also use the Landau symbols $O(\cdot)$ and $o(\cdot)$.

2. Simplex constraint in high-dimensional problems: Basic analysis

Before designing schemes that promote sparsity under the constraint $\beta \in \Delta^p$, it is worth deriving basic performance bounds and establishing adap-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

tivity to underlying sparsity when only simplex constraints are used for the estimation, without explicitly enforcing sparse solutions. Note that the constraint $\beta \in \Delta^p$ is stronger than the ℓ_1-ball constraint, $\|\beta\|_1 \leq 1$. As a result, ERM enjoys properties known from analyses of (unconstrained) ℓ_1-regularized estimations, including the adaptivity to sparsity under certain conditions. This already sets a substantial limit on what can be achieved by sparsity-promoting schemes.

Let $\{Z_i\}_{i=1}^n$ be independently and identically distributed (i.i.d.) copies of a random variable Z following a distribution P on a sample space $\mathcal{Z} \subseteq \mathbb{R}^d$. Let $L : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}$ be a loss function, such that $\forall z \in \mathcal{Z}$, $L(\cdot, z)$ is convex and differentiable. For $\beta \in \mathbb{R}^p$, $R(\beta) = \mathbf{E}[L(\beta, Z)]$ denotes the expected risk, and $R_n(\beta) = n^{-1} \sum_{i=1}^n L(\beta, Z_i)$ denotes its empirical counterpart. The goal is to recover $\beta^* = \operatorname{argmin}_{\beta \in \Delta^p} \mathbf{E}[L(\beta, Z)]$. ERM yields $\widehat{\beta} \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta)$. Figure 1 provides an overview of the key quantities and their relationships.

In addition to ERM, our analysis simultaneously covers all elements of the set

$$\mathcal{D}(\lambda) = \{\beta \in \Delta^p : \|\nabla R_n(\beta)\|_\infty \leq \lambda\}, \tag{2.1}$$

for suitably chosen $\lambda \geq 0$, as discussed below. The construction of $\mathcal{D}(\lambda)$ is inspired by the constraint set of the Dantzig selector (Candes and Tao,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

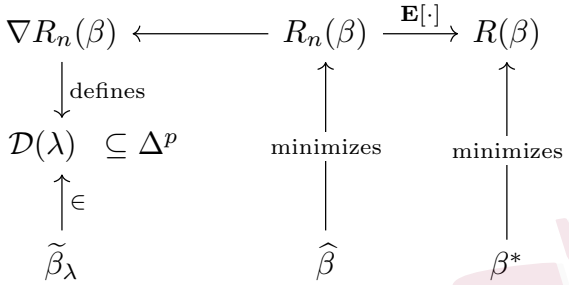


Figure 1: Diagram summarizing the relationships between the quantities employed in this study.

2007), which is extended to general loss functions by Lounici (2008); James and Radchenko (2009); Fan (2013). Elements of $\mathcal{D}(\lambda)$ are shown to have performance comparable to $\widehat{\beta}$. The set $\mathcal{D}(\lambda)$ need not be convex, in general. For squared loss, it becomes a convex polyhedron, which is nonempty as long as $\lambda \geq \lambda_*$, where $\lambda_* = \|\nabla R_n(\beta^*)\|_\infty$. In many settings of interest (cf., Lounici (2008); Negahban et al. (2012)), it can be shown that

$$\mathbf{P}(\lambda_* \gtrsim \sqrt{\log(p)/n}) = o(1) \quad \text{as } n \to \infty. \tag{2.2}$$

2.1 Excess risk

The first result bounds the excess risk of $\widehat{\beta}$ and $\widetilde{\beta}_\lambda$, where in what follows, $\widetilde{\beta}_\lambda$ represents an arbitrary element of $\mathcal{D}(\lambda)$ in (2.1).

Proposition 1. *For $\beta \in \mathbb{R}^p$, let $\psi_n(\beta) = R_n(\beta) - R(\beta)$ and $\overline{\psi}_n(\beta) =$*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

$\psi_n(\beta) - \psi_n(\beta^*)$. *For* $r > 0$, *let* $\mathbb{B}_1^p(r;\beta^*) = \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_1 \le r\}$ *denote the* ℓ_1*-ball of radius* r *centered at* β^* *and* $\overline{\Psi}_n(r) = \sup\{|\overline{\psi}_n(\beta)| : \beta \in \mathbb{B}_1^p(r;\beta^*)\}$. *We then have*

$$R(\widehat{\beta}) - R(\beta^*) \le \overline{\Psi}_n(\|\widehat{\beta} - \beta^*\|_1) \le \overline{\Psi}_n(2),$$

$$R(\widetilde{\beta}_\lambda) - R(\beta^*) \le \overline{\Psi}_n(\|\widetilde{\beta}_\lambda - \beta^*\|_1) + \lambda\|\widetilde{\beta}_\lambda - \beta^*\|_1 \le \overline{\Psi}_n(2) + 2\lambda.$$

The excess risk of ERM and points in $\mathcal{D}(\lambda)$ can thus be bounded by controlling $\overline{\Psi}_n(r)$, the supremum of the empirical process $\overline{\psi}_n(\beta)$ over all β, with ℓ_1-distance at most r from β^*. This supremum is well studied in the literature on ℓ_1-regularization. For example, for linear regression with a fixed or random sub-Gaussian design and sub-Gaussian errors, as well as for a Lipschitz loss (e.g., logistic loss), it can be shown that (van de Geer, 2008)

$$\mathbf{P}(\overline{\Psi}_n(r) \gtrsim r\sqrt{\log(p)/n}) = o(1) \quad \text{as } n \to \infty. \tag{2.3}$$

Using $\widehat{\beta} \in \Delta^p$ and $\widetilde{\beta}_\lambda \in \Delta^p$, choosing $\lambda \asymp \lambda_*$, and invoking (2.2), Proposition 1 yields that the excess risk of ERM and points in $\mathcal{D}(\lambda)$ scale as $O(\sqrt{\log(p)/n})$. As a result, ERM and finding a point in $\mathcal{D}(\lambda)$ constitute persistent procedures, in the sense of Greenshtein and Ritov (2004).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

2.2 Adaptation to sparsity

Proposition 1 does not entail further assumptions on β^* or R_n. In this subsection, we suppose that $\beta^* \in \Delta_0^p(s)$ and that R_n obeys a restricted strong convexity (RSC) condition, defined as follows. Consider the set

$$\mathcal{C}^{\Delta}(s) = \{\delta \in \mathbb{R}^p : \exists J \subseteq \{1, \ldots, p\}, |J| \leq s \quad \text{s.t. } \mathbf{1}^\top \delta_{J^c} = -\mathbf{1}^\top \delta_J, \;\; \delta_{J^c} \succeq 0\}. \tag{2.4}$$

Observe that $\{\beta - \beta^* : \; \beta \in \Delta^p\} \subseteq \mathcal{C}^{\Delta}(s)$. For the next result, we require R_n to be strongly convex over $\mathcal{C}^{\Delta}(s)$.

Condition 1. We say that the Δ-*RSC* condition is satisfied for sparsity level $1 \leq s \leq p$ and constant $\kappa > 0$ if, for all $\beta \in \Delta_0^p(s)$ and $\delta \in \mathcal{C}^{\Delta}(s)$,

$$R_n(\beta + \delta) - R_n(\beta) - \nabla R_n(\beta)^\top \delta \geq \kappa \|\delta\|_2^2.$$

Condition 1 is an adaptation of a condition employed in Negahban et al. (2012) for the analysis of (unconstrained) ℓ_1-regularized ERM. Note that for squared loss, Condition 1 becomes the restricted eigenvalue condition in Bickel, Ritov, and Tsybakov (2009), the range of validity of which has been investigated by, among others, Raskutti, Wainwright, and Yu (2010); Rudelson and Zhou (2013); Lecue and Mendelson (2017). Our condition here is weaker, because the RSC condition in Negahban et al. (2012) is over

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

the larger set

$$\mathcal{C}(\alpha, s) = \{\delta \in \mathbb{R}^p : \exists J \subseteq \{1, \ldots, p\}, |J| \leq s \quad \text{s.t. } \|\delta_{J^c}\|_1 \leq \alpha \|\delta_J\|_1\},$$

for $\alpha \geq 1$. We can now state a second set of bounds.

Proposition 2. *Let the* Δ-RSC *condition hold for sparsity level* s *and* $\kappa > 0$. *We then have*

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq \frac{4s\lambda_*^2}{\kappa^2}, \quad \|\widetilde{\beta}_\lambda - \beta^*\|_2^2 \leq \frac{4s(\lambda + \lambda_*)^2}{\kappa^2},$$
$$\|\widehat{\beta} - \beta^*\|_1 \leq \frac{4s\lambda_*}{\kappa}, \quad \|\widetilde{\beta}_\lambda - \beta^*\|_1 \leq \frac{4s(\lambda + \lambda_*)}{\kappa}.$$

Invoking (2.2) and choosing $\lambda \asymp \lambda_*$, we recover the rates $O(s\log(p)/n)$ for the squared ℓ_2-error and $O(s\sqrt{\log(p)/n})$ for the ℓ_1-error, respectively. Combining the bounds on the ℓ_1-error with (2.3) and Proposition 1, we obtain

$$R(\widehat{\beta}) - R(\beta^*) \lesssim \frac{s\log p}{n}, \quad R(\widetilde{\beta}_\lambda) - R(\beta^*) \lesssim \frac{s\log p}{n}.$$

The above rates are known to be minimax optimal for the parameter set $\mathbb{B}_0^p(s)$ and squared loss (Ye and Zhang, 2010). Thus under the Δ-RSC condition, there does not seem to be much room for improving over $\widehat{\beta}$ and $\widetilde{\beta}_\lambda$ as far as the ℓ_1-error, ℓ_2-error, and excess risk are concerned. An additional advantage of $\widehat{\beta}$ is that it does not depend on any hyperparameters.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

3. Two-stage Procedures

While $\widehat{\beta}$ has appealing adaptation properties with regard to underlying sparsity, $\|\widehat{\beta}\|_0$ may be significantly larger than the sparsity level s. Note that the ℓ_2-bound of Proposition 2 yields that $S(\widehat{\beta}) \supseteq S(\beta^*)$ as long as $b^*_{\min} \gtrsim \lambda^* \sqrt{s}$, where $b^*_{\min} = \min\{\beta^*_j : j \in S(\beta^*)\}$. If the aim is an estimator $\widehat{\theta}$ that achieves support recovery, that is, $S(\widehat{\theta}) = S(\beta^*)$, $\widehat{\beta}$ needs to be further sparsified by a suitable form of post-processing. Here, we consider two schemes, namely thresholding and weighted ℓ_1-regularization:

Stage 1	**Stage 2**	
Compute $\widehat{\beta}$	thresholding: $\widehat{\beta}_\tau = (\widehat{\beta}_j \cdot \mathbf{I}(\widehat{\beta}_j \geq \tau))_{1 \leq j \leq p}$	(3.1)
	or weighted ℓ_1: $\widehat{\beta}^w_\lambda \in \operatorname*{argmin}_{\beta \in \Delta^p} R_n(\beta) + \lambda \langle w, \beta \rangle$,	(3.2)

where $\mathbf{I}(\cdot)$ denotes the indicator function and $w = (w_j)_{j=1}^p$ are non-negative weights. We restrict ourselves to the common choice $w_j = 1/\widehat{\beta}_j$ if $\widehat{\beta}_j > 0$, and $w_j = +\infty$ otherwise (s.t. $(\widehat{\beta}^w_\lambda)_j = 0$), for $j = 1, \ldots, p$. Note that weighted ℓ_1-regularization is often referred to as the "adaptive lasso" method (Zou, 2006).

While its primary purpose is model selection, thresholding (3.1) can optionally be complemented by a refitting step with fixed support, that is, ERM with the additional constraints $\beta_j = 0 \; \forall j \notin S(\widehat{\beta}_\tau)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

A third approach is to ignore the unit sum constraint first, such that ℓ_1-regularization has a sparsity-promoting effect, and then to divide the output by its sum as a simple way to satisfy the following constraint:

Stage 1	**Stage 2**	
$\widehat{\beta}^{\ell_1}_\lambda \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p_+} R_n(\beta) + \lambda \mathbf{1}^\top \beta$	Normalize: $\widehat{\beta}^{\ell_1}_\lambda / (\mathbf{1}^\top \widehat{\beta}^{\ell_1}_\lambda)$.	(3.3)

From the point of view of optimization, (3.3) offers several advantages. Non-negativity constraints alone tend to be easier to handle than simplex constraints. For projected gradient-type algorithms, the projection on the constraint set becomes trivial. Moreover, coordinate descent is applicable because non-negativity constraints do not couple several variables (whereas simplex constraints do). Coordinate descent is one of the fastest algorithms for sparse estimation (Friedman, Hastie, and Tibshirani, 2010; Mazumder, Friedman, and Hastie, 2011), particularly for large values of λ. On the other hand, from a statistical perspective, (3.3) is an ad hoc rather than a well-grounded approach. It is advisable to prefer $\widehat{\beta}$ because it incorporates all given constraints into the optimization problem, which leads to a weaker RSC condition and eliminates the need to specify λ appropriately. Indeed, taking a large value of λ in (3.3) in order to obtain a highly sparse solution increases the bias and may lead to false negatives. In addition, (3.3) may also lead to false positives if the "irrepresentable condition" (Zhao and Yu,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

2006) is violated. Our experimental results (cf., Supplementary Material, Li, Rangapuram, and Slawski (2018)) confirm that (3.3) has a considerably larger estimation error than that of ERM.

Model selection. In this paragraph, we briefly discuss a data-driven approach for selecting the parameters τ and λ in (3.1) and (3.2) when the aim is support recovery. It suffices to pick τ from $T = \{\widehat{\beta}_j\}_{j=1}^p$, whereas for (3.2), we consider a finite set $\Lambda \subset \mathbb{R}^+$. We first obtain $\{\widehat{\beta}_\tau,\ \tau \in T\}$ or $\{\widehat{\beta}^w_\lambda,\ \lambda \in \Lambda\}$, and then select one of the candidate models induced by the support set $\{S(\widehat{\beta}_\tau),\ \tau \in T\}$ or $\{S(\widehat{\beta}^w_\lambda),\ \lambda \in \Lambda\}$, respectively. Model selection can be performed using a hold-out data set or an appropriate model selection criterion, such as the RIC in the case of squared loss (Foster and George, 1994). Specifically, let $Z_i = (X_i, Y_i)$, for $i = 1, \ldots, n$, and suppose that

$$Y_i = X_i^\top \beta^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n. \tag{3.4}$$

Then, for $S \subseteq \{1, \ldots, p\}$, the RIC is defined as

$$\mathrm{RIC}(S) = \min_{\beta \in \mathbb{R}^p : \beta_{S^c} = 0} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 + \frac{2\sigma^2 \log p}{n} |S|. \tag{3.5}$$

While computationally intractable in general, minimizing (3.5) has been shown to be model-selection consistent in high-dimensional regimes (Kim et al., 2012; Zhang and Zhang, 2013). Here, we minimize (3.5) over $\{S(\widehat{\beta}_\tau)\}_{\tau \in T}$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

or $\{S(\widehat{\beta}^w_\lambda)\}_{\lambda\in\Lambda}$ only. The rationale is that support recovery is still achieved whenever the RIC is satisfied, provided that

$$S(\beta^*) \in \{S(\widehat{\beta}_\tau)\}_{\tau\in T} \quad (3.6) \qquad \text{or} \quad S(\beta^*) \in \{S(\widehat{\beta}^w_\lambda)\}_{\lambda\in\Lambda}. \quad (3.7)$$

Condition (3.6) is met if $\min_{j\in S(\beta^*)} \widehat{\beta}_j > \max_{j\in S(\beta^*)^c} \widehat{\beta}_j$, which can in turn be deduced from a bound on $\|\widehat{\beta} - \beta^*\|_2$ (cf., Proposition 2) and a corresponding lower bound on $b^*_{\min} = \min\{\beta^*_j : j \in S(\beta^*)\}$. For weighted ℓ_1-regularization, (3.7) is implied by a similar, albeit slightly more stringent condition.

Proposition 3. *Consider model* (3.4) *with* $\{X_i\}_{i=1}^n$ *deterministic, such that* $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ *for all* j*, and* $\widehat{\beta}^w_\lambda$ *in* (3.2) *with* $R_n(\beta) = \frac{1}{2n}\sum_{i=1}^n (Y_i - X_i^\top\beta)^2$*. Then,* (3.7) *is satisfied with probability at least* $1 - O(p^{-1})$ *if*

i) $\min\limits_{j\in S(\beta^*)} \widehat{\beta}_j \gtrsim \max\limits_{j\in S(\beta^*)^c} \widehat{\beta}_j$,

ii) $\Lambda \ni \lambda$ *s.t.* $\lambda = \min\limits_{j\in S(\beta^*)} \widehat{\beta}_j\lambda_0$ *with* $\sigma\sqrt{\log(p)/n} \lesssim \lambda_0 \lesssim b^*_{\min}$.

The constants hidden in $\gtrsim$ and $O(\cdot)$ are provided in the proof of the above statement.

On a practical note, we point out that consistent model selection based on the RIC (3.5) presumes knowledge of σ, or an estimator $\widehat{\sigma}$ obeying at least $\widehat{\sigma} \asymp \sigma$ (Kim et al., 2012). We refer to Sun and Zhang (2012); Fan,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Guo, and Hao (2012); Dicker (2014); Reid, Tibshirani, and Friedman (2016) for specific estimators $\widehat{\sigma}$.

4. Regularization with the negative ℓ_2-norm

A concern with ERM (optionally followed by a sparsification step) is that potential prior knowledge about sparsity is not incorporated into the estimation. The hope is that by taking sparsity into account, the guarantees of §2 can be improved. In particular, it may be possible to weaken Condition 1.

It turns out that any sparsity-promoting regularizer Ω on Δ^p cannot be convex. To see this, note that if Ω is sparsity-promoting and homogeneous across coordinates, it should assign strictly smaller values to any of the vertices $\{e_j\}_{j=1}^p$ of Δ^p (which are maximally sparse) than to its barycentre (which is maximally dense); that is,

$$\Omega(e_j) < \Omega(\{e_1 + \ldots + e_p\}/p), \quad j = 1, \ldots, p. \tag{4.1}$$

However, (4.1) contradicts the convexity of Ω, because by Jensen's inequality,

$$\Omega(\{e_1 + \ldots + e_p\}/p) \leq \{\Omega(e_1) + \ldots \Omega(e_p)\}/p.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

4.1 Approach

For $0 \neq \beta \in \mathbb{R}^p$, consider $\Omega(\beta) = \|\beta\|_1^2/\|\beta\|_2^2$. Here, Ω can be viewed as a "robust" measure of sparsity. We have $\|\beta\|_0 \geq \Omega(\beta)$, with equality holding iff $\{|\beta_j|,\ j \in S(\beta)\}$ is constant. By "robustness" we mean that Ω is small for vectors that have few entries of large magnitude, whereas the number of nonzero elements may be as large as p. From $\|\beta\|_2^2 \leq \|\beta\|_\infty \|\beta\|_1$, we have the alternative, $\overline{\Omega}(\beta) = \|\beta\|_1/\|\beta\|_\infty$. As $\|\beta\|_1 = 1\ \forall \beta \in \Delta^p$, we have

$$\frac{1}{\|\beta\|_\infty} \leq \frac{1}{\|\beta\|_2^2} \leq \|\beta\|_0 \quad \forall \beta \in \Delta^p. \tag{4.2}$$

The map $\beta \mapsto 1/\|\beta\|_\infty$ is proposed as a sparsity-promoting regularizer on Δ^p by Pilanci, Ghaoui, and Chandrasekaran (2012). It yields a looser lower bound on $\beta \mapsto \|\beta\|_0$ than that of the map $\beta \mapsto 1/\|\beta\|_2^2$ advocated in the present work. Both lower bounds are sparsity-promoting on Δ^p as indicated by Figure 2.

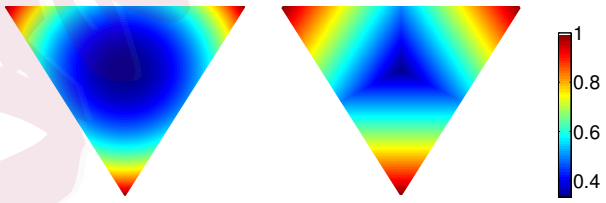


Figure 2: Contours of $\beta \mapsto \|\beta\|_2^2$ (left) and $\beta \mapsto \|\beta\|_\infty$ (right) on Δ^3.

This lets us propose the following modifications of $\widehat{\beta}$ and $\widetilde{\beta}_\lambda$, respec-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

tively,

$$\widehat{\beta}^{\ell_2}_{\lambda} \in \operatorname*{argmin}_{\beta \in \Delta^p} R_n(\beta) - \lambda \|\beta\|_2^2, \tag{4.3}$$

$$\widetilde{\beta}^{\ell_2}_{\lambda} \in \operatorname*{argmin}_{\beta \in \mathcal{D}(\lambda)} -\|\beta\|_2^2, \text{ with } \mathcal{D}(\lambda) \text{ as in (2.1).} \tag{4.4}$$

Note the correspondence of (4.3) / (4.4) on the one hand, and the lasso (resp., Dantzig selector) on the other hand.

For (4.3), it appears to be better to use $1/\|\beta\|_2^2$ rather than of $-\|\beta\|_2^2$, given (4.2). Eventually, this becomes a matter of parameterization. Although $\beta \mapsto \|\beta\|_0$ is the canonical measure of sparsity, $\beta \mapsto -1/\|\beta\|_0$ provides another measure. It is lower bounded by $\beta \mapsto -1/\|\beta\|_2^2$. We prefer the negative over the inverse, for computational reasons: the optimization problem in (4.3) is a DC program (Pham Dinh and Le Thi, 1997) and, hence, is more amenable to optimization. The problem in (4.4) is also a DC program if $\mathcal{D}(\lambda)$ is convex. Note that for (4.4), minimizing the negative ℓ_2-norm is equivalent to minimizing the inverse ℓ_2-norm.

4.2 Least squares denoising

In order to show that the negative ℓ_2-norm combined with simplex constraints promotes exactly sparse solutions, we elaborate on (4.3) in the simple setup of denoising. Let $Z_i = \beta^*_i + \varepsilon_i$, for $i = 1, \ldots, n = p$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

$\beta^* \in \Delta_0^n(s)$ and $\{\varepsilon_i\}_{i=1}^n$ represents random noise. We consider squared loss, i.e., $L(\beta, Z_i) = (Z_i - \beta)^2$, $i = 1, \ldots, n$. This yields the optimization problem

$$\min_{\beta \in \Delta^n} \frac{1}{n}\|\mathbf{Z} - \beta\|_2^2 - \lambda\|\beta\|_2^2, \quad \mathbf{Z} = (Z_i)_{i=1}^n. \tag{4.5}$$

As stated below, (4.5) can be recast as a Euclidean projection of $\mathbf{Z}/\gamma$ on Δ^n, where γ is a function of λ. Using this property, we derive conditions on β^* and λ such that $\widehat{\beta}_\lambda^{\ell_2}$ achieves support recovery.

Proposition 4. *Consider* (4.5) *and suppose that* $z_{(1)} > \ldots > z_{(n)}$, *where* $\{z_{(i)}\}_{i=1}^n$ *denotes the ordered realizations of* $\{Z_i\}_{i=1}^n$. *For all* $\lambda \geq 1/n$, *we have* $\widehat{\beta}_\lambda^{\ell_2} = \big(\mathbf{I}(Z_i = z_{(1)})\big)_{i=1}^n$. *For all* $0 \leq \lambda < 1/n$, *we have* $\widehat{\beta}_\lambda^{\ell_2} = \operatorname{argmin}_{\beta \in \Delta^n}\|\mathbf{Z}/\gamma - \beta\|_2^2$, *where* $\gamma = 1 - n\lambda$. *Moreover, if* $2s \max_{1 \leq i \leq n} |\varepsilon_i|/n < \lambda < 1/n$ *and* $b^*_{\min} > (n\lambda)/s + 2\max_{1 \leq i \leq n} |\varepsilon_i|$, *we have* $S(\widehat{\beta}_\lambda^{\ell_2}) = S(\beta^*)$.

In particular, for $\lambda = (1 + \delta)2s \max_{1 \leq i \leq n} |\varepsilon_i|/n$, for any $\delta > 0$, the required lower bound on $b^*_{\min}$ becomes $4(1 + \delta)\max_{1 \leq i \leq n} |\varepsilon_i|$. For the sake of reference, note that in the Gaussian sequence model with $\varepsilon_i \sim N(0, \sigma^2/n)$ (cf., Johnstone (2013)), we have $\max_{1 \leq i \leq n} |\varepsilon_i| \asymp \sqrt{\log(n)/n}$.

The denoising problem (4.5) can be viewed as a least squares regression problem in which the design matrix is the identity matrix. For general design matrices, the analysis becomes more difficult, particularly because the optimization problem may be neither convex nor concave. In the latter case, the minimum is attained at one of the vertices of Δ^p.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

4.3 Optimization

Both (4.3) and (4.4) are nonconvex in general. Furthermore, maximizing the Euclidean norm over a convex set is NP-hard in general (Pardalos and Vavasis, 1991). To solve these two problems, we exploit the fact that both objectives are in DC form, that is, they can be represented as $f(\beta) = g(\beta) - h(\beta)$, with g and h both being convex. Linearizing $-h$ at a given point yields a convex majorant of f that is tight at that point. Repeatedly minimizing the majorant yields an iterative procedure known as the concave-convex procedure (CCCP, Yuille and Rangarajan (2003)), which falls into the more general framework of majorization-minimization (MM) algorithms (Lange, Hunter, and Yang, 2000). When applied to (4.3) and (4.4), this approach yields Algorithm 1.

For the second part of Algorithm 1 to be practical, we assume that $\mathcal{D}(\lambda)$ is convex. The above algorithms can be shown to yield strict descent until convergence to a limit point satisfying the first-order optimality condition of problems (4.3)/(4.4). This is the content of the next proposition.

Proposition 5. *Let f denote the objective in* (4.3) *or* (4.4)*. The elements of the sequence $\{\beta^k\}_{k\geq 0}$ produced by Algorithm 1 satisfy $f(\beta^{k+1}) < f(\beta^k)$ until convergence. Moreover, the limit satisfies the first-order optimality condition of the respective problem.*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Algorithm 1

(4.3): $\min_{\beta\in\Delta^p} R_n(\beta) - \lambda\|\beta\|_2^2$

Initialization: $\beta^0 \in \Delta^p$

repeat $\beta^{k+1} \in \operatorname{argmin}_{\beta\in\Delta^p} R_n(\beta) - 2\left\langle \beta^k, \beta - \beta^k \right\rangle$

until $R_n(\beta^{k+1}) - 2\left\langle \beta^k, \beta^{k+1} \right\rangle = R_n(\beta^k)$

(4.4): $\min_{\beta\in\mathcal{D}(\lambda)} -\|\beta\|_2^2$

Initialization: $\beta^0 \in \mathcal{D}(\lambda)$

repeat $\beta^{k+1} \in \operatorname{argmin}_{\beta\in\mathcal{D}(\lambda)} -2\left\langle \beta^k, \beta - \beta^k \right\rangle$

until $\left\langle \beta^k, \beta^{k+1} - \beta^k \right\rangle = 0$

An appealing feature of Algorithm 1 is that solving each subproblem in the repeat step involves only minor modifications to the computational approaches used for ERM (resp. finding a feasible point in $\mathcal{D}(\lambda)$). With R_n assumed to be convex, ERM is a convex optimization problem. If R_n is also smooth, off-the-shelf algorithms such as interior point methods, projected gradient descent, and conditional gradient descent (Bertsekas, 1999) can be employed. For common nonsmooth losses, such as an absolute loss or hinge loss, ERM can be converted into a linear program. For the squared loss and absolute loss, specialized algorithms are proposed in Vila and Schniter (2014).

When selecting the parameter λ using a grid search, we suggest solving

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

the associated instances of (4.3)/(4.4) from the smallest to the largest value of λ, using the solution from the current instance as the initial iterate for the next one. For the smallest value of λ, we recommend using $\widehat{\beta}$ and any point from $\mathcal{D}(\lambda)$ as the initial iterate for (4.3) and (4.4), respectively. Running Algorithm 1 for formulation (4.4) has the advantage that all iterates are contained in $\mathcal{D}(\lambda)$, and thus enjoy at least the statistical guarantees of $\widetilde{\beta}_\lambda$ derived in §2. According to our numerical results, formulation (4.3) achieves better performance (cf., supplement Li, Rangapuram, and Slawski (2018)).

5. Extension to the matrix case

As pointed out in the introduction, there is a matrix counterpart to the aforementioned problem in which the object of interest is a low-rank Hermitian positive semidefinite matrix of unit trace. This set of matrices includes the density matrices of quantum systems (Nielsen and Chuang, 2000). The task of reconstructing such density matrices from so-called observables (e.g., noisy linear measurements) is termed quantum state tomography (Paris and Rehacek, 2004). In the past few years, quantum state tomography based on Pauli measurements has attracted considerable interest in the field of mathematical signal processing and statistics (Gross et al., 2010; Gross, 2011; Koltchinskii, 2011; Wang, 2013; Cai et al., 2016).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Specifically, the setup we employ throughout this section is as follows. Let $\mathbb{H}^m = \{B \in \mathbb{C}^{m\times m} : B = B^H\}$ be the Hilbert space of complex Hermitian matrices with inner product $\langle F, G\rangle = \mathrm{tr}(FG)$, $(F, G) \in \mathbb{H} \times \mathbb{H}$, and, henceforth, let $\|\cdot\|_q$, for $0 \le q \le \infty$, denote the Schatten q-"norm" of a Hermitian matrix, defined as the ℓ_q-norm of its eigenvalues. Here, $\|\cdot\|_0$ denotes the number of nonzero eigenvalues, or equivalently, the rank. We suppose that the target B^* is contained in $\boldsymbol{\Delta}_0^m(r) := \boldsymbol{B}_0^m(r) \cap \boldsymbol{\Delta}^m$, where

$$\boldsymbol{B}_0^m(r) := \{B \in \mathbb{H}^m : \ \|B\|_0 \le r\}, \qquad \boldsymbol{\Delta}^m := \{B \in \mathbb{H}^m : B \succeq 0,\ \mathrm{tr}(B) = 1\}.$$

That is, B^* is also positive semidefinite, of unit trace, and has rank at most r. In low-rank matrix recovery, the Schatten 1-norm (typically referred to as the nuclear norm) is commonly used as a convex surrogate for the rank (Recht, Fazel, and Parillo, 2010). Because the nuclear norm is constant over $\boldsymbol{\Delta}^m$, a different strategy is needed to promote low-rankedness under that constraint. In the sequel, we carry over our treatment of the vector case to the matrix case. The analogies are mostly direct; at certain points, however, the matrix case yields additional complications, as detailed below. For simplicity, we restrict ourselves to the setup in which $Z_i = (X_i, Y_i)$ are such that

$$Y_i = \langle X_i, B^*\rangle + \varepsilon_i, \ \ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n, \tag{5.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

with $\{X_i\}_{i=1}^n \subset \mathbb{H}^m$. Equivalently,

$$\mathbf{Y} = \mathcal{X}(B^*) + \varepsilon, \;\; \mathbf{Y} = (Y_i)_{i=1}^n, \; \varepsilon = (\varepsilon_i)_{i=1}^n,$$

where $\mathcal{X} : \mathbb{H}^m \to \mathbb{R}^n$ is a linear operator defined by $(\mathcal{X}(B))_i = \langle X_i, B\rangle$, $B \in \mathbb{H}^m$, for $i = 1, \ldots, n$. We consider squared loss; that is, for $B \in \boldsymbol{\Delta}^m$, the empirical risk is given by

$$R_n(B) = \|\mathbf{Y} - \mathcal{X}(B)\|_2^2/n.$$

Basic estimators. As basic estimators, we consider the empirical risk minimization given by $\widehat{B} \in \operatorname{argmin}_{B\in\boldsymbol{\Delta}^m} R_n(B)$, as well as $\widetilde{B}_\lambda$, where $\widetilde{B}_\lambda$ is any point in the set

$$\begin{aligned} \boldsymbol{D}(\lambda) &= \{B \in \boldsymbol{\Delta}^m : \|\nabla R_n(B)\|_\infty \le \lambda\} \\ &= \Big\{B \in \boldsymbol{\Delta}^m : \frac{2}{n}\|\mathcal{X}^\star(\mathcal{X}(B) - y)\|_\infty \le \lambda\Big\}, \end{aligned} \tag{5.2}$$

where $\mathcal{X}^\star : \mathbb{R}^n \to \mathbb{H}^m$ is the adjoint of $\mathcal{X}$. Both $\widehat{B}$ and $\widetilde{B}_\lambda$ adapt to the rank of B^* under a restricted strong convexity condition. For $B \in \boldsymbol{B}_0^m(r)$, let $\mathbb{T}(B)$ be the tangent space of $\boldsymbol{B}_0^m(r) \subset \mathbb{H}^m$ at B (see Definition 1 in Supplementary Material, Li, Rangapuram, and Slawski (2018)), and let Π_V denote the projection on a subspace V of $\mathbb{H}^m$.

Condition 2. We say that the $\boldsymbol{\Delta}$-*RSC* condition is satisfied for rank r and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

constant $\kappa > 0$ if $\forall \Phi \in \mathcal{K}^{\mathbf{\Delta}}(r)$, it holds that $\|\mathcal{X}(\Phi)\|_2^2/n \geq \kappa \|\Phi\|_2^2$, where

$$\mathcal{K}^{\mathbf{\Delta}}(r) = \{\Phi \in \mathbb{H}^m : \exists B \in \boldsymbol{B}_0^m(r) \text{ s.t.}$$
$$\operatorname{tr}(\Pi_{\mathbb{T}(B)^\perp}(\Phi)) = -\operatorname{tr}(\Pi_{\mathbb{T}(B)}\Phi) \text{ and } \Pi_{\mathbb{T}(B)^\perp}(\Phi) \succeq 0\}.$$

The $\mathbf{\Delta}$-RSC condition is weaker than the corresponding condition employed in Negahban and Wainwright (2011), which, in turn, is weaker than the matrix RIP condition (Recht et al., 2010). The next statement parallels Proposition 2, asserting that the constraint $B \in \mathbf{\Delta}^m$ alone is strong enough to take advantage of low-rankedness.

Proposition 6. *Suppose that the* $\mathbf{\Delta}$-RSC *condition is satisfied for rank* r *and* $\kappa > 0$. *Set* $\lambda_* = 2\|\mathcal{X}^\star(\varepsilon)\|_\infty/n$, *where* $\mathcal{X}^\star : \mathbb{R}^n \to \mathbb{H}^m$ *is the adjoint of* $\mathcal{X}$. *We then have*

$$\|\widehat{B} - B^*\|_2^2 \leq \frac{4s\lambda_*^2}{\kappa^2}, \quad \|\widetilde{B}_\lambda - B^*\|_2^2 \leq \frac{4s(\lambda + \lambda_*)^2}{\kappa^2},$$
$$\|\widehat{B} - B^*\|_1 \leq \frac{4s\lambda_*}{\kappa}, \quad \|\widetilde{B}_\lambda - B^*\|_1 \leq \frac{4s(\lambda + \lambda_*)}{\kappa}.$$

Obtaining solutions of low rank. Although $\widehat{B}$ may have a low estimation error, its rank can far exceed that of B^*, even though the extra nonzero eigenvalues of $\widehat{B}$ tend to be small. The simplest approach to obtaining solutions of low rank is to threshold the spectrum of $\widehat{B} = \widehat{U}\widehat{\Phi}\widehat{U}^\top$ (the r.h.s. representing the usual spectral decomposition); that is, $\widehat{B}_\tau = \widehat{U}\widehat{\Phi}_\tau\widehat{U}^\top$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

where $\widehat{\Phi}_\tau = \mathrm{diag}(\{\mathbf{I}(\widehat{\phi}_j \geq \tau)\,\widehat{\phi}_j\}_{j=1}^m)$ for a threshold $\tau > 0$. Similarly, we may use the following analog to weighted ℓ_1-regularization:

$$\begin{aligned} &\widehat{B}_w = \widehat{U}\mathrm{diag}(\{\widehat{\phi}_{w,j}\}_{j=1}^m)\widehat{U}^\top \\ &\text{with} \quad \widehat{\phi}_w \in \operatorname*{argmin}_{\phi \in \Delta^p} \frac{1}{n}\|\mathbf{Y} - \mathcal{X}(\widehat{U}\mathrm{diag}(\{\phi_j\}_{j=1}^m)\widehat{U}^\top)\|_2^2 + \lambda \langle w, \phi \rangle , \end{aligned} \tag{5.3}$$

for non-negative weights $\{w_j\}_{j=1}^m$, as in the vector case. Note that the matrix of eigenvectors $\widehat{U}$ is kept fixed at the second stage; the optimization is only over the eigenvalues. Alternatively, we can consider optimization over $\mathbf{\Delta}^m$, with regularizer $B \mapsto \|B\|_w = \sum_{j=1}^m w_j \phi_j(B)$, for eigenvalues $\phi_1(B) \geq \ldots \geq \phi_m(B) \geq 0$ of B, in decreasing order. However, from the point of view of optimization $\|\cdot\|_w$ poses difficulties, including possible nonconvexity (depending on w).

Regularization with the negative ℓ_2-norm. An additional positive aspect about the regularization scheme proposed in §4 is that it allows a straightforward extension to the matrix case, including the algorithm used for optimization (Algorithm 1). In contrast, for regularization with the inverse ℓ_∞-norm, which can be reduced to p convex optimization problems in the vector case, no such reduction seems to be possible in the matrix case. The analogs of (4.3) and (4.4) are given by

$$\widehat{B}_\lambda^{\ell_2} \in \operatorname*{argmin}_{B \in \mathbf{\Delta}^m} R_n(\beta) - \lambda\|B\|_2^2, \tag{5.4}$$

$$\widetilde{B}_\lambda^{\ell_2} \in \operatorname*{argmin}_{B \in \boldsymbol{D}(\lambda)} -\|B\|_2^2. \tag{5.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Algorithm 1 can be employed for optimization mutatis mutandis. In the vector case and for squared loss, formulations (4.3) and (4.4) are comparable in terms of their computational requirements: each minimization problem inside the repeat-loop becomes a quadratic (resp., a linear) program, with a comparable number of variables/constraints. In the matrix case, however, (5.4) appears to be preferable because the subproblems are better suited to the proximal gradient method. In contrast, the constraint set in (5.5) requires a more sophisticated approach.

Denoising. Negative ℓ_2-regularization, together with the constraint set $\mathbf{\Delta}^m$ enforces a solution of low rank, as exemplified here in the special case of denoising of a real-valued matrix (i.e., $B^* \in \mathbb{H}^m \cap \mathbb{R}^{m\times m}$) contaminated by Gaussian noise. Specifically, the sampling operator $\mathcal{X}(\cdot) = (\langle X_i, \cdot\rangle)_{i=1}^n$, for $n = m(m+1)/2$, is equal to the symmetric vectorization operator; that is

$$X_1 = e_1e_1^\top,\ \ X_2 = \frac{e_1e_2^\top + e_2e_1^\top}{\sqrt{2}}, \ldots, X_m = \frac{e_1e_m^\top + e_me_1^\top}{\sqrt{2}},\ \ X_{m+1} = e_2e_2^\top, \ldots,$$
$$X_{2m-1} = \frac{e_2e_m^\top + e_me_2^\top}{\sqrt{2}}, \ldots, X_{m(m+1)/2} = \frac{e_{m-1}e_m^\top + e_me_{m-1}^\top}{\sqrt{2}}. \tag{5.6}$$

The following proposition uses a result in random matrix theory of Peng (2012).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Proposition 7. *Let $B^* \in \mathbf{\Delta}_0^m(r) \cap \mathbb{R}^{m\times m}$ with eigenvalues $\phi_1^* \geq \ldots \geq \phi_r^* > 0$ and $\phi_{r+1}^* = \ldots = \phi_m^* = 0$, let $\mathcal{X}$ be defined according to (5.6), and let $\varepsilon \sim N(0, \sigma^2 I_m/m)$, $\mathbf{Y} = \mathcal{X}(B^*) + \varepsilon$. Consider the optimization problem*

$$\min_{B\in\mathbf{\Delta}^m} \frac{1}{n}\|\mathbf{Y} - \mathcal{X}(B)\|_2^2 - \lambda\|B\|_2^2,$$

with minimizer $\widehat{B}_\lambda^{\ell_2}$, and define $\Upsilon = B^* + \mathcal{X}^\star(\varepsilon)$. Then, for all $\lambda \geq 1/n$, we have $\widehat{B}_\lambda^{\ell_2} = u_1 u_1^\top$, where u_1 is the eigenvector of Υ corresponding to its largest eigenvalue. For all $0 \leq \lambda < 1/n$, we have $\widehat{B}_\lambda^{\ell_2} = \operatorname{argmin}_{B\in\mathbf{\Delta}^m}\|\Upsilon/\gamma - B\|_2^2$, where $\gamma = 1 - n\lambda$. Moreover, there exist constants $c_0, c, C > 0$ so that if $r < c_0 m$, $\lambda \geq 6\sigma r/n$, and $\phi_r^* \geq 5\sigma + n\lambda/r$, we have $\|\widehat{B}_\lambda^{\ell_2}\|_0 = r$, with probability at least $1 - C\exp(-cm)$.

In particular, for $\lambda = (1+\delta)6\sigma r/n$ for some $\delta > 0$, the required lower bound on ϕ_r^* becomes $11(1+\delta)\sigma$, which is proportional to the noise level of the problem, as follows from the proof of the proposition.

6. Conclusion

Simplex constraints are beneficial in high-dimensional estimations, empirically achieving lower estimation errors than when using ℓ_1-norm regularization in place of the constraint. In order to enhance the sparsity of the solution, simple two-stage methods (i.e., thresholding and weighted ℓ_1-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

regularization) are effective. A more principled way to incorporate sparsity is to use a suitable regularizer. We have pointed out that under simplex constraints, sparsity cannot be promoted by convex regularizers. We have therefore considered nonconvex alternatives, among which regularization using the negative ℓ_2-norm turns out to be a natural approach, lending itself to a straightforward computational strategy. As an attractive feature, there is a direct and practical generalization to the matrix counterpart, in contrast to the two-stage methods.

Supplementary Material

The Supplementary Material (Li, Rangapuram, and Slawski, 2018) contains proofs of all statements, as well as extensive numerical results and simulations illustrating the central aspects of this work.

Acknowledgments

The work was partially supported by NSF-DMS-1444124, NSF-III-1360971, and AFOSR-FA9550-13-1-0137. The work of Syama Sundar Rangapuram was also partially supported by the ERC starting grant NOLEPRO. The authors would like to thank Anastasios Kyrillidis for clarifications regarding

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

step size selection for the iterative hard thresholding method discussed in the present work.

We thank the three reviewers for their careful reading and thoughtful comments, which led to several improvements to the paper.

References

Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*, 1705–1732.

Blumensath, T. and M. Davies (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis 27*, 265–274.

Bunea, F., A. Tsybakov, M. Wegkamp, and A. Barbu (2010). SPADES and mixture models. *The Annals of Statistics 38*, 2525–2558.

Cai, T., D. Kim, Y. Wang, M. Yuan, and H. Zhou (2016). Optimal large-scale quantum state tomography with Pauli measurements. *The Annals of Statistics 44*, 681–712.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

Candes, E., M. Wakin, and S. Boyd (2007). Enhancing sparsity by reweighted ℓ_1-minimization. *Journal of Fourier Analysis and Applications 14*, 877–905.

Dicker, L. (2014). Variance estimation in high-dimensional linear models. *Biometrika 101*,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

269–284.

Fan, J. (2013). Features of big data and sparsest solution in high confidence set. Technical report, Department of Operations Research and Financial Engineering, Princeton University.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society Series B 74*, 37–65.

Foster, D. and E. George (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics 22*, 1947–1975.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli 6*, 971–988.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory 57*, 1548–1566.

Gross, D., Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert (2010). Quantum State Tomography via Compressed Sensing. *Physical Review Letters 105*, 150401–15404.

James, G., C. Paulson, and P. Rusmevichientong (2015). Penalized and Constrained Regression. Manuscript, University of Southern California.

James, G. and P. Radchenko (2009). A Generalized Dantzig Selector with Shrinkage tuning.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Biometrika 96, 323–337.

Johnstone, I. (2013). Gaussian estimation: Sequence and wavelet models. `http://statweb.stanford.edu/~imj/GE06-11-13.pdf`.

Jojic, V., S. Saria, and D. Koller (2011). Convex envelopes of complexity controlling penalties: the case against premature envelopement. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 15 of *JMLR W&CP*, pp. 399–406.

Keshava, N. (2003). A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal 14*, 55–78.

Kim, Y., S. Kwon, and H. Choi (2012). Consistent Model Selection Criteria on High Dimensions. *Journal of Machine Learning Research 13*, 1037–1057.

Koltchinskii, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics 39*, 2936–2973.

Kyrillidis, A., S. Becker, V. Cevher, and C.Koch (2013). Sparse projections onto the simplex. In *International Conference on Machine Learning (ICML)*, Volume 28 of *JMLR W&CP*, pp. 235–243.

Lange, K., D. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics 9*, 1–20.

Larsson, M. and J. Ugander (2011). A concave regularization technique for sparse mixture models. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 24, pp.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

1890–1898.

Lecue, G. and S. Mendelson (2017). Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society 19*, 881–904.

Li, P., S. Rangapuram, and M. Slawski (2018). Supplementary material of "methods for sparse and low-rank recovery under simplex constraint".

Lin, W., P. Shi, R. Feng, and H. Li (2014). Variable selection in regression with compositional covariates. *Biometrika 101*, 785–797.

Lounici, K. (2008). High-dimensional stochastic optimization with the generalized Dantzig estimator. arXiv:0811.2281.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance 7*, 77–91.

Mazumder, R., J. Friedman, and T. Hastie (2011). *SparseNet*: Coordinate Descent with Non-Convex Penalties. *Journal of the American Statistical Association 106*, 1125–1138.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *The Electronic Journal of Statistics 7*, 1607–1631.

Negahban, S., P. Ravikumar, M. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*, 538–557.

Negahban, S. and M. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics 39*, 1069–1097.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Nemirovski, A. (2000). *Ecole d'Ete de Probabilites de Saint-Flour XXVIII*, Chapter 'Topics in non-parametric statistics'. Springer.

Nielsen, M. and I. Chuang (2000). *Quantum Computation and Quantum Information*. Cambridge University Press.

Pardalos, P. and S. Vavasis (1991). Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global Optimization 1*, 15–22.

Paris, M. and J. Rehacek (Eds.) (2004). *Quantum State Estimation*. Springer.

Peng, M. (2012). Eigenvalues of Deformed Random Matrices. arXiv:1205.0572.

Pham Dinh, T. and H. Le Thi (1997). Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica 22*, 289–355.

Pilanci, M., L. E. Ghaoui, and V. Chandrasekaran (2012). Recovery of Sparse Probability Measures via Convex Programming. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 25, pp. 2420–2428.

Raskutti, G., M. Wainwright, and B. Yu (2010). Restricted nullspace and eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research 11*, 2241–2259.

Recht, B., M. Fazel, and P. Parillo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review 52*, 471–501.

Reid, S., R. Tibshirani, and J. Friedman (2016). A study of error variance estimation in lasso regression. *Statistica Sinica 26*, 35–67.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Rudelson, M. and S. Zhou (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory 59*, 3434–3447.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*, 461–464.

Shashanka, M., B. Raj, and P. Smaragdis (2008). Sparse Overcomplete Latent Variable Decomposition of Counts Data. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 20, pp. 1313–1320.

Shen, J. and P. Li (2018). A tight bound of hard thresholding. *Journal of Machine Learning Research 18*(208), 1–42.

Shi, P., A. Zhang, and H. Li (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics 10*, 1019–1040.

Slawski, M. and M. Hein (2013). Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. *The Electronic Journal of Statistics 7*, 3004–3056.

Sun, T. and C. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*, 879–898.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*, 614–645.

Vila, J. and P. Schniter (2014). An Empirical-Bayes Approach to Recovering Linearly Constrained Non-Negative Sparse Signals. *IEEE Transactions on Signal Processing 62*, 4689–4703.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0220

Wang, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *The Annals of Statistics 41*, 2462–2504.

Ye, F. and C. Zhang (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research 11*, 3519–3540.

Yuille, A. and A. Rangarajan (2003). The concave-convex procedure. *Neural Computation 15*, 915–936.

Zhang, C. and T. Zhang (2013). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science 27*, 576–593.

Zhao, P. and B. Yu (2006). On model selection consistency of the lasso. *Journal of Machine Learning Research 7*, 2541–2567.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Baidu Research, Bellevue WA, 98004, United States

E-mail: pingli98@gmail.com

Amazon Development Center, Germany.

E-mail: rangapur@amazon.de

George Mason University, United States.

E-mail: mslawsk3@gmu.edu