On signal detection and confidence sets for low rank inference problems

Alexandra Carpentier[*]

Institut für Mathematik
Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
Germany
e-mail: carpentier@maths.uni-potsdam.de

and

Richard Nickl

Statistical Laboratory
Center for Mathematical Sciences
University of Cambridge
Wilberforce Road
CB3 0WB Cambridge
United Kingdom
e-mail: r.nickl@statslab.cam.ac.uk

Abstract: We consider the signal detection problem in the Gaussian design trace regression model with low rank alternative hypotheses. We derive the precise (Ingster-type) detection boundary for the Frobenius and the nuclear norm. We then apply these results to show that honest confidence sets for the unknown matrix parameter that adapt to all low rank sub-models in nuclear norm do not exist. This shows that recently obtained positive results in [5] for confidence sets in low rank recovery problems are essentially optimal.

1. Introduction

Consider the Gaussian design trace regression model

$$Y_i = \text{tr}(X^i\theta) + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\epsilon \sim N(0, I_n)$ is an i.i.d. vector of Gaussian noise. Here the matrices X^i are $d \times d$ square matrices with i.i.d. entries $X^i_{mk} \sim N(0, 1)$, and θ is the unknown

---

$d \times d$ matrix we want to make inference on. We are interested in the case where the model dimension d^2 is possibly large compared to sample size n, but where θ has low rank k, in which case we write $\theta \in R(k), 1 \le k \le d$. This setting serves as a prototype for various matrix inference problems such as those occurring in compressed sensing [4] or in quantum tomography [7]. We consider here a high-dimensional regime where $\min(d, n) \to \infty$, reflecting contemporary statistical challenges.

The first problem we study in this paper is the *signal detection problem* with low-rank alternatives: We want to test the hypothesis

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta \ne 0, \ \theta \in R(k), \ \|\theta\| \ge \rho,$$

where $\| \cdot \|$ equals either the Frobenius norm $\| \cdot \|_F$ or the nuclear norm $\| \cdot \|_*$ (defined in detail below), and where ρ should be the minimal 'signal strength' condition for the above hypothesis testing problem to have a consistent solution (in the sense of Ingster, see [10]). We will show that the minimax optimal detection boundary in Frobenius norm is of the form

$$\rho \approx \min\left(\sqrt{\frac{d}{n}}, n^{-1/4} \right)$$

whereas in nuclear norm it is

$$\rho \approx \min\left(\sqrt{\frac{kd}{n}}, \sqrt{\frac{k}{n^{1/2}}} \right).$$

A remarkable feature is that for the Frobenius norm the detection rate *does not depend at all* on the complexity of the alternative hypothesis (the rank k), whereas for the nuclear norm it does. The phase transition between the two regimes in these rates depends precisely on whether the sample size n exceeds the dimension d^2 of the maximal parameter space $R(d)$ or not. The upper bounds in our proofs are related to the papers [9, 1] about the detection boundary in the sparse regression setting, and our main contribution consists in deriving the matching lower bounds for low rank alternatives.

Our interest in the detection boundary is triggered by the second problem we investigate here: the question of existence and non-existence of adaptive confidence sets for low rank parameters. It follows from general decision-theoretic principles (see Chapter 8.3 in [6] and also [8, 2]) that the answer to this question is closely related to a 'composite version' of the detection problem (see (15) below). This approach was employed in [14] to prove that adaptive and honest confidence sets for the parameter θ do not exist in *sparse* regression models if an ℓ_2-risk performance beyond $O(n^{-1/4})$ is desired. In contrast in the recent paper [5] it was shown that if sparsity constraints are replaced by low rank conditions, then adaptive and fully honest confidence sets exist over the entire parameter space $R(d)$. Adaptation means here that the expected Frobenius norm diameter of the confidence set reflects the minimax risk over arbitrary low rank sub-models $R(k), 1 \le k \le d$. The fact that the detection rates obtained here

in Frobenius norm are independent of the rank constraint $\theta \in R(k)$ provides another heuristic explanation of the result in [5].

Moreover [5] constructed another confidence set whose diameter adapts to low rank sub-models in the stronger nuclear norm distance, and that is honest for all θ's that are *non-negative definite and have trace equal to one*, that is, whenever θ is the density matrix of a *quantum state*. Such a constraint on θ is natural in a quantum physics context considered in [5], but not in general. The question arises whether it is essentially necessary or not. In the present paper we show that indeed the existence results of [5] are specific to the geometry induced by the Frobenius norm or to the quantum state constraint, and that nuclear-norm adaptive and honest confidence sets over general low rank parameter spaces *do not exist* in the model (1). For example, our results imply that if one requires coverage of a confidence set over all of $R(d)$ then the worst case nuclear norm diameter for rank-one parameters can be off the minimax estimation rate over $R(1)$ by as much as $\sqrt{d}$. Our results thus further illustrate the subtleties involved in the theory of confidence sets for high-dimensional parameters, and that the positive results in [5] are of a rather specific nature.

Our proofs are given in the simplest model where both the design and the noise are Gaussian, and the matrices involved are of square type. As usual, our results extend without major difficulty to sub-Gaussian design and noise, to certain correlated random designs, and also to non-square matrices, at the expense of slightly more technical proofs. Generalisations of our results to the matrix completion problem are currently under investigation.

2. Main results

2.1. Notation

We write $\mathbb{M}_d$ for the set of $d \times d$ matrices with real elements. If $\mathcal{X} : \mathbb{M}_d \to \mathbb{R}^n$ denotes the 'sampling operator'

$$\theta \mapsto \mathcal{X}\theta = \big(\mathrm{tr}(X^1\theta), \ldots, \mathrm{tr}(X^n\theta)\big)^T,$$

then the model (1) can be written as

$$Y = \mathcal{X}\theta + \epsilon,$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. We write E^X for the expectation over the distribution of $\mathcal{X}$ only, and E_θ for the expectation conditional on $\mathcal{X}$. The full expectation is denoted by $\mathbb{E}_\theta = E^X E_\theta$. The corresponding probability laws are denoted by $P^X, P_\theta, \mathbb{P}_\theta$ and we employ the usual $o/O/o_P/O_P$-notation with $\min(n, d) \to \infty$.

We denote the standard norm on Euclidean space by $\|\cdot\|_2$, and the associated inner product by $\langle \cdot, \cdot \rangle_2$. Let $\|.\|_F$ be the Frobenius norm over $\mathbb{M}_d$, i.e.

$$\|M\|_F = \sqrt{\mathrm{tr}(M^T M)} = \sqrt{\sum_{j \leq d} \lambda_j^2},$$

where λ_i^2 are the eigenvalues of $M^T M$. The associated inner product is

$$\langle U, V \rangle_F = \operatorname{tr}(U^T V).$$

We also define the nuclear norm of M as

$$\|M\|_* = \sum_{j \leq d} |\lambda_j|.$$

These two norms are in fact defined also for matrices that are not of square type. Finally we recall that for any matrix $M \in R(k)$, we have

$$\|M\|_F \leq \|M\|_* \leq \sqrt{k}\|M\|_F.$$

2.2. Signal detection for low rank alternatives

We consider first the following hypothesis testing problem, also known as the signal detection problem:

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta \in R(k), \ \|\theta\| \geq \rho. \tag{2}$$

Here the alternative space is restricted to a 'low rank' hypothesis $\theta \in R(k)$ for some $1 \leq k \leq d$. Moreover, for a separation constant $\rho > 0$, the detection boundary is described by a 'signal strength' condition measured in terms of the size $\|\theta\| \geq \rho$ of the Frobenius-, or of the nuclear norm of θ. In the high-dimensional regime where $\min(n, d) \to \infty$, we want to find the minimal sequence $\rho \equiv \rho_{n,d}$ such that for any $\alpha > 0$ a level α-test $\Psi = \Psi(Y, \mathcal{X}, \alpha)$ exists:

$$\left[\mathbb{E}_0[\Psi] + \sup_{\theta \in H_1} \mathbb{E}_\theta[1 - \Psi] \right] = \mathbb{P}_0(\text{reject } H_0) + \sup_{\theta \in H_1} \mathbb{P}_\theta(\text{accept } H_0) \leq \alpha. \tag{3}$$

Recall that a test is simply a random indicator function $\psi = 1_A$ where the rejection event A depends only on $Y, \mathcal{X}, \alpha$, and we require the sum of the type-one and the type-two error of the test to be controlled at any fixed level $\alpha > 0$.

Theorem 1. *Consider the testing problem (2) with norm $\|\cdot\|$. Define*

$$r_{n,d} = \begin{cases} \min(\sqrt{d/n}, n^{-1/4}) & \text{if } \|\cdot\| = \|\cdot\|_F \\ \min(\sqrt{kd/n}, \sqrt{k}/n^{1/4}) & \text{if } \|\cdot\| = \|\cdot\|_*. \end{cases}$$

1) *Suppose $\rho \geq D r_{n,d}$. Then for every $\alpha > 0$ there exists a test $\Psi = \Psi(Y, \mathcal{X}, \alpha)$ and finite constants $D = D_\alpha > 0, n_\alpha \in \mathbb{N}$ such that (3) holds for every $n \geq n_\alpha$.*
2) *Conversely, suppose $\rho = o(r_{n,d})$ and $k = o(d)$ as $\min(n, d) \to \infty$. Then no test satisfying (3) for every $\alpha > 0$ exists. In fact*

$$\liminf_{n,d} \inf_\Psi \left[\mathbb{E}_0[\Psi] + \sup_{\theta \in H_1} \mathbb{E}_\theta[1 - \Psi] \right] \geq 1 \tag{4}$$

where the infimum extends over all test functions $\Psi = \Psi(Y, \mathcal{X})$.

The tests Ψ constructed in the proof are given in (9) below and straightforward to implement. Note also that the $\|\cdot\|_*$-separated alternatives are a subset of the $\|\cdot\|_F$-separated alternatives (see (10) below), and our results imply that an optimal test for the case $\|\cdot\| = \|\cdot\|_F$ is essentially optimal also for $\|\cdot\|_*$.

2.3. Confidence sets for low rank recovery

Low rank recovery algorithms are well-studied in compressed sensing and high-dimensional statistics, see e.g., [4, 7, 11, 12, 13, 3] and the references therein. In the setting of model (1) they provide minimax optimal estimators $\tilde{\theta}$ of $\theta \in R(k)$ with (high probability) performance guarantees

$$\|\tilde{\theta} - \theta\|_F^2 \lesssim \frac{kd}{n}, \qquad \|\tilde{\theta} - \theta\|_* \lesssim k\sqrt{\frac{d}{n}}. \tag{5}$$

The question we study here is whether associated uncertainty quantification methodology exists, that is, whether we can find confidence sets $C_n \subset \mathbb{M}_d$ such that

$$\inf_{\theta \in \mathbb{M}_d} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \tag{6}$$

at least for $\min(n, d)$ large enough, and such that the diameter $|C_n|$ of C_n reflects the accuracy of adaptive estimation in the sense that $|C_n|$ shrinks, with high probability, at the optimal rates from (5) whenever $\theta \in R(k)$. We insist here on an *adaptive* confidence set that does not require knowledge of the unknown rank k of θ.

A first result that is proved in the paper [5] is that such adaptive confidence sets do exist in the model (1) if the diameter is measured in Frobenius distance. The construction of this set is straightforward, see [5] for details.

Theorem 2 (Theorem 2 in [5]). *For every $\alpha > 0$ there exists a confidence set $C_n = C_n(Y, \mathcal{X}, \alpha)$ such that for all $n \in \mathbb{N}$, (6) holds, and such that uniformly in $\theta \in R(k_0)$ for any $1 \leq k_0 \leq k$, with high $\mathbb{P}_\theta$-probability the Frobenius-norm diameter $|C_n|_F$ of C_n satisfies*

$$|C_n|_F \lesssim \sqrt{k_0 \frac{d}{n}}.$$

A second result that is proved in the paper [5] is that an (asymptotic) adaptive confidence set exists also in nuclear norm provided that the "quantum state constraint" is satisfied, namely, provided it is known a priori that θ is non-negative definite and has nuclear norm one, and provided the coverage requirement in (6) is relaxed to hold only over a maximal model $R(k)$ in which asymptotically consistent estimation of θ is possible (i.e., $k\sqrt{d/n} = o(1)$). Define

$$R^+(k) = R(k) \cap \{\theta \text{ is non-negative definite, } \operatorname{tr}(\theta) = 1\},$$

the set of quantum state density matrices of rank at most k.

Theorem 3 (Theorem 4 in [5])**.** *Assume $k\sqrt{d/n} = o(1)$ for some $1 \le k \le d$, and let $\alpha > 0$ be given. Then there exists a confidence set $C_n = C_n(Y, \mathcal{X}, \alpha)$ such that*

$$\liminf_{\min(n,d) \to \infty} \inf_{\theta \in R^+(k)} \mathbb{P}_\theta(\theta \in C_n) \ge 1 - \alpha,$$

and such that uniformly in $\theta \in R^+(k_0)$ for any $1 \le k_0 \le k$, with high $\mathbb{P}_\theta$-probability the nuclear norm diameter $|C_n|_*$ of C_n satisfies

$$|C_n|_* \lesssim k_0 \sqrt{\frac{d}{n}}.$$

In fact it is not difficult to generalise the above theorem to the case where the condition $tr(\theta) = 1$ is relaxed to $\|\theta\|_* \le 1$.

The next theorem, which is the main result of this subsection, implies that no analogue of Theorem 2 can hold true if the Frobenius norm there is replaced by the nuclear norm, and it also shows that Theorem 3 cannot hold true if $R^+(k)$ is replaced by $R(k)$, that is, if the 'quantum state constraint' is relaxed. More precisely, we show that if a confidence set C_n is required to have coverage over the maximal model $R(k_1)$, then the worst case expected nuclear norm diameter of C_n over arbitrary sub-models $R(k_0), k_0 = o(k_1)$, depends on the maximal model dimension k_1 and does not improve as $k_0 \downarrow 1$. The proof of Theorem 4 is based on Part 2) of Theorem 1 and lower bound techniques for adaptive confidence sets from [8, 2].

Theorem 4. *Let $k_1 \to \infty$ such that $k_1 = o(d)$ as $\min(n, d) \to \infty$. Suppose that for any $0 < \alpha < 1/3$ the confidence set $C_n = C_n(Y, \mathcal{X}, \alpha)$ is asymptotically honest over the maximal model $R(k_1)$, that is, it satisfies*

$$\liminf_{\min(n,d) \to \infty} \inf_{\theta \in R(k_1)} \mathbb{P}_\theta(\theta \in C_n) \ge 1 - \alpha. \tag{7}$$

Then for every $k_0 = o(k_1)$ and some constant $c > 0$ depending on α, we have

$$\sup_{\theta \in R(k_0)} \mathbb{E}_\theta |C_n|_* \ge c\sqrt{\frac{k_1 d}{n}} \tag{8}$$

for every $\min(n, d)$ large enough. In particular no confidence set exists that is honest over all of $\mathbb{M}_d$ and that adapts in nuclear norm to any model $R(k_0), k_0 = o(\sqrt{d})$.

For notational simplicity we have lower bounded the *expected diameter* $|C_n|_*$ in (8), but the proof actually contains a stronger 'in probability version' of this lower bound.

Remark 1. *A few remarks on Theorem 4 are in order:*

  i) *In the least favourable case where one wants coverage over the entire $R(d) = \mathbb{M}_d$ while still adapting to rank-one matrices (i.e., $k_0 = 1$), the performance of any honest confidence set is off the minimax optimal adaptive estimation rate $\sqrt{d/n}$ over $R(1)$ by a diverging factor that can be as close to $\sqrt{d}$ as desired.*

ii) Even if one restricts coverage to hold only for 'consistently estimable models' $R(k_1)$ with $k_1\sqrt{d/n} \to 0$ (as in Theorem 3), the diameter $|C_n|_*$ can be off the minimax rate of estimation over $R(1)$ by a factor of $\sqrt{k_1}$.

iii) We also note that the above result does not disprove the existence of adaptive confidence sets for sub-models $R(k_0)$ of 'moderate rank' where $k_0 \geq \sqrt{d}$. While more of technical interest – note that this rules out $n < d^2$ for consistent recovery to be possible – this regime currently remains open (it is related to the apparently hard problem of finding optimal separation rates in the composite testing problem (15) below).

3. Proofs

3.1. Proof of Theorem 1, upper bounds

When $n < d^2$ then define

$$\hat{r}_n = \frac{1}{n}\|Y\|_2^2 - 1, \quad \tau_n = n^{-1/2}$$

but when $n \geq d^2$ set

$$\hat{r}_n = \frac{2}{n(n-1)} \sum_{i<j} \sum_{1 \leq m \leq d, 1 \leq k \leq d} Y_i X_{mk}^i Y_j X_{mk}^j, \quad \tau_n = d/n.$$

The test statistic is

$$\Psi_n = 1\left\{\hat{r}_n \geq z_\alpha \tau_n\right\} \tag{9}$$

where z_α are quantile constants chosen below.

These tests work for Frobenius norm separation, by effectively the same proofs as in [9], using that we can embed the matrix regression model into a vector regression model with $p = d^2$ parameters, and since the separation rates only depend on the model dimension (and not on low rank or sparsity degrees). However, to provide intuition, we give some details, first for the case $n < d^2$: Under H_0 we have $Y = \epsilon$ and so

$$\mathbb{E}_0 \Psi_n = \Pr\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i^2 - E\varepsilon_i^2) > z_\alpha\right) \leq \alpha/2$$

for every $n \in \mathbb{N}$ and z_α large enough (using either Chebyshev's inequality and $E\varepsilon_i^4 = 3$, or Theorem 4.1.9 in [6] for a more precise non-asymptotic bound). Now for the alternatives $\theta \in H_1$ we use the basic concentration result Lemma 1a) in [5] which implies that for any fixed θ the event

$$\mathcal{E} = \left\{ \left|(1/n)\|\mathcal{X}\theta\|_2^2 - \|\theta\|_F^2\right| \leq \|\theta\|_F^2/2\right\}$$

has P^X-probability at least $1 - 2\exp(-n/24)$, and so, for $n \geq n_\alpha$ such that $2\exp(-n/24) < \alpha/6$,

$$\mathbb{E}_\theta(1 - \Psi_n) = \mathbb{P}_\theta\left(\hat{r}_n < z_\alpha \tau_n\right)$$

$$= \Pr\left(\frac{1}{n}\|\mathcal{X}\theta + \epsilon\|_2^2 - 1 < \frac{z_\alpha}{\sqrt{n}}\right)$$

$$= \Pr\left(\frac{1}{n}\|\mathcal{X}\theta\|_2^2 - \frac{z_\alpha}{\sqrt{n}} < -\frac{2}{n}\epsilon^T\mathcal{X}\theta - \frac{1}{n}\sum_{i=1}^n(\varepsilon_i^2 - 1)\right)$$

$$\leq \Pr\left(\frac{\|\theta\|_F^2}{2} - \frac{z_\alpha}{\sqrt{n}} < -\frac{2}{n}\epsilon^T\mathcal{X}\theta - \frac{1}{n}\sum_{i=1}^n(\varepsilon_i^2 - 1), \mathcal{E}\right)$$

$$+ 2\exp(-n/24)$$

$$\leq \Pr\left(\left|\frac{2}{n}\epsilon^T\mathcal{X}\theta\right| > \|\theta\|_F^2/8, \mathcal{E}\right) + \Pr\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\varepsilon_i^2 - E\varepsilon_i^2) > z_{\alpha/3}\right)$$

$$+ \alpha/6,$$

since, by the hypothesis on ρ, we have for D large enough that

$$\frac{\|\theta\|_F^2}{2} - \frac{z_\alpha}{\sqrt{n}} \geq \frac{\|\theta\|_F^2}{4} \geq \frac{2z_{\alpha/3}}{n^{1/2}}.$$

The last probability is bounded by $\alpha/6$ as under H_0 and the last but one probability is also bounded by $\alpha/6$ by a direct (conditional on $\mathcal{X}$) Gaussian tail inequality (restricting to the event $\mathcal{E}$: just as in term II of the proof of Theorem [5] with $\tilde{\theta} = 0$ there), so that in total we have bounded the testing errors in (3) by $\alpha/2 + (3/6)\alpha = \alpha$, as desired. The case $n \geq d^2$ follows from similar but slightly more technical arguments, adapting the arguments from proof of Theorem 3 in [5], or arguing directly as in Theorem 4.3 in [9] with $p = d^2$.

The test (9) also works for nuclear-norm separation since

$$H_1^* = \theta \in R(k) : \|\theta\|_* \geq c\sqrt{k}\rho$$

is a subset of

$$H_1^F = \theta \in R(k) : \|\theta\|_F \geq c\rho$$

in view of the inequality

$$\|\theta\|_F \geq (1/\sqrt{k})\|\theta\|_* \quad \forall\theta \in R(k), \tag{10}$$

so that

$$\mathbb{E}_0\Psi_n + \sup_{\theta \in H_1^*} \mathbb{E}_\theta(1 - \Psi_n) \leq \mathbb{E}_0\Psi_n + \sup_{\theta \in H_1^F} \mathbb{E}_\theta(1 - \Psi_n) \leq \alpha.$$

We now turn to the more difficult lower bounds.

3.2. Proof of Theorem 1, lower bounds

Let Ψ be any test – any measurable function of $Y, \mathcal{X}$ that takes values in $\{0, 1\}$. Assume $\rho = o(r_{n,d})$ as $\min(n, d) \to \infty$ and let $H_1 = H_1(\rho)$ be the corresponding alternative hypothesis.

Step I: Reduction to averaged likelihood ratios: Let $\pi = \pi_{n,d}$ be a sequence of finitely supported probability distributions on $\mathbb{M}_d$ such that $\pi_{n,d}(H_1) \to 1$, and denote by $\pi|H_1$ that measure restricted to H_1 and re-normalised to unit mass. Define

$$Z = \mathbb{E}_{\theta \sim \pi} \prod_{i \leq n} \frac{dP_i^{(\theta)}}{dP_i^{(0)}} \equiv \int \prod_{i \leq n} \frac{dP_i^{(\theta)}}{dP_i^{(0)}} d\pi(\theta),$$

where $dP_i^{(\theta)}$ is the distribution of $Y_i|\mathcal{X}$ when the parameter generating the data is θ, and $dP_i^{(0)}$ is the distribution of $Y_i|\mathcal{X}$ when the parameter generating the data is 0. Then, by a standard testing lower bound (e.g., (6.23) in [6]), for any $\eta > 0$,

$$\mathbb{E}_0 \Psi + \sup_{\theta \in H_1} \mathbb{E}_\theta (1 - \Psi) \geq \mathbb{E}_0 \Psi + \mathbb{E}_{\theta \sim \pi|H_1} \mathbb{E}_\theta (1 - \Psi)$$

$$\geq \mathbb{E}_0 \Psi + \mathbb{E}_{\theta \sim \pi} \mathbb{E}_\theta (1 - \Psi) - o(1)$$

$$= E^X \left[E_0 \Psi + \mathbb{E}_{\theta \sim \pi} E_\theta (1 - \Psi) \right] - o(1)$$

$$\geq (1 - \eta) \left[1 - \left[\frac{\sqrt{\mathbb{E}_0(Z-1)^2}}{\eta} \right] \right] - o(1).$$

Now since

$$\mathbb{E}_0[Z-1]^2 = \mathbb{E}_0[Z^2] - 1,$$

if we show that $\mathbb{E}_0[Z^2] \leq 1 + o(1)$ as $\min(n,d) \to \infty$ for a suitable choice of π, then the lower bound (4) will follow by letting $\eta \to 0$. Recall the notation $\mathbb{E}_\theta = E^X E_\theta$.

Step II: Computation of $E_0[Z^2]$: The (Y_i) are independent with distribution $\mathcal{N}((\mathcal{X}\theta)_i, 1)$ conditional on the design $\mathcal{X}$, hence

$$Z = \mathbb{E}_{\theta \sim \pi} \left[\prod_{i \leq n} \frac{\exp(-\frac{1}{2}(y_i - (\mathcal{X}\theta)_i)^2)}{\exp(-\frac{1}{2}y_i^2)} \right]$$

$$= \mathbb{E}_{\theta \sim \pi} \left[\prod_{i \leq n} \exp(y_i(\mathcal{X}\theta)_i) \exp(-\frac{1}{2}((\mathcal{X}\theta)_i)^2) \right]$$

and can hence write

$$E_0[Z^2] = \int_{\mathbb{R}^n} \left(\mathbb{E}_{\theta \sim \pi} \left[\prod_{i \leq n} \exp(y_i(\mathcal{X}\theta)_i) \exp(-\frac{1}{2}((\mathcal{X}\theta)_i)^2) \right] \right)^2$$

$$\times \prod_{i \leq n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y_i^2}{2}) dy_1 \ldots dy_n$$

$$= \int_{\mathbb{R}^n} \left(\mathbb{E}_{\theta \sim \pi} \left[\exp(-\frac{1}{2}\|\mathcal{X}\theta\|_2^2) \prod_{i \leq n} \exp(y_i(\mathcal{X}\theta)_i) \right] \right)^2$$

$$\times \prod_{i \leq n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y_i^2}{2}) dy_1 \dots dy_n.$$

Thus, if θ, θ' are independent copies of joint law π^2, then we have

$$E_0\big[Z^2\big]$$
$$= \int_{\mathbb{R}^n} \mathbb{E}_{\pi^2} \Big[\exp(-\frac{1}{2}(\|\mathcal{X}\theta\|_2^2 - \frac{1}{2}(\|\mathcal{X}\theta'\|_2^2) \prod_{i \leq n} \frac{1}{\sqrt{2\pi}} \exp\big(y_i(\mathcal{X}(\theta + \theta'))_i - \frac{y_i^2}{2}\big) \Big]$$
$$\times dy_1 \dots dy_n$$
$$= \mathbb{E}_{\pi^2} \Bigg[\exp(-\frac{1}{2}\|\mathcal{X}\theta\|_2^2 - \frac{1}{2}\|\mathcal{X}\theta'\|_2^2)$$
$$\times \prod_{i \leq n} \int_{y_i} \Big(\frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(y_i - (\mathcal{X}(\theta + \theta'))_i)^2 \big) dy_i \exp\Big(\frac{1}{2}(\mathcal{X}(\theta + \theta'))_i^2\Big) \Bigg]$$
$$= \mathbb{E}_{\pi^2} \Big[\exp\Big(\frac{1}{2}\|\mathcal{X}(\theta + \theta')\|_2^2 - \frac{1}{2}\|\mathcal{X}\theta\|_2^2 - \frac{1}{2}\|\mathcal{X}\theta'\|_2^2\Big) \Big]$$

Step III: Integrating over $\mathcal{X}$: The E^X-expectation of the last expression can be bounded by

$$\mathbb{E}_{\pi^2} \Big[\exp\Big(\frac{n}{2}(\|\theta + \theta'\|_F^2 - \|\theta\|_F^2 - \|\theta'\|_F^2)\Big) E^X \exp\Big(\frac{1}{2}(Z_1 - Z_2 - Z_3)\Big) \Big]$$

where

$$Z_\ell = \|\mathcal{X}\vartheta_\ell\|_2^2 - n\|\vartheta_\ell\|_F^2, \quad \text{with} \quad \vartheta_1 = \theta + \theta', \vartheta_2 = \theta, \vartheta_3 = \theta'.$$

The last factor can be bounded, by applying the Cauchy-Schwarz inequality twice, by

$$(E^X \exp(Z_1))^{1/2} (E^X \exp(2Z_2))^{1/4} (E^X \exp(2Z_3))^{1/4}. \tag{11}$$

Since $\mathcal{X}\vartheta_\ell \sim N(0, \|\vartheta_\ell\|_F^2 I_n)$ the distribution of Z_ℓ is the one of $\|\vartheta_\ell\|_F^2 \sum_{i=1}^n (g_i^2 - 1)$ where the g_i are i.i.d. $N(0, 1)$. Applying Theorem 3.1.9 in [6] with $\tau_i \equiv 1$ and $\lambda = \|\vartheta_1\|_F^2$ or $\lambda = 2\|\vartheta_\ell\|_F^2, \ell = 2, 3$, (and hence setting $\|A\| = 1, \|A\|_{HS} = n$ in that theorem) we see that if $\max_\ell \|\vartheta_\ell\|_F^2 \leq 1/4$ then

$$E^X \exp(Z_1) \leq \exp\Big(\frac{n\|\vartheta_1\|_F^4}{1 - 2\|\vartheta_1\|_F^2} \Big), \quad \text{and}$$

$$E^X \exp(2Z_\ell) \leq \exp\Big(\frac{2n\|\vartheta_\ell\|_F^4}{1 - 4\|\vartheta_\ell\|_F^2} \Big), \quad \ell = 2, 3.$$

As a consequence if

$$\max_{\ell=1,2,3} \|\vartheta_\ell\|_F = o(n^{-1/4}) \tag{12}$$

then the the product (11) is bounded above by $1 + o(1)$. We conclude that if the prior π satisfies (12) almost surely then

$$\mathbb{E}_0[Z^2] = E^X E_0[Z^2] \leq (1 + o(1)) \times E_{\pi^2} \exp\left(\frac{n}{2}(\|\theta + \theta'\|_F^2 - \|\theta\|_F^2 - \|\theta'\|_F^2)\right)$$

$$= (1 + o(1)) \times E_{\pi^2} \exp\left(n\langle\theta, \theta'\rangle_F\right).$$

Step IV: Construction of π and bounds for $\mathbb{E}_0[Z^2]$: Assume for notational simplicity that d is an integer multiple of k, the general case needs only minor notational adjustment. Pick independent random $d \times 1$ vectors $v_\ell : \ell = 1, \ldots, k$ each of which consists of i.i.d. Rademacher entries (i.e., taking values ± 1 with probability $1/2$). Create a matrix W as follows: In the first d/k columns insert v_1 times a random sign $B_{1,j}, j = 1, \ldots, d/k$. Then, in the ℓ-th block repeat the same with v_1 replaced by v_ℓ, and random signs $B_{\ell,j}, j = 1, \ldots, d/k$. If $\|\cdot\| = \|\cdot\|_F$ let $\gamma_n = \rho_n/d$ and if $\|\cdot\| = \|\cdot\|_*$ set $\gamma_n = 2\rho_n/(\sqrt{k}d)$, so that in either case

$$\gamma_n = o\left(\min(\sqrt{1/dn}, d^{-1}n^{-1/4})\right).$$

Define the random matrix $\theta = \gamma_n W$ and let θ' be an independent copy of it. Thus

$$n\langle\theta, \theta'\rangle_F = n\gamma_n^2 \sum_{\ell=1}^{k} \sum_{m=1}^{d} \sum_{j=1}^{d/k} v_{\ell,m} B_{\ell,j} v'_{\ell,m} B'_{\ell,j}$$

$$= n\gamma_n^2 \sum_{\ell} \sum_{m} v_{\ell,m} v'_{\ell,m} \sum_{j} B_{\ell,j} B'_{\ell,j}.$$

As products of Rademacher variables are again Rademacher variables we have, for $\epsilon_{\ell,m}, \tilde{\epsilon}_{\ell,j}$ i.i.d. Rademacher variables (all defined on a suitable product probability space),

$$E_{\pi^2} \exp\left(n\langle\theta, \theta'\rangle_F\right) = E_\epsilon E_{\tilde{\epsilon}} \exp\left(n\gamma_n^2 \sum_{\ell} \sum_{m} \epsilon_{\ell,m} \sum_{j} \tilde{\epsilon}_{\ell,j}\right)$$

$$= \left(E_\epsilon E_{\tilde{\epsilon}} \exp\left(n\gamma_n^2 \sum_{m} \epsilon_{\ell,m} \sum_{j} \tilde{\epsilon}_{\ell,j}\right)\right)^k. \quad (13)$$

Conditional on the values of ϵ we set $\lambda = n\gamma_n^2 \sum_{m=1}^{d} \epsilon_{\ell,m}$ and note that

$$|\lambda| \leq nd\gamma_n^2 = o(1).$$

By Taylor expansion or standard properties of the hyperbolic cosine (as, e.g., in the proof of Theorem 6.2.9 in [6])

$$E_{\tilde{\epsilon}} \exp\left(\lambda \sum_{j=1}^{d/k} \tilde{\epsilon}_{\ell,j}\right) = \cosh(\lambda^2)^{d/k} \leq \exp\left(\lambda^2 d/k\right)$$

and thus, since $[EU]^k \le E[U^k]$ for any non-negative random variable U, the right hand side in (13) is bounded above by

$$
\left(E_\epsilon \exp\left(\lambda^2 d/k\right) \right)^k \le E_\epsilon \exp\left(\lambda^2 d\right)
$$

$$
= E_\epsilon \exp\left(n^2 \gamma_n^4 d \left(\sum_{m=1}^{d} \epsilon_m \right)^2 \right) \equiv E \exp\left(Z^2/c^2 \right),
$$

where the Rademacher sum $Z = \sum_{m=1}^{d} \epsilon_m$ is a sub-Gaussian random variable with variance proxy $\sigma^2 = d$ (cf. Section 2.3 in [6]). Thus by (2.24) in [6] we have

$$
E \exp\left(Z^2/c^2 \right) \le 1 + \frac{2}{c^2/2\sigma^2 - 1} = 1 + o(1)
$$

since

$$
\frac{c^2}{\sigma^2} = \frac{1}{d^2 n^2 \gamma_n^4} \to \infty
$$

as $n, d \to \infty$. Summarising all steps so far we conclude

$$
0 \le E^X E_0 [Z - 1]^2 = \mathbb{E}[Z^2] - 1 \le 1 - 1 + o(1) = o(1)
$$

noting that (12) holds π-almost surely in view of

$$
\|\theta\|_F^2 = \gamma_n^2 \|W\|_F^2 = \gamma_n^2 d^2 = o(n^{-1/2}).
$$

Step V: Asymptotic concentration of π on H_1: Finally we show that for the above prior we have indeed $\Pi(H_1) \to 1$. First since θ consists of columns that are linear combinations of at most k distinct vectors v_ℓ we immediately have $\theta \in R(k)$ almost surely. Moreover, for the case $\|\cdot\| = \|\cdot\|_F$ we have from the last display and by definition of γ_n that $\|\theta\|_F^2 = \rho_n^2$, so $\Pi(H_1) = 1$ follows.

For the case $\|\cdot\| = \|\cdot\|_*$ we have to show that

$$
\pi_{n,d}(\|\theta\|_* \ge \rho_n) \to 1
$$

as $\min(n, d) \to \infty$. We can transform θ into the $d \times k$ matrix θU consisting of k column vectors $\gamma_n \sqrt{d/k} v_\ell, \ell = 1, \dots, k$. The corresponding $d \times k$ matrix U consists of k column vectors, the ℓ-th of which has zero entries except for the indices $m \in [\ell d/k, \dots, -1 + (\ell + 1)d/k]$, where it equals $\sqrt{k/d} B_{\ell,m}$. Thus, U is an orthonormal projection matrix and we deduce that

$$
\|\theta\|_* \ge \|\theta U\|_*.
$$

We can renormalise the column vectors of θU so that

$$
\theta U = \gamma_n \frac{d}{\sqrt{k}} \left(\dots \frac{1}{\sqrt{d}} v_\ell \dots \right) \equiv \gamma_n \frac{d}{\sqrt{k}} V.
$$

The $d \times k$ matrix V consists of scaled i.i.d. Rademacher entries, and hence the proof of Lemma 1 in [14] (with $n = d, k = k_1 = p$ in the first display on p.2868

there) implies that, if $k/d \to 0$, then with probability as close to one as desired, the smallest singular value of V is bounded below by $1/2$ for d large enough. As a consequence $\|V\|_* \geq k/2$ and so, with probability approaching one,

$$\|\theta\|_* \geq \gamma_n d\sqrt{k}/2 = \rho_n.$$

Note that the same lower bound holds for

$$\|\theta - R(k_0)\|_* = \inf_{\theta' \in R(k_0)} \|\theta - \theta'\|_* \geq \sum_{j=k_0+1}^{k} |\lambda_j| \geq (k - k_0)/2 \qquad (14)$$

for any $k_0 < k$, if the absolute eigenvalues in the last display are assumed to be in decreasing order.

3.3. Proof of Theorem 4

Consider the composite testing problem

$$H_0 : \theta \in R(k_0) \quad vs. \quad H_1^c : \theta \in R(k_1), \ \|\theta - R(k_0)\|_* = \inf_{\theta' \in R(k_0)} \|\theta - \theta'\|_* \geq \rho.$$
$$(15)$$

From (14) with $k = k_1$ and $k_0 = o(k_1)$ we see that for $\min(n, d)$ large enough such that $(k_1 - k_0)/2 \geq k_1/4$, the prior π from the previous proof with $\gamma_n = 4\rho_n/(\sqrt{k}d)$ asymptotically concentrates on H_1^c. As a consequence testing (15) is no easier than when $H_0 = \{0\}$, so that when $\rho = o(\sqrt{k_1 d/n})$ then the proof of Part 2 of Theorem 1 implies

$$\liminf_{n,d} \inf_{\Psi} \left[\sup_{\theta \in H_0} \mathbb{E}_\theta \Psi + \sup_{\theta \in H_1^c} \mathbb{E}_\theta (1 - \Psi) \right] \geq 1. \qquad (16)$$

Now assume by way of contradiction that there exists C_n that satisfies (7) with $\alpha < 1/3$ and such that for every $c > 0$ there exist infinitely many n, d such that

$$\sup_{\theta \in H_0} \mathbb{E}_\theta |C_n|_* < c\sqrt{k_1 d/n}.$$

Passing to the infinite subsequence $\min(n, d) \to \infty$ along which the last inequalities hold, we deduce from Markov's inequality that

$$\sup_{\theta \in R(k_0)} \mathbb{P}_\theta(|C_n|_* \geq \alpha\sqrt{k_1 d/n}) \leq c/\alpha < \alpha$$

for c small enough depending only on α. Then, by Proposition 8.6.3 in [6] we can construct a test for (15) for which the testing errors in (16) are no more than $3\alpha < 1$ along the chosen subsequence, a contradiction that completes the proof.

Acknowledgement

References

[1] E. Arias-Castro, E.J. Candès, and Y. Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.*, 39:2533–2556, 2011. MR2906877

[2] A.D. Bull and R. Nickl. Adaptive confidence sets in L^2. *Probability Theory and Related Fields*, 156:889–919, 2013. MR3078289

[3] T.T. Cai and A. Zhang. ROP: matrix recovery via rank-one projections. *Ann. Statist.*, 43(1):102–138, 2015. MR3285602

[4] E.J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011. MR2809094

[5] A. Carpentier, J. Eisert, D. Gross, and R. Nickl. Uncertainty Quantification for Matrix Compressed Sensing and Quantum Tomography Problems. *arXiv:1504.03234*, 2015.

[6] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. to appear, Cambridge University Press, 2015.

[7] D. Gross, Y.-K. Liu, S.T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(15):150401, 2010.

[8] M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *Ann. Statist.*, 39:2382–2409, 2011. MR2906872

[9] Y.I. Ingster, A.B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Elec. J. Stat.*, 4:1476–1526, 2010. MR2747131

[10] Y.I. Ingster and I.A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003. MR1991446

[11] V. Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.*, 39(6):2936–2973, 2011. MR3012397

[12] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. MR2906869

[13] S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097, 2011. MR2816348

[14] R. Nickl and S. van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013. MR3161450