Statistica Sinica **27** (2017), 879-906
doi: https://doi.org/10.5705/ss.202016.0074

OPTIMAL ESTIMATION OF A QUADRATIC FUNCTIONAL UNDER THE GAUSSIAN TWO-SEQUENCE MODEL

T. Tony Cai and Xin Lu Tan

University of Pennsylvania

Abstract: This paper studies the problem of optimal estimation of a quadratic functional of two normal mean vectors, $Q(\mu, \theta) = (1/n)\sum_{i=1}^{n} \mu_i^2\theta_i^2$, with a particular focus on the case where both mean vectors are sparse. We propose optimal estimators of $Q(\mu, \theta)$ for different regimes and establish the minimax rates of convergence over a family of parameter spaces. The optimal rates exhibit interesting phase transitions in this family. We also include a simulation study to complement the theoretical results in the paper.

Key words and phrases: Gaussian sequence model, minimax estimation, quadratic functional, signal detection, sparse means.

1. Introduction

The problem of estimating the quadratic functional $\int f^2$ occupies an important position in nonparametric statistical inference literature. In the density estimation setting where one observes an i.i.d. sample from a distribution with density function f, Bickel and Ritov (1988) was the first to show that there is an interesting phase transition where the minimax rate of convergence for estimating $\int f^2$ under mean squared error is the usual parametric rate when the Hölder smoothness parameter of the density function is greater than 1/4, and is otherwise slower than the parametric rate. Giné and Nickl (2008) constructed an adaptive estimator of $\int f^2$ in the density estimation setting. Donoho and Nussbaum (1990) developed a minimax theory for estimating quadratic functionals of periodic functions in the nonparametric regression model.

Quadratic functional estimation has been particularly well studied in the Gaussian sequence model:

$$Y_i = \theta_i + \sigma_n z_i, \qquad i = 1, 2, \ldots, \tag{1.1}$$

where $z_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. The model (1.1) is equivalent to the white noise with drift model and can be used to approximate other nonparametric function estimation models. Estimating the quadratic functional $Q(\theta) = \sum \theta_i^2$ under (1.1) is the analog of estimating $\int f^2$ in the density estimation or nonparametric regression model. Fan (1991) and Efromovich and Low (1996) developed a minimax theory for estimating $Q(\theta) = \sum \theta_i^2$ over quadratically convex parameter spaces such as hyperrectangles and Sobolev balls. Cai and Low (2005, 2006b) further extended this theory to minimax and adaptive estimation over parameter spaces that are not necessarily quadratically convex. It is shown that the problem exhibits different phase transition phenomena in such a setting. A more recent paper by Collier, Comminges and Tsybakov (2015) gave a non-asymptotic analysis of estimation of the quadratic functional over ellipsoids and classes of sparse vectors. The focus so far has been on the one-sequence case.

There are close connections between the problem of quadratic functional estimation and that of signal detection under (1.1). Specifically, for a mean vector θ, we say that there is a signal at location i if $\theta_i \neq 0$. The problem of signal detection is then to distinguish between $\theta = 0$ and $\theta \neq 0$. Since $Q(\theta) = 0$ if and only if $\theta = 0$, it is not surprising that estimators of $Q(\theta)$ can be used to construct procedures that are effective for detecting signals. See, for instance, Cai and Low (2005) and the references therein. The results on estimating the quadratic functional $Q(\theta)$ also have important implications on hypothesis testing and construction of confidence balls. See, for example, Li (1989), Dümbgen (1998), Lepski and Spokoiny (1999), Ingster and Suslina (2003), Baraud (2004), Genovese and Wasserman (2005), and Cai and Low (2006a,b).

In this paper, we consider the estimation of the quadratic functional

$$Q(\mu, \theta) = \frac{1}{n} \sum_{i=1}^{n} \mu_i^2 \theta_i^2 \tag{1.2}$$

under the Gaussian two-sequence model,

$$X_i = \mu_i + \sigma z_i', \quad Y_i = \theta_i + \sigma z_i, \qquad i = 1, \ldots, n, \tag{1.3}$$

where $z_1', \ldots, z_n', \ z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$ and σ is the noise level. The goal is to optimally estimate $Q(\mu, \theta)$ based on the observed data (X_i, Y_i), $i = 1, ..., n$. Strictly speaking, $Q(\mu, \theta)$ is a quartic functional, but we will refer to it as a quadratic functional in the two-sequence case, as it is quadratic in μ given θ, and vice versa. We are particularly interested in the case where both mean vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\theta = (\theta_1, \ldots, \theta_n)$ are sparse.

In addition to being of significant theoretical interest in its own right, this estimation problem is also motivated by the problem of simultaneous signal detection in integrative genomics, where it is of interest to test whether there are single nucleotide polymorphisms (SNPs) that are simultaneously associated with multiple human traits or disorders (Consortium (2011); Cotsapas et al. (2011); Sivakumaran et al. (2011); Rankinen et al. (2015); Li et al. (2015)). More specifically, let X_i be the Z-score of the association between trait 1 and the i^{th} SNP, and let Y_i be the Z-score of the association between trait 2 and the i^{th} SNP, for $i = 1, \ldots, n$. When the SNPs are chosen from different linkage equilibrium blocks, then it is approximately true that the X_i's are independent, as are the Y_i's. Moreover, when X_i and Y_i are calculated in independent datasets, then for each i, X_i is independent of Y_i. In a simplified statistical framework, the simultaneous signal detection problem can then be studied under the Gaussian two-sequence model (1.3), where the goal is to detect the presence of location i with $\mu_i\theta_i \neq 0$. Equivalently, we want to distinguish between $\mu \star \theta = 0$ and $\mu \star \theta \neq 0$, where $\mu \star \theta = (\mu_1\theta_1, \ldots, \mu_n\theta_n)$ is the coordinate-wise product of μ and θ. Of particular interest is the setting where the proportion of signals is small, and the signal strengths are relatively weak. This is indeed the setting in the genomics context, as only a small number of SNPs are expected to be associated with both traits. Moreover, the association, if it exists, is weak. Since $Q(\mu, \theta) = 0$ if and only if $\mu \star \theta = 0$, one might expect a connection similar to that in the single Gaussian sequence model to exist between the estimation problem and the simultaneous signal detection problem. More discussions on the application of quadratic functional estimators to the problem of simultaneous signal detection are given in Section 4.

In this paper, we focus on studying the estimation of $Q(\mu, \theta)$. We propose optimal estimators of $Q(\mu, \theta)$ over a family of parameter spaces to be introduced, and establish the minimax rates of convergence. It is shown that the optimal rate exhibits interesting phase transitions in this family. Along with the establishment of the minimax rates of convergence, we explain the intuition behind the construction of the optimal estimators.

The rest of the paper is organized as follows: Section 2 considers estimation of the functional $Q(\mu, \theta)$ and establishes the minimax rates of convergence. Section 3 complements our theoretical study with some simulation results. We conclude the paper with a discussion in Section 4. Supplement S1 contains additional results that are not included in the main text. We present the proofs of some of the main results in Section 5, and relegate the rest to supplement S2 for the reason of space.

2. Optimal Estimation of $Q(\mu, \theta)$

In this section, we consider the estimation of the quadratic functional $Q(\mu, \theta)$

$= (1/n)\sum_{i=1}^{n}\mu_i^2\theta_i^2$ of two sparse normal mean vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\theta = (\theta_1, \ldots, \theta_n)$ under the Gaussian two-sequence model (1.3). An additional constraint is imposed on the number of coordinates that are simultaneously nonzero for both mean vectors. The noise level σ in model (1.3) is assumed to be known. Estimation of the noise level, σ, is relatively easy under the sparse sequence model (1.3) and will be discussed in Section 3.

We begin by introducing some notation that will be used throughout the paper. Given a vector $\theta = (\theta_1, \ldots, \theta_n)$, we denote by $\|\theta\|_0 = \mathrm{Card}(\{i : \theta_i \neq 0\})$ the ℓ_0-quasi-norm of θ, $\|\theta\|_2 = \sqrt{\sum_{i=1}^{n}\theta_i^2}$ its ℓ_2-norm, and $\|\theta\|_\infty = \max_{1\leq i\leq n}|\theta_i|$ its ℓ_∞-norm. For any real numbers a and b, we set $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$ and $a_+ = a \vee 0$. Throughout, the notation $a_n \asymp b_n$ means that there exists some numerical constants c and C such that $c \leq a_n/b_n \leq C$ when n is large. By "numerical constants" we usually mean constants that might depend on the characteristics of the problem but whose specific values are of little interest to us. The precise values of the numerical constants c and C may also vary from line to line.

Adopting an asymptotic framework where the vector size n is the driving variable, we parameterize the signal strength, sparsity, and simultaneous sparsity of μ and θ as functions of n. Specifically, we consider the family of parameter spaces

$$\Omega(\beta, \epsilon, b) = \{(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}^n : \|\mu\|_0 \leq k_n, \|\mu\|_\infty \leq s_n, \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n, \\ \|\mu \star \theta\|_0 \leq q_n\}, \tag{2.1}$$

indexed by three parameters β, ϵ, and b. We have the sparsity parameterization

$$k_n = n^\beta, \qquad 0 < \beta < \frac{1}{2}, \tag{2.2}$$

the simultaneous sparsity parameterization

$$q_n = n^\epsilon, \qquad 0 < \epsilon \leq \beta, \tag{2.3}$$

and the signal strength parametrization

$$s_n = n^b, \qquad b \in \mathbb{R}. \tag{2.4}$$

In principle, β can take any value between 0 and 1. We are primarily interested in the estimation problem for the range $0 < \beta < 1/2$, as it is well-known that this corresponds to the case of rare signals (Donoho and Jin (2004)).

Our goal is to derive the minimax rate of convergence for $Q(\mu, \theta)$ over $\Omega(\beta, \epsilon, b)$:

$$R^*(n, \Omega(\beta, \epsilon, b)) = \inf_{\widehat{Q}} \sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q} - Q(\mu, \theta))^2.$$

We will show that $R^*(n, \Omega(\beta, \epsilon, b))$ satisfies

$$R^*(n, \Omega(\beta, \epsilon, b)) \asymp \gamma_n(\beta, \epsilon, b), \tag{2.5}$$

where $\gamma_n(\beta, \epsilon, b)$ is a function of n indexed by β, ϵ and b. There are two main tasks in establishing the minimax rate of convergence. For each triple (β, ϵ, b) satisfying $0 < \epsilon \le \beta < 1/2$ and $b \in \mathbb{R}$, we construct an estimator $\widehat{Q}^*$ that satisfies

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}^* - Q(\mu, \theta))^2 \le C\gamma_n(\beta, \epsilon, b),$$

and show that $R^*(n, \Omega(\beta, \epsilon, b)) \ge c\gamma_n(\beta, \epsilon, b)$, where C and c are numerical constants that depend only on β, ϵ, b, and σ. Combining these upper and lower bounds yields the minimax rate of convergence (2.5). In this case, we say that the estimator $\widehat{Q}^*$ attains the minimax rate of convergence over the parameter space $\Omega(\beta, \epsilon, b)$.

Interestingly, the estimation problem exhibits different phase transitions for the minimax rate $\gamma_n(\beta, \epsilon, b)$ in three regimes: the *sparse* regime where $0 < \epsilon < \beta/2$, the *moderately dense* regime where $\beta/2 \le \epsilon \le 3\beta/4$, and the *strongly dense* regime where $3\beta/4 < \epsilon \le \beta$. Collectively, we call $\beta/2 \le \epsilon \le \beta$ the *dense* regime. In the sparse regime, the simultaneous signal is sparse in the sense that $q_n \ll \sqrt{k_n}$, while in the dense regime, the simultaneous signal is dense in the sense that $q_n \gg \sqrt{k_n}$. This is analogous to the terminology used in the one-sequence model, where the signal is called sparse if $0 < \beta < 1/2$ ($k_n \ll \sqrt{n}$), and dense if $1/2 \le \beta \le 1$ ($k_n \gg \sqrt{n}$). The key distinction is that, in the two-sequence case, sparseness or denseness is used to describe the relationship between simultaneous sparsity q_n and sparsity k_n, as opposed to between k_n and the vector size n. We remark that our use of the terminology is not superficial — a detailed analysis of lower and upper bounds for the estimation problem does reveal an intimate connection to the corresponding regimes in the one-sequence case. In particular, when the signal is moderately strong, the hardness of the two-sequence estimation problem is essentially characterized by an underlying one-sequence problem that displays different behavior in the sparse and the dense regimes. On the other hand, we construct optimal estimators for $Q(\mu, \theta)$, borrowing intuition from optimal estimators for $Q(\theta)$ in respective regimes.

Intuitively, when b is very small (i.e., signal is very weak), we are better off estimating $Q(\mu, \theta)$ by

$$\widehat{Q}_0 = 0, \tag{2.6}$$

since any attempt to estimate $Q(\mu, \theta)$ will incur a greater estimation risk. On the other hand, when b is sufficiently large (i.e., signal is strong), it is desirable to estimate $Q(\mu, \theta)$ based on the observed data (X_i, Y_i), $i = 1, \ldots, n$. With a slight abuse of terminology, we say that the signal is weak if it corresponds to the region where $\widehat{Q}_0$ is optimal, and we say that the signal is strong otherwise. In Sections 2.1 and 2.2, we construct two estimators of $Q(\mu, \theta)$ that respectively attain the minimax rates of convergence over the sparse and dense regimes when the signal is sufficiently large.

It is possible to generalize our parametrization to the case where μ and θ have different levels of both sparsity and signal strengths. This amounts to estimating $Q(\mu, \theta)$ over the parameter space

$$\begin{aligned}\Omega(\alpha, \beta, \epsilon, a, b) = \{(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}^n : \|\mu\|_0 \leq j_n, \|\mu\|_\infty \leq r_n, \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n, \\ \|\mu \star \theta\|_0 \leq q_n\}, \qquad (2.7)\end{aligned}$$

where $j_n = n^\alpha, k_n = n^\beta, q_n = n^\epsilon$ with $0 < \epsilon \leq \alpha \wedge \beta < 1/2$, and $r_n = n^a, s_n = n^b$ with $a, b \in \mathbb{R}$. In this section, however, we will focus on the simplest case where $j_n = k_n = n^\beta$ and $r_n = s_n = n^b$, since the technical analysis is similar to that for the more general case (2.7), but less tedious. We did derive the minimax rates of convergence for the case where $j_n = k_n = n^\beta$ but r_n and s_n are allowed to differ. As the phase transitions for the minimax rates of convergence in this case are much more sophisticated, but also are less easily digestible, we opt to defer its presentation to supplement S1. The analysis for the general case (2.7) where no equality constraint is imposed on either the sparsity or signal strength of μ and θ follows similarly, provided that the magnitude of the simultaneous sparsity ϵ is compared to α if $a \geq b$, and to β if $b \geq a$, for the determination of sparse and dense regimes.

2.1. Estimation in the sparse regime

We begin with the estimation of $Q(\mu, \theta) = (1/n) \sum \mu_i^2 \theta_i^2$ over the parameter space $\Omega(\beta, \epsilon, b)$ in the sparse regime, where q_n is calibrated as in expression (2.3) with $0 < \epsilon < \beta/2$.

To construct an optimal estimator for $Q(\mu, \theta)$, we base our intuition on the estimation of the quadratic functional $Q(\theta) = (1/n) \sum \theta_i^2$, in the case where we only have one sequence of observations Y_i, $i = 1, \ldots, n$, from model (1.3). Consider the family of parameter spaces indexed by $k_n = n^\beta, 0 < \beta < 1$ and $s_n = n^b, b \in \mathbb{R}$:

$$\Theta(\beta, b) = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n\}. \qquad (2.8)$$

It can be shown that for $0 < \beta < 1/2$, the minimax rate of convergence for $Q(\theta)$ over $\Theta(\beta, b)$ satisfies

$$R^*(n, \Theta(\beta, b)) := \inf_{\widehat{Q}} \sup_{\theta \in \Theta(\beta,b)} E_\theta(\widehat{Q} - Q(\theta))^2 \asymp \gamma_n(\beta, b), \tag{2.9}$$

where

$$\gamma_n(\beta, b) = \begin{cases} n^{2\beta+4b-2} & \text{if } b \leq 0, \\ n^{2\beta-2}(\log n)^2 & \text{if } 0 < b \leq \dfrac{\beta}{2}, \\ n^{\beta+2b-2} & \text{if } b > \dfrac{\beta}{2}. \end{cases} \tag{2.10}$$

When $0 < \beta < 1/2$, we have $k_n \ll \sqrt{n}$. Thus, we anticipate only very few coordinates of θ to be nonzero. If, in addition, $b < 0$, then the signal is both rare and weak, and one can do no better than simply estimating $Q(\theta)$ by $\widehat{Q}_0 = 0$. Nonetheless, when $b > 0$, the signal is rare but sufficiently strong, and the estimator

$$\widehat{Q}_1 = \frac{1}{n} \sum_{i=1}^{n} [(Y_i^2 - \sigma^2 \tau_n)_+ - \theta_0], \quad \text{where } \theta_0 := E(Z^2 - \sigma^2\tau_n)_+, Z \sim N(0, \sigma^2), \tag{2.11}$$

that performs coordinate-wise thresholding on Y_i^2 with choice of tuning parameter $\tau_n = 2 \log n$ is optimal. Each term θ_i^2 is estimated independently by $(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0$, since the sparsity pattern is unstructured. The estimator (2.11) involves a thresholding step, $(Y_i^2 - \sigma^2\tau_n)_+$, for denoising, and a de-bias step by subtracting θ_0 from the thresholded term so that we estimate the zero coordinates of θ unbiasedly. This is important because the proportion of zero entries in this case is relatively large, and a biased estimator for these coordinates will unnecessarily inflates the estimation risk.

The results on the estimation of one-sequence quadratic functional over classes of sparse vectors in (2.8)-(2.11) (and that over classes of dense vectors in (2.16)-(2.17)) are new, though we were made aware of the appearance of similar results in the concurrent work of Collier, Comminges and Tsybakov (2015). The focus and main contribution of our paper is on the estimation of the quadratic functional $Q(\mu, \theta)$ in the two-sequence case.

We now return to the sparse regime in the two-sequence setting, where $0 < \epsilon < \beta/2$ and $0 < \beta < 1/2$. In this case, $k_n \ll \sqrt{n}$, so the signal of individual sequences is rare. Moreover, the simultaneous sparsity $q_n \ll \sqrt{k_n}$ implies that we rarely have signals occurring simultaneously at the same coordinate of each sequence. This means that if we know for sure that μ_i is nonzero, it is unclear if θ_i is nonzero unless $|\theta_i|$ is large enough (and vice versa). Such an intuition motivates the estimator

$$\widehat{Q}_2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i^2 - \sigma^2\tau_n)_+ - \mu_0][(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0], \tag{2.12}$$

where $\mu_0 = \theta_0 := E(Z^2 - \sigma^2\tau_n)_+$ with the threshold level $\tau_n = \log n$, where $Z \sim N(0, \sigma^2)$. The construction of $\widehat{Q}_2$ is a straightforward extension of the construction of $\widehat{Q}_1$: each term $\mu_i^2\theta_i^2$ is estimated independently by the product $[(X_i^2 - \sigma^2\tau_n)_+ - \mu_0][(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0]$. Since $q_n \ll \sqrt{k_n}$, following our previous argument, thresholding X_i^2 and Y_i^2 *independently* at a common threshold level is natural.

We now present a theorem on the upper bound of the mean squared error of $\widehat{Q}_2$.

Theorem 1 (Sparse Regime: Upper Bound)**.** *For $b > 0$, the estimator $\widehat{Q}_2$, as in* (2.12) *with $\tau_n = \log n$, satisfies*

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2 - Q(\mu,\theta))^2 \leq C\Big[n^{2\epsilon+4b-2}(\log n)^2 + n^{\epsilon+6b-2}\Big]. \tag{2.13}$$

Straightforward calculation shows that for the estimator $\widehat{Q}_0 = 0$,

$$\begin{aligned}\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_0 - Q(\mu,\theta))^2 &= \sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} \left(\frac{1}{n}\sum_{i=1}^{n}\mu_i^2\theta_i^2\right)^2 \\ &= q_n^2 s_n^8 n^{-2} = n^{2\epsilon+8b-2}, \end{aligned}\tag{2.14}$$

for $0 < \epsilon \leq \beta < 1/2$ and $b \in \mathbb{R}$. We now show that the combination of $\widehat{Q}_0$ (when $b < 0$) and $\widehat{Q}_2$ (when $b \geq 0$) is optimal, by providing a matching lower bound.

Theorem 2 (Sparse Regime: Lower Bound)**.** *Let $0 < \epsilon < \beta/2$ and $0 < \beta < 1/2$. Then*

$$R^*(n, \Omega(\beta,\epsilon,b)) \geq c\gamma_n(\beta,\epsilon,b),$$

where

$$\gamma_n(\beta,\epsilon,b) = \begin{cases} n^{2\epsilon+8b-2} & \textit{if } b \leq 0, \\ n^{2\epsilon+4b-2}(\log n)^2 & \textit{if } 0 < b \leq \dfrac{\epsilon}{2}, \\ n^{\epsilon+6b-2} & \textit{if } b > \dfrac{\epsilon}{2}. \end{cases} \tag{2.15}$$

Crucial to the derivation of the lower bound is the Constrained Risk Inequality (CRI) given in Brown and Low (1996). To apply CRI, it suffices to construct two priors supported on $\Omega(\beta,\epsilon,b)$ that have small chi-square distance but a large difference in the expected values of the resulting quadratic functionals. The cases $b \leq \epsilon/2$ and $b > \epsilon/2$ correspond to choices of distinct pairs of priors. For $b > \epsilon/2$, the CRI boils down to the standard technique of inscribing a hardest hyperrectangle, with the Bayes risk for a simple prior supported on the hyperrectangle being a lower bound for the minimax risk. Nevertheless, the case $b \leq \epsilon/2$ requires the use of a rich collection of hyperrectangles and a mixture prior which mixes over the vertices of the hyperrectangles in this collection. Mixing increases the difficulty of the Bayes estimation problem and is needed here to attain a sharp lower bound.

Remark 1. Combining (2.13), (2.14) and (2.15), we see that when $0 < \epsilon < \beta/2$ and $0 < \beta < 1/2$, $\widehat{Q}_2$ attains the optimal rate of convergence over $\Omega(\beta, \epsilon, b)$ when $b > 0$. On the other hand, $\widehat{Q}_0$ attains the optimal rate of convergence over $\Omega(\beta, \epsilon, b)$ when $b \leq 0$.

Remark 2. So far, we have implicitly assumed that β is fixed and we characterize each regime by the relative magnitude of ϵ to β. It is possible to turn this view the other way around, to assume that ϵ is fixed and to characterize each regime by the relative magnitude of β to ϵ. We then see from (2.15) that within the sparse regime where $0 < 2\epsilon < \beta < 1/2$, the minimax rate of convergence $\gamma_n(\beta, \epsilon, b)$ for a fixed ϵ does not involve β. Such a lack of dependency on β is also highlighted in the two plot panels in the bottom row of Figure 1.

2.2. Estimation in the dense regime

We now consider estimating $Q(\mu, \theta)$ in the dense regime, where q_n is calibrated as in expression (2.3) with $\beta/2 \leq \epsilon \leq \beta$. The dense regime is subdivided into two cases: the moderately dense case with $\beta/2 \leq \epsilon \leq 3\beta/4$ and the strongly dense case with $3\beta/4 < \epsilon \leq \beta$.

In the dense regime, the estimator $\widehat{Q}_2$ defined in (2.12) is suboptimal, as the thresholding step in both X_i^2 and Y_i^2 ends up thresholding too many coordinates when the signal is weak. Note that the simultaneous sparsity $q_n \gg \sqrt{k_n}$ suggests that for each coordinate i with $\mu_i \neq 0$, it is more often the case that $\theta_i \neq 0$ (compared to when $q_n \ll \sqrt{k_n}$), and vice versa. Therefore, it is no longer reasonable to perform thresholding on X_i^2 and Y_i^2 independently. The additional knowledge of relatively high proportion of simultaneous nonzero entries suggests that whenever we observe a large value of X_i^2 (an implication of $\mu_i \neq 0$), then even if Y_i^2 is small, we should still estimate $\mu_i^2\theta_i^2$ rather than setting it equals zero. The same reasoning applies to the case where X_i^2 is small but Y_i^2 is large.

To construct an optimal estimator in the dense regime, we again borrow some intuition from the estimation of the quadratic functional $Q(\theta) = (1/n)\sum \theta_i^2$ in the one-sequence case. We consider the family of parameter spaces given in (2.8), but for $1/2 \leq \beta < 1$. The minimax rate of convergence once again satisfies (2.9), but with

$$\gamma_n(\beta, b) = \begin{cases} n^{2\beta+4b-2} & \text{if } b \leq \dfrac{1-2\beta}{4}, \\ n^{-1} & \text{if } \dfrac{1-2\beta}{4} < b \leq \dfrac{1-\beta}{2}, \\ n^{\beta+2b-2} & \text{if } b > \dfrac{1-\beta}{2}. \end{cases} \tag{2.16}$$

When $1/2 \leq \beta < 1$, we have $k_n \gg \sqrt{n}$, meaning that θ contains a relatively large number of non-zero coordinates compared to the case when $0 < \beta < 1/2$. The characterization of weak and strong signal is no longer $b < 0$ versus $b \geq 0$ as in the case of $0 < \beta < 1/2$, but $b \leq (1-2\beta)/4$ versus $b > (1-2\beta)/4$. That is, given the same signal strength b, the relatively large number of nonzero coordinates of θ when $k_n \gg \sqrt{n}$ collectively represents a stronger signal as compared to the case when $k_n \ll \sqrt{n}$. Thus, the threshold of "strong" signal as encoded by b is lowered when $k_n \gg \sqrt{n}$. It is not surprising that for the range of weak signal $b \leq (1-2\beta)/4$, the estimator $\widehat{Q}_0 = 0$ is optimal. On the other hand, when $b > (1-2\beta)/4$, the optimal estimator for $Q(\theta)$ is the unbiased estimator

$$\widehat{Q}_3 = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - \sigma^2). \tag{2.17}$$

An optimal estimator is often one that strikes an appropriate balance between bias and variance in its mean squared error. The estimators $\widehat{Q}_0$ and $\widehat{Q}_3$ represent two extremes in terms of bias-variance tradeoff. We see that the $\widehat{Q}_0$ that is optimal for exceedingly weak signal has zero variance, while the $\widehat{Q}_3$ that is optimal for sufficiently strong signal has zero bias. Due to the denseness of nonzero coordinates when $k_n \gg \sqrt{n}$, one could not afford to introduce bias to the estimator in the hope of achieving smaller variance. Without additional information about the sparsity structure, the unbiased estimator $\widehat{Q}_3$ is necessary for optimal estimation of $Q(\theta)$.

We now return to the two-sequence setting for the estimation of $Q(\mu, \theta)$, for the case $\beta/2 \leq \epsilon \leq \beta$ and $0 < \beta < 1/2$. Although the signal for individual sequences is sparse ($k_n \ll \sqrt{n}$), the simultaneous signal is dense in the sense that $q_n \gg \sqrt{k_n}$. The intuition garnered from the one-sequence case motivates the estimator

$$\widehat{Q}_4 = \frac{1}{n}\sum_{i=1}^{n}\big[(X_i^2-\sigma^2)(Y_i^2-\sigma^2)\mathbb{1}(X_i^2\vee Y_i^2>\sigma^2\tau_n)-\eta\big], \tag{2.18}$$

where

$$\eta = E[(Z_1^2-\sigma^2)(Z_2^2-\sigma^2)\mathbb{1}(Z_1^2\vee Z_2^2>\sigma^2\tau_n)], \qquad Z_1, Z_2 \overset{i.i.d.}{\sim} N(0,\sigma^2).$$

From $\widehat{Q}_4$, we see that each term $\mu_i^2\theta_i^2$ is estimated unbiasedly (modulo η) by $(X_i^2-\sigma^2)(Y_i^2-\sigma^2)$ whenever at least one of X_i^2 and Y_i^2 is sufficiently large. This is in accordance with our previous argument that estimation should be done whenever we have at least one large value of X_i^2 or Y_i^2. The threshold τ_n is a tuning parameter whose value is yet to be determined during the analysis of the mean squared error of $\widehat{Q}_4$, though it turns out that $\tau_n = c\log n$ for any $c \geq 4$ attains the optimal rate of convergence. The subtraction of η from $(X_i^2 - \sigma^2)(Y_i^2-\sigma^2)\mathbb{1}(X_i^2\vee Y_i^2>\sigma^2\tau_n)$ is needed because the majority of coordinates i has $\mu_i=\theta_i=0$. A biased estimator for these coordinates unavoidably inflates the estimation risk. The naive unbiased estimator

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^2-\sigma^2)(Y_i^2-\sigma^2)$$

does not seem to perform well when $0<\beta<1/2$ due to the rarity of nonzero coordinates in individual sequences. A thresholding step $\mathbb{1}(X_i^2\vee Y_i^2>\sigma^2\tau_n)$ is needed to guard against estimating entries with $\mu_i=\theta_i=0$ with noise.

Note that $\widehat{Q}_2$ defined in (2.12) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}[(X_i^2-\sigma^2\tau_n)\mathbb{1}(X_i^2>\sigma^2\tau_n)-\mu_0][(Y_i^2-\sigma^2\tau_n)\mathbb{1}(Y_i^2>\sigma^2\tau_n)-\theta_0].$$

Comparing this expression with $\widehat{Q}_4$, we see that when both X_i^2 and Y_i^2 are large, the term $\mu_i^2\theta_i^2$ is roughly estimated as $(X_i^2-\sigma^2\tau_n)(Y_i^2-\sigma^2\tau_n)$. Moreover, $(X_i^2-\sigma^2\tau_n)(Y_i^2-\sigma^2\tau_n)$ is a biased estimator of $\mu_i^2\theta_i^2$ when $\tau_n>1$.

We present an upper bound on the mean squared error of $\widehat{Q}_4$ in the following theorem.

Theorem 3 (Dense Regime: Upper Bound)**.** *For $b>0$, the estimator $\widehat{Q}_4$, as in* (2.18) *with $\tau_n = 4\log n$, satisfies*

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_4-Q(\mu,\theta))^2 \leq C\max\left\{n^{2\epsilon-2}(\log n)^4, n^{\epsilon+6b-2}, n^{\beta+4b-2}\right\}. \tag{2.19}$$

We now provide a matching lower bound to complement the upper bound in the dense regime.

Theorem 4 (Dense Regime: Lower Bound)**.** *Let $\beta/2 \le \epsilon \le \beta$ and $0 < \beta < 1/2$. Then*

$$R^*(n, \Omega(\beta, \epsilon, b)) \ge c\gamma_n(\beta, \epsilon, b),$$

where

$$\gamma_n(\beta, \epsilon, b) = \begin{cases} n^{2\epsilon+8b-2} & \text{if } b \le 0, \\ n^{2\epsilon-2}(\log n)^4 & \text{if } 0 < b \le \dfrac{2\epsilon-\beta}{4}, \\ n^{\beta+4b-2} & \text{if } \dfrac{2\epsilon-\beta}{4} < b \le \dfrac{\beta-\epsilon}{2}, \\ n^{\epsilon+6b-2} & \text{if } b > \dfrac{\beta-\epsilon}{2}, \end{cases} \tag{2.20}$$

when $\beta/2 \le \epsilon \le 3\beta/4$, and

$$\gamma_n(\beta, \epsilon, b) = \begin{cases} n^{2\epsilon+8b-2} & \text{if } b \le 0, \\ n^{2\epsilon-2}(\log n)^4 & \text{if } 0 < b \le \dfrac{\epsilon}{6}, \\ n^{\epsilon+6b-2} & \text{if } b > \dfrac{\epsilon}{6}, \end{cases} \tag{2.21}$$

when $3\beta/4 < \epsilon \le \beta$.

The minimax rates of convergence display different phase transitions within the two subdivisions of the dense regime. In the moderately dense regime where $\beta/2 \le \epsilon \le 3\beta/4$, there are phase transitions at $b = (2\epsilon - \beta)/4$ and $b = (\beta - \epsilon)/2$, given in (2.20). Note that $(2\epsilon - \beta)/4 \le (\beta - \epsilon)/2$ if and only if $\epsilon \le 3\beta/4$. In the strongly dense regime where $\epsilon > 3\beta/4$, the phase $(2\epsilon - \beta)/4 < b \le (\beta - \epsilon)/2$ is non-existent, and we only have one intermediate phase, $0 < b \le \epsilon/6$, given in (2.21).

We establish the lower bound by constructing least favorable priors and applying CRI. Except for the rate $n^{\epsilon+6b-2}$, which is obtained through the inscription of a hardest hyperrectangle, all other cases require some forms of mixing over the vertices of a rich collection of hyperrectangles.

Remark 3. Combining (2.14), (2.19), (2.20), and (2.21), we see that for the parameter space $\Omega(\beta, \epsilon, b)$ with $\beta/2 \le \epsilon \le \beta < 1/2$, $\widehat{Q}_4$ attains the minimax rate of convergence when $b > 0$. On the other hand, $\widehat{Q}_0 = 0$ attains the minimax rate of convergence when $b \le 0$.

Remark 4. Following Remark 2, we see that similar to the sparse regime, the minimax rate of convergence $\gamma_n(\beta, \epsilon, b)$ for a fixed ϵ does not involve β in the strongly dense regime where $\epsilon \leq \beta < 4\epsilon/3$. In contrast, $\gamma_n(\beta, \epsilon, b)$ for a fixed ϵ depends explicitly on β in the moderately dense regime where $4\epsilon/3 \leq \beta \leq 2\epsilon$. The dependency or lack of dependency of $\gamma_n(\beta, \epsilon, b)$ on β within each regime is also illustrated in the two plot panels in the bottom row of Figure 1.

Interestingly, in the two-sequence case, the regions $\{b : b \leq 0\}$ and $\{b : b > 0\}$ appear to constitute the regions of weak signal and strong signal, respectively, regardless of the level of simultaneous sparsity. This is in contrast to the one-sequence case where the dividing line is $b = 0$ when $k_n \ll \sqrt{n}$, and $b = (1 - 2\beta)/4$ when $k_n \gg \sqrt{n}$. We caution that this apparent "reconciliation" in the two-sequence case is simply because the signal strengths are taken to be the same for both sequences μ and θ in the simplified results presented above.

Remark 5. When the signal strengths $r_n = n^a$ and $s_n = n^b$ of μ and θ are allowed to differ, it turns out that $\{(a, b) : a \wedge b \leq 0\}$ characterizes the region of weak signal when $q_n \ll \sqrt{k_n}$, while $\{(a, b) : a \vee b \leq 0\} \cup \{(a, b) : a \wedge b \leq (\beta - 2\epsilon)/4\}$ comprises the region of weak signal when $q_n \gg \sqrt{k_n}$. We refer the readers to supplement S1 for more details.

2.3. Phase transitions in the minimax rates of convergence

We see from Sections 2.1 and 2.2 that within each regime, the minimax rates of convergence exhibit several phase transitions. In addition, each transition is governed by a change in the relative magnitudes of the sparsity parameter β, the simultaneous sparsity parameter ϵ, and the signal strength parameter b. In fact, it is the way phase transitions occur within each regime that characterizes the regime itself. Furthermore, the phase transitions actually display "continuity" across the boundaries of different regimes.

To depict what we meant graphically, first note that from Sections 2.1 and 2.2, the minimax rates of convergence

$$\gamma_n(\beta, \epsilon, b) \asymp n^{r(\beta, \epsilon, b)}, \tag{2.22}$$

modulo a factor involving $\log n$ when applicable. In Figure 1, we plot the rate exponent $r(\beta, \epsilon, b)$ against b for the sparse, moderately dense, and strongly dense regimes.

Specifically, in the top row of Figure 1, we fix $\beta = 0.45$ and plot $r(\beta, \epsilon, b)$ against b for a range of ϵ values in $(0, \beta)$. The top left panel of Figure 1 provides a continuum view of $r(\beta, \epsilon, b)$, as ϵ increases from 0 to β. Each piecewise straight line corresponds to an ϵ value in the considered range. To highlight the discrepancy among the three regimes, we color the sparse regime $(0 < \epsilon < \beta/2)$ in red,

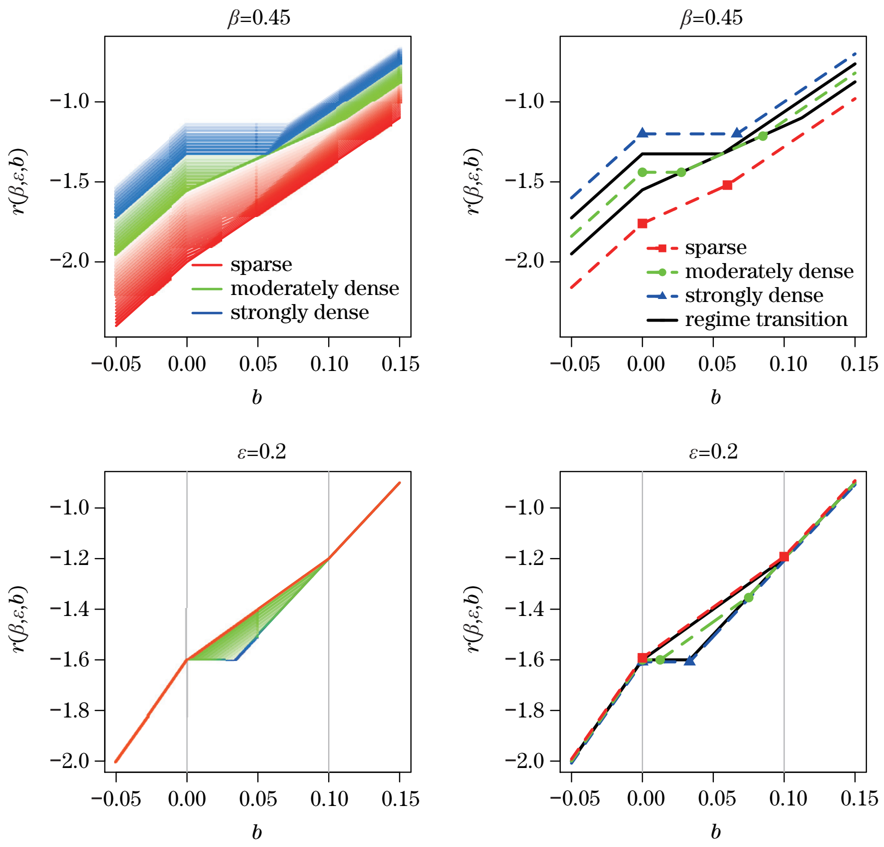

Figure 1. Plot of the rate exponent $r(\beta, \epsilon, b)$ against the signal strength b. In the sparse regime (🟥), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon + 4b - 2, \epsilon + 6b - 2$. In the moderately dense regime (🟩), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon - 2, \beta + 4b - 2, \epsilon + 6b - 2$. In the strongly dense regime (🟦), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon - 2, \epsilon + 6b - 2$. Top row, left panel: a continuum view of $r(\beta, \epsilon, b)$ as ϵ increases from 0 to $\beta = 0.45$ (color changes from red to blue). Top row, right panel: a static view of each regime: sparse ($\epsilon = 0.12$), moderately dense ($\epsilon = 0.28$), and strongly dense ($\epsilon = 0.4$). Transition points are indicated by the knots on the dashed lines. Bottom row, left panel: a continuum view of $r(\beta, \epsilon, b)$ as β increases from $\epsilon = 0.2$ to 0.5 (color changes from blue to red). Grey vertical lines indicate $b = 0$ and $b = \epsilon/2$. Bottom row, right panel: a static view of each regime: strongly dense ($\beta = 0.25$), moderately dense ($\beta = 0.35$), and sparse ($\beta = 0.45$).

the moderately dense regime ($\beta/2 \leq \epsilon \leq 3\beta/4$) in green, and the strongly dense regime ($3\beta/4 < \epsilon \leq \beta$) in blue. We see that the three regimes have somewhat different behaviors for small positive values of b. In particular, the sparse regime and the strongly dense regime experience two transitions (three different slopes), while the moderately dense regime experiences three transitions (four different slopes). Note that the difference in the number of transitions is restored at the intersection of the blue region and the red region. Thus, the phase transition is in some sense "continuous" across the regime boundaries — the piecewise straight lines corresponding to $r(\beta, \epsilon, b)$'s exhibit smooth transition as ϵ increases from 0 to β. The top right panel of Figure 1 provides a static view for each regime. We plot $r(\beta, \epsilon, b)$ against b for three values of ϵ corresponding to three different regimes: $\epsilon = 0.12$ (sparse regime), $\epsilon = 0.28$ (moderately dense regime), and $\epsilon = 0.4$ (strongly dense regime). The knots on each dashed line indicate the transition points for the slope of the line.

On the other hand, in the bottom row of Figure 1, we fix $\epsilon = 0.2$ and plot $r(\beta, \epsilon, b)$ against b for a range of β values in $(\epsilon, 0.5)$. The bottom left panel of Figure 1 provides a continuum view of $r(\beta, \epsilon, b)$, as β increases from ϵ to 0.5. Again, the strongly dense regime ($\epsilon \leq \beta < 4\epsilon/3$) is colored in blue, the moderately dense regime ($4\epsilon/3 \leq \beta \leq 2\epsilon$) in green, and the sparse regime ($\beta > 2\epsilon$) in red, with each piecewise straight line corresponding to a β value in the considered range. The two grey vertical lines indicate the locations $b = 0$ and $b = \epsilon/2$. Note that all the red lines overlap (so do all the blue lines), indicating that $r(\beta, \epsilon, b)$ for a fixed ϵ is independent of β in the sparse regime and the strongly dense regime. In the moderately dense regime, $r(\beta, \epsilon, b)$ only depends on β when $0 < b < \epsilon/2$. The bottom right panel of Figure 1 provides a static view for each regime. We plot $r(\beta, \epsilon, b)$ against b for three values of β: $\beta = 0.25$ (strongly dense regime), $\beta = 0.35$ (moderately dense regime), and $\beta = 0.45$ (sparse regime). Due to the overlap of all lines in the range $b \leq 0$ and $b > \epsilon/2$, we shift the dashed lines corresponding to $\beta = 0.45$ and $\beta = 0.25$ (in red and in blue, respectively) slightly to aid distinguishing the changes of $r(\beta, \epsilon, b)$ in different regimes.

3. Simulation

In this section, we report on simulation studies to compare the performance of the three estimators $\widehat{Q}_0 = 0$, $\widehat{Q}_2$ as in (2.12), and $\widehat{Q}_4$ as in (2.18), under different scenarios. We computed the mean squared error (MSE) of the three estimators to show that our simulation results are compatible with the theoretical results given in Section 2.

So far, we have assumed that the noise level σ is known. In practice, σ is typically unknown and needs to be estimated. Under the sparse setting of the present paper, σ is easily estimable. Let $M \in \mathbb{R}^{2n}$ have $M_{2i-1} = X_i$ and $M_{2i} = Y_i$ for $i = 1, \ldots, n$. A simple robust estimator of the noise level σ can be obtained from the median absolute deviation (MAD) of the combined sample:

$$\hat{\sigma} = \frac{\text{median}_j |M_j - \text{median}_k(M_k)|}{0.6745}.$$

Such an estimator has been used in Donoho and Johnstone (1994) for wavelet estimation.

We considered simulation studies over a range of sample size n, sparsity $k_n = n^{\beta}$, simultaneous sparsity $q_n = n^{\epsilon}$, and signal strength $s_n = n^b$. More specifically, we took $n \in \{10^3, 10^4, \ldots, 10^7\}$, $\beta = 0.45$ for individual sequences, $b \in \{-0.1, 0.15, 0.2\}$, and three values of simultaneous sparsity, one for each regime: $\epsilon = 0.02$ (sparse regime), $\epsilon = 0.3$ (moderately dense regime) and $\epsilon = 0.44$ (strongly dense regime). For each (n, β, ϵ, b), we generated data from the gaussian two-sequence model (1.3) with $\mu, \theta \in \{0, \pm n^b\}^n$, $\|\mu\|_0 = \|\theta\|_0 = [n^{\beta}]$, and $\|\mu \star \theta\|_0 = [n^{\epsilon}]$, where $[\cdot]$ denotes rounding to the nearest integer. Figure 2 is the plot of the MSE (averaged over 200 replications) of the three estimators against sample size in the log-log scale, for each combination of simultaneous sparsity and signal strength.

The theoretical results in Section 2 indicate that for $\widehat{Q} = \widehat{Q}_0, \widehat{Q}_2$, or $\widehat{Q}_4$,

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E(\widehat{Q} - Q(\mu,\theta))^2 \asymp n^{r(\beta,\epsilon,b)}$$

for some rate exponent $r(\beta, \epsilon, b)$ (modulo a logarithmic factor when applicable). Thus, it is not surprising that the results in Figure 2 (mostly) exhibit a linear pattern. When the signal is weak with $b = -0.1$ (see the first row of Figure 2), we see that $\widehat{Q}_0$ (wide-dashed line) and $\widehat{Q}_4$ (dotted line) have the lowest mean squared error. Note that we expect $\widehat{Q}_0$ to be optimal when the signal is weak. We observe that $\widehat{Q}_4$ is nearly as good as $\widehat{Q}_0$ from Figure 2. This is because when the signal is weak, the thresholding step $\mathbb{1}(X_i^2 \vee Y_i^2 \geq \sigma^2 \tau_n)$ thresholds both noise and weak signals, and the de-bias term η is extremely small when n is moderately large, resulting in $\widehat{Q}_4 \approx \widehat{Q}_0 = 0$. As the signal becomes sufficiently strong ($b \in \{0.15, 0.2\}$), $\widehat{Q}_2$ starts to dominate in the sparse regime ($\epsilon = 0.02$) while $\widehat{Q}_4$ dominates in the moderately dense and strongly dense regimes ($\epsilon \in \{0.3, 0.44\}$). When the signal is sufficiently large ($b \in \{0.15, 0.2\}$), $\widehat{Q}_0$ is clearly suboptimal. In particular, in the case where signal is both dense and strong ($b = 0.2, \epsilon \in \{0.3, 0.44\}$), the MSE of $\widehat{Q}_0$ diverges to infinity, as indicated by the positive slope of the wide-dashed line. Note also that as either ϵ or b increases, MSE increases, as can be seen by the flattening or reversing of slopes towards the right end or bottom of the plot panel. This is compatible with the fact that $r(\beta, \epsilon, b)$ increases with respect to both ϵ and b.

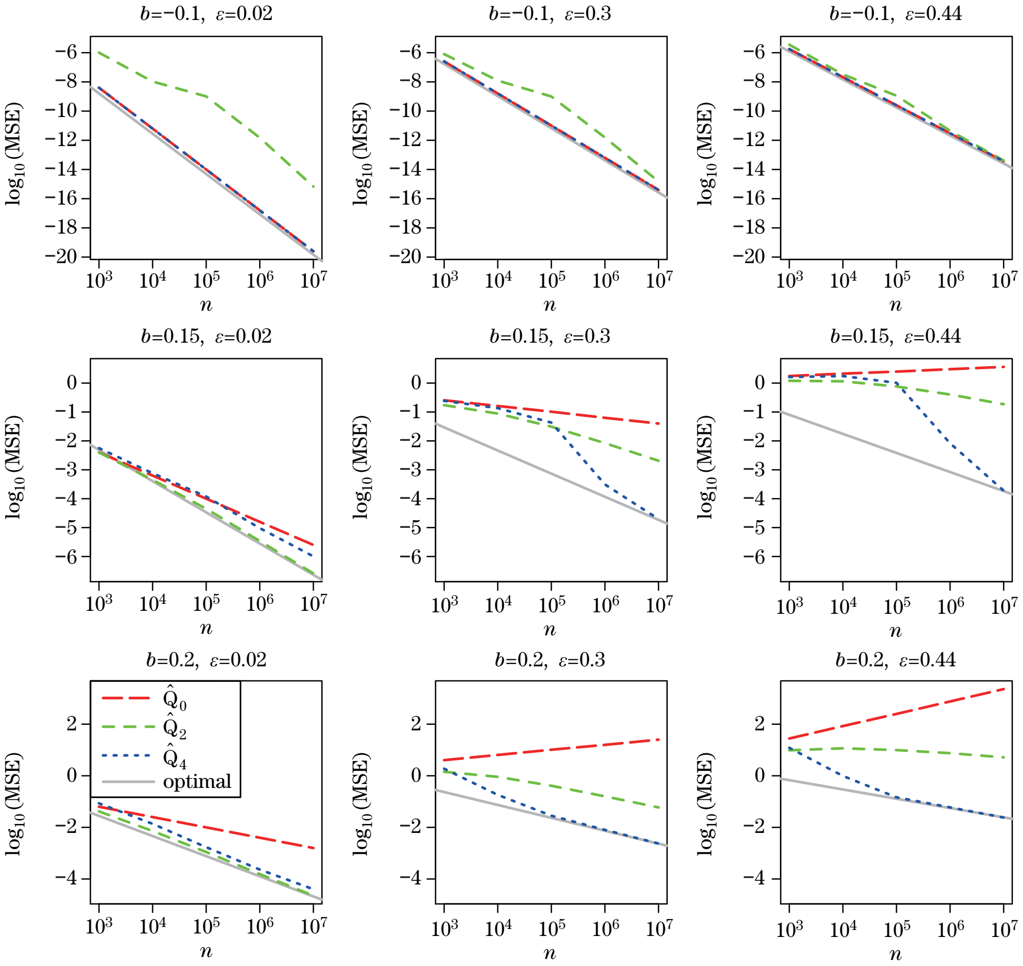

Figure 2. Plot of MSE for the estimators $\widehat{Q}_0$, $\widehat{Q}_2$, and $\widehat{Q}_4$ over different sample sizes $n \in \{10^3, \ldots, 10^7\}$, in the log-log scale. Fixing $\beta = 0.45$, the columns are ordered from left to right as $\epsilon = 0.02$ (sparse regime), $\epsilon = 0.3$ (moderately dense regime), and $\epsilon = 0.44$ (strongly dense regime). The rows are ordered from top to bottom in increasing signal strength: $b \in \{-0.1, 0.15, 0.2\}$. Solid line has a slope equal to that of the optimal rate exponent $r(\beta, \epsilon, b)$.

For each combination (β, ϵ, b), the solid line has a slope equal to the optimal rate exponent $r(\beta, \epsilon, b)$, and an intercept deliberately selected so that it lies close to the line corresponding to the optimal estimator. We see from Figure 2 that for all combinations of (b, ϵ) except $b = 0.15, \epsilon \in \{0.3, 0.44\}$, the slope of the solid line aligns well with that of the optimal estimator, confirming the validity of our theoretical results. We conjecture that in the case $b = 0.15, \epsilon \in \{0.3, 0.44\}$, the worst case rate of the optimal estimator $\widehat{Q}_4$ in $\Omega(\beta, \epsilon, b)$ is not attained at the configuration of location and magnitude of nonzero entries in μ, θ considered in the simulation. This can be seen from the fact that $\widehat{Q}_4$ has a steeper slope than the optimal one (i.e., faster rate of convergence) for sufficiently large n.

4. Discussion

In this paper, we discuss the estimation of the quadratic functional $Q(\mu, \theta) = (1/n)\sum \mu_i^2\theta_i^2$ over a family of parameter spaces where μ and θ are constrained in terms of the magnitude, sparsity, and simultaneous sparsity. Similar to the one-sequence estimation problem, we show that the minimax rates of convergence display different phase transitions over the sparse regime and the dense regime. Different from the one-sequence estimation problem, in the two-sequence case, the dense regime can be further subdivided into the moderately dense regime and the strongly dense regime. Despite the similarity in terminology, we emphasize that denseness and sparseness refer to the relationship between simultaneous sparsity and individual sparsity in the two-sequence problem, rather than that between sparsity and vector size as in the one-sequence problem. The construction of the optimal estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ are inspired by their one-sequence correspondence in respective regimes, with appropriate modification that accounts for the structure of the two-sequence problem.

Our study of the two-sequence estimation problem can be generalized in several aspects. In supplement S1, we show that the optimal rates of convergence for estimation of $Q(\mu, \theta)$ continue to subsume the aforementioned regimes, when μ and θ are allowed unequal signal strengths. Moreover, the optimal rates are attained by the same estimators in respective regimes. Nonetheless, the distinction between the sparse and dense regimes is more apparent in this setting. In the sparse regime, estimation is only desirable when the signal strengths of both sequences are sufficiently strong. In contrast, in the dense regime, estimation is desirable whenever at least one sequence admits a sufficiently strong signal (and the signal strength of the other sequence is not too weak). Throughout the paper, we assume that the sequences $\{X_i : 1 \le i \le n\}$ and $\{Y_i : 1 \le i \le n\}$ have a common noise level σ. Our analysis can be easily extended to the case where $\sigma_X \neq \sigma_Y$, by appropriately replacing the threshold levels in the proposed estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ with ones that involve σ_X or σ_Y. Such a modification yields estimators which attain minimax rates of convergence that are identical to that given in the paper. When σ_X and σ_Y are unknown, we can use MAD to estimate the noise level of each sequence and plug in to the modified estimators.

The focus of this paper is on minimax rates of convergence for the estimation of $Q(\mu, \theta)$. Adaptive estimation of $Q(\mu, \theta)$ is an interesting but technically challenging problem. Cai and Low (2005) introduced a block thresholding estimator for adaptive estimation of the quadratic functional in the one-sequence setting. It would be interesting to explore whether a similar idea could be used for adaptively estimating the quadratic functional in the two-sequence setting. In this paper, we consider the estimation of $Q(\mu, \theta)$ over the parameter space defined in (2.1), where signal strengths are incorporated through the ℓ_∞-norm. For future work, it would also be interesting to study the behavior of the estimation problem under an ℓ_p-norm constraint on the signal strengths, where $p \in (0, \infty)$.

A problem that is closely related to the estimation of the quadratic functional $Q(\mu, \theta)$ is the simultaneous signal detection problem, where the goal is to distinguish between $\mu \star \theta = 0$ and $\mu \star \theta \neq 0$. In the single Gaussian sequence setting where one observes $Y_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$, it is of interest to test $\theta = 0$ against $\theta \neq 0$, and there are two natural approaches: the sum of squares type test statistic $\sum Y_i^2$ and the max-type test statistic $\max |Y_i|$. Simultaneous signal detection generalizes the one-sequence testing problem and arises frequently in the context of integrative genomics. In genetics, for instance, it is often of interest to identify polymorphisms that are associated with multiple related conditions (Rankinen et al. (2015); Li et al. (2015)). The problem of simultaneous signal detection has been studied by Zhao, Cai and Li (2014) under a mixture model framework, and a max-type statistic, $\max(|X_i| \wedge |Y_i|)$, is proposed for detecting sparse simultaneous signals. On the other hand, in this paper we study the estimation of quadratic functional under the sequence model framework. The proposed estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ can be applied to the simultaneous signal detection problem as well. Similar to the problem of quadratic functional estimation, it turns out that the simultaneous signal detection problem behaves differently over two regimes. In the dense regime, a signal is detectable provided the signal strength of at least one of the sequences is sufficiently strong and the signal strength of the other sequence is not too weak. In contrast, in the sparse regime, a signal is only detectable when both sequences admit sufficiently strong signals. A crude analysis shows that the test procedures based on the statistics $\widehat{Q}_2$ and $\widehat{Q}_4$ are effective in detecting simultaneous signals over the respective detectable regions. A complete analysis of the optimality and adaptivity of such a test procedure is an interesting but challenging problem which we leave for future work.

5. Proofs of Theorems 1 and 2

In this section, we present the proofs of Theorems 1 and 2 and, for reasons of space, we relegate the proofs of Theorems 3 and 4 to supplement S2.

Henceforth, we omit the subscripts n in k_n, q_n, s_n and τ_n that signifies their dependence on the sample size. We denote by ψ_μ the density of a Gaussian distribution with mean μ and variance σ^2, and we denote by $\ell(n,k)$ the class of all subsets of $\{1,\ldots,n\}$ of k distinct elements. We let $\phi(z)$ and $\Phi(z) = P(Z \le z)$ be the density and cumulative distribution function of a standard normal random variable Z, respectively. Finally, c and C denote generic positive constants whose values may vary for each occurrence.

5.1. Proof of Theorem 1

We need a lemma from Cai and Low (2005) (Lemma 1, page 2939) for proving Theorem 1.

Lemma 1. *Let $Y \sim N(\theta,\sigma^2)$ and let $\theta_0 = E(Z^2 - \sigma^2\tau)_+$, where $Z \sim N(0,\sigma^2)$. Then for $\tau \ge 1$ and $\widehat{\theta^2} = (Y^2 - \sigma^2\tau)_+ - \theta_0$,*

$$|\theta_0| \le \frac{4\sigma^2}{\sqrt{2\pi}\tau^{1/2}e^{\tau/2}},$$
$$|E(\widehat{\theta^2}) - \theta^2| \le \min\{2\sigma^2\tau, \theta^2\},$$
$$\operatorname{Var}(\widehat{\theta^2}) \le 6\sigma^2\theta^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}.$$

Lemma 2 is an immediate consequence of Lemma 1.

Lemma 2. *Let $Y \sim N(\theta,\sigma^2)$ and let $\theta_0 = E(Z^2 - \sigma^2\tau)_+$, where $Z \sim N(0,\sigma^2)$. Then for $\tau \ge 1$,*

$$(E(Y^2 - \sigma^2\tau)_+ - \theta_0)^2 \le \max\left\{6\sigma^2\theta^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}, 10\theta^4\right\}. \tag{5.1}$$

Proof. Let $B(\theta) = E(Y^2 - \tau\sigma^2)_+ - \theta_0$. We first note that $B(-\theta) = B(\theta) \ge 0$ for $\theta \ge 0$. This follows from

$$\begin{aligned} B'(\theta) &= 2\sigma[\phi(\tau^{1/2} - \frac{\theta}{\sigma}) - \phi(\tau^{1/2} + \frac{\theta}{\sigma})] \\ &\quad - 2\theta[\Phi(\tau^{1/2} - \frac{\theta}{\sigma}) - \Phi(-\tau^{1/2} - \frac{\theta}{\sigma}) - 1] \\ &\ge 0 \end{aligned}$$

and $B(0) = 0$. So we have $B(\theta) = E(Y^2 - \tau\sigma^2)_+ - \theta_0 \ge 0$ for all $\theta \in \mathbb{R}$. It follows that $(E[(Y^2 - \tau\sigma^2)_+ - \theta_0])^2 \le (E(Y^2 - \tau\sigma^2)_+)^2 \le E[(Y^2 - \tau\sigma^2)_+^2]$. To bound the term $E[(Y^2-\tau\sigma^2)_+^2]$, we consider two cases: $\theta \le \sigma$ and $\theta > \sigma$. It follows from the proof of Lemma 1 in Cai and Low (2005) that when $\theta \le \sigma$, then

$$E[(Y^2-\tau\sigma^2)_+^2] \le 6\sigma^2\theta^2 + \sigma^4 \frac{4\tau^{1/2}+18}{e^{\tau/2}}.$$

On the other hand, when $\theta > \sigma$, we have

$$E[(Y^2-\tau\sigma^2)_+^2] \le E[Y^4] = \theta^4 + 6\sigma^2\theta^2 + 3\sigma^4 \le 10\theta^4.$$

If follows that (5.1) holds.

Proof of Theorem 1. We first bound the bias of the estimator $\widehat{Q}_2$ defined in (2.12). Using the equality

$$AB - ab = (A-a)(B-b) + a(B-b) + b(A-a),$$

the independence of X_i and Y_i, and the triangle inequality, we get

$$\begin{aligned}
&\left|E_{(\mu_i,\theta_i)}\{[(X_i^2-\sigma^2\tau)_+ - \mu_0][(Y_i^2-\sigma^2\tau)_+ - \theta_0]\} - \mu_i^2\theta_i^2\right| \\
&\quad \le \left|E_{\mu_i}[(X_i^2-\sigma^2\tau)_+ - \mu_0] - \mu_i^2\right| \cdot \left|E_{\theta_i}[(Y_i^2-\sigma^2\tau)_+ - \theta_0] - \theta_i^2\right| \\
&\qquad\quad + \mu_i^2\left|E_{\theta_i}[(Y_i^2-\sigma^2\tau)_+ - \theta_0] - \theta_i^2\right| + \theta_i^2\left|E_{\mu_i}[(X_i^2-\sigma^2\tau)_+ - \mu_0] - \mu_i^2\right| \\
&\quad \le \min\{2\sigma^2\tau, \mu_i^2\}\min\{2\sigma^2\tau, \theta_i^2\} + \mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + \theta_i^2\min\{2\sigma^2\tau, \mu_i^2\} \\
&\quad \le 2\mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + 2\theta_i^2\min\{2\sigma^2\tau, \mu_i^2\},
\end{aligned}$$

the second inequality follows from Lemma 1. It follows that, for $(\mu,\theta) \in \Omega(\beta,\epsilon,b)$ and $\tau \ge 1$,

$$\begin{aligned}
&|E_{(\mu,\theta)}(\widehat{Q}_2) - Q(\mu,\theta)| \\
&\quad = \left|\frac{1}{n}\sum_{i=1}^n E_{(\mu_i,\theta_i)}\{[(X_i^2-\sigma^2\tau)_+ - \mu_0][(Y_i^2-\sigma^2\tau)_+ - \theta_0]\} - \frac{1}{n}\sum_{i=1}^n \mu_i^2\theta_i^2\right| \\
&\quad \le \frac{2}{n}\sum_{i=1}^n \left[\mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + \theta_i^2\min\{2\sigma^2\tau, \mu_i^2\}\right] \\
&\quad \le \frac{4}{n}\min\{2\sigma^2 q s^2\tau, qs^4\},
\end{aligned}$$

the second inequality follows from the fact that, for $(\mu,\theta) \in \Omega(\beta,\epsilon,b)$, there are at most q entries that are simultaneously nonzero for μ and θ.

We now proceed to bound the variance of $\widehat{Q}_2$. Applying the equality

$$\mathrm{Var}(AB) = \mathrm{Var}(A)\mathrm{Var}(B) + [E(A)]^2\mathrm{Var}(B) + [E(B)]^2\mathrm{Var}(A),$$

for $\tau \ge 1$, we have

$$\mathrm{Var}_{(\mu_i,\theta_i)}\{[(X_i^2-\sigma^2\tau)_+ - \mu_0][(Y_i^2-\sigma^2\tau)_+ - \theta_0]\}$$

$$\begin{aligned}
&= \mathrm{Var}_{\mu_i}[(X_i^2-\sigma^2\tau)_+-\mu_0]\mathrm{Var}_{\theta_i}[(Y_i^2-\sigma^2\tau)_+-\theta_0]\\
&\quad+[E_{\mu_i}(X_i^2-\sigma^2\tau)_+-\mu_0]^2\mathrm{Var}_{\theta_i}[(Y_i^2-\sigma^2\tau)_+-\theta_0]\\
&\quad+[E_{\theta_i}(Y_i^2-\sigma^2\tau)_+-\theta_0]^2\mathrm{Var}_{\mu_i}[(X_i^2-\sigma^2\tau)_+-\mu_0]\\
&\le 3\left[6\sigma^2\mu_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\left[6\sigma^2\theta_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\\
&\quad+10\mu_i^4\left[6\sigma^2\theta_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]+10\theta_i^4\left[6\sigma^2\mu_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right],
\end{aligned}$$

the inequality follows from Lemma 1 and Lemma 2. Thus, for $(\mu,\theta)\in\Omega(\beta,\epsilon,b)$ and $\tau\ge 1$,

$$\begin{aligned}
&\mathrm{Var}_{(\mu,\theta)}(\widehat{Q}_2)\\
&=\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_{(\mu_i,\theta_i)}\{[(X_i^2-\sigma^2\tau)_+-\mu_0][(Y_i^2-\sigma^2\tau)_+-\theta_0]\}\\
&\le\frac{3}{n^2}\sum_{i=1}^n\left[6\sigma^2\mu_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\left[6\sigma^2\theta_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\\
&\quad+\frac{10}{n^2}\sum_{i=1}^n\mu_i^4\left[6\sigma^2\theta_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]+\frac{10}{n^2}\sum_{i=1}^n\theta_i^4\left[6\sigma^2\mu_i^2+\sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\\
&\le\frac{3}{n^2}\left[36\sigma^4qs^4+12\sigma^6ks^2\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)+n\sigma^8\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\right]\\
&\quad+\frac{20}{n^2}\left[6\sigma^2qs^6+\sigma^4ks^4\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)\right].
\end{aligned}$$

Combining the bias and variance term, we get, for $\tau\ge 1$,

$$\begin{aligned}
&\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2-Q(\mu,\theta))^2\\
&\le\frac{C}{n^2}\Bigg[\min\{q^2s^4\tau^2,q^2s^8\}+\max\Bigg\{qs^4,qs^6,ks^2\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right),\\
&\qquad\qquad ks^4\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right),n\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\Bigg\}\Bigg]\\
&=\frac{C}{n^2}\Bigg[\min\{n^{2\epsilon+4b}\tau^2,n^{2\epsilon+8b}\}+\max\Bigg\{n^{\epsilon+4b},n^{\epsilon+6b},n^{\beta+2b}\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right),\\
&\qquad\qquad n^{\beta+4b}\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right),n\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\Bigg\}\Bigg].
\end{aligned}$$

Suppose that $b > 0$. Then letting $\tau = \log n$ leads to

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2 - Q(\mu,\theta))^2 \leq C\left[n^{2\epsilon+4b-2}(\log n)^2 + n^{\epsilon+6b-2}\right].$$

5.2. Proof of Theorem 2

To prove Theorem 2, it suffices to show that for $0 < \beta < 1/2$,

$$\gamma_n(\beta,\epsilon,b) \geq \begin{cases} n^{2\epsilon+4b-2}(\log n)^2 & \text{if } b > 0, \quad \text{for } 0 < \epsilon < \frac{\beta}{2}, \qquad \text{(Case 1)} \\ n^{2\epsilon+8b-2} & \text{if } b \leq 0, \quad \text{for } 0 < \epsilon \leq \beta, \qquad \text{(Case 2)} \\ n^{\epsilon+6b-2} & \text{if } b > 0, \quad \text{for } 0 < \epsilon \leq \beta. \qquad \text{(Case 3)} \end{cases}$$

For individual regions in $\{(\beta,\epsilon,b) : 0 < \epsilon < \beta/2, 0 < \beta < 1/2, b \in \mathbb{R}\}$, the minimax rate of convergence is then given by the sharpest rate among all cases in which the region belongs. For instance, the region $\{(\beta,\epsilon,b) : 0 < \epsilon < \beta/2, 0 < \beta < 1/2, b > \epsilon/2\}$ is included in Case 1 and Case 3, hence $\gamma_n(\beta,\epsilon,b) \geq \max\{n^{2\epsilon+4b-2}(\log n)^2, n^{\epsilon+6b-2}\} = n^{\epsilon+6b-2}$.

To establish the desired lower bounds, for each case we construct two priors f and g that have small chi-square distance but a large difference in the expected values of the resulting quadratic functionals, then apply the Constrained Risk Inequality (CRI) in Brown and Low (1996). The choice of priors f and g is crucial in deriving sharp lower bound for the estimation problem. In fact, the fundamental difference between different phases in the sparse regime for the estimation of $Q(\mu,\theta)$ can be seen from the choices f and g. For some background on lower bound technique, see supplement S2.

Proof of Case 1. Our proof builds on arguments similar to that used in Cai and Low (2004) and Baraud (2002), who considered the one-sequence estimation problem. We first follow the lines of the proof of Theorem 7 in Cai and Low (2004), and then apply a result from Aldous (1985) as was done in Baraud (2002). Let

$$f(x_1,\ldots,x_n,y_1,\ldots,y_n) = \prod_{i=1}^{k}\psi_s(x_i)\prod_{i=k+1}^{n}\psi_0(x_i)\prod_{i=1}^{n}\psi_0(y_i).$$

For $I \in \ell(k,q)$, let

$$g_I(x_1,\ldots,x_n,y_1,\ldots,y_n) = \prod_{i=1}^{k}\psi_s(x_i)\prod_{i=k+1}^{n}\psi_0(x_i)\prod_{i=1}^{k}\psi_{\theta_i}(y_i)\prod_{i=k+1}^{n}\psi_0(y_i),$$

where $\theta_i = \rho\mathbb{1}(i \in I)$ with $\rho > 0$, and let

$$g = \frac{1}{\binom{k}{q}} \sum_{I \in \ell(k,q)} g_I.$$

In both f and g, the sequence $\mu = (s, \ldots, s, 0, \ldots, 0)$ is taken to be the same. However, θ is taken to be all zeros in f but is taken as a mixture in g. The nonzero coordinates of θ are mixed uniformly over the support of μ at a common magnitude ρ, whose value is yet to be determined. Our choice of f and g essentially reduces the two-sequence problem to the case where we only have one Gaussian mean sequence of length k with q nonzero coordinates, hence explains the correspondence between the sparse regime in the two-sequence case ($q \ll \sqrt{k}$) and the sparse regime in the one-sequence case ($k \ll \sqrt{n}$).

We now compute the chi-square affinity between f and g,

$$\int \frac{g^2}{f} = \frac{1}{\binom{k}{q}^2} \sum_{I \in \ell(k,q)} \sum_{J \in \ell(k,q)} \int \frac{g_I g_J}{f}. \tag{5.2}$$

For $I, J \in \ell(k,q)$, let $m = \text{Card}(I \cap J)$. Then

$$\begin{aligned} \int \frac{g_I g_J}{f} &= \prod_{i=1}^{k} \int \frac{\psi_{\rho \mathbb{1}(i \in I)}(y_i) \cdot \psi_{\rho \mathbb{1}(i \in J)}(y_i)}{\psi_0(y_i)}\ dy_i \\ &= \left[\int \psi_0(y)\ dy\right]^{k-2q+m} \left[\int \psi_\rho(y)\ dy\right]^{2q-2m} \left[\int \frac{\psi_\rho^2(y)}{\psi_0(y)}\ dy\right]^{m} \\ &= \exp\left(\frac{m\rho^2}{\sigma^2}\right). \end{aligned}$$

It follows that

$$\int \frac{g^2}{f} = E\left[\exp\left(\frac{M\rho^2}{\sigma^2}\right)\right],$$

where M has the hypergeometric distribution

$$P(M = m) = \frac{\binom{q}{m}\binom{k-q}{q-m}}{\binom{k}{q}}. \tag{5.3}$$

As shown in Aldous (1985), M has the same distribution as the conditional expectation $E(\tilde{M}|\mathcal{B})$, where $\tilde{M}$ is a Binomial(q, q/k) random variable and $\mathcal{B}$ is a suitable σ-algebra. Coupled with Jensen's inequality, this implies that

$$\int \frac{g^2}{f} \le E\left[\exp\left(\frac{\tilde{M}\rho^2}{\sigma^2}\right)\right] = \left(1 - \frac{q}{k} + \frac{q}{k} e^{\rho^2/\sigma^2}\right)^q.$$

Taking $\rho = \sigma\sqrt{(\beta - 2\epsilon)\log n}$ gives

$$e^{\rho^2/\sigma^2} = n^{\beta - 2\epsilon} = \frac{k}{q^2},$$

hence

$$\int \frac{g^2}{f} \leq \left(1 + \frac{1}{q}\right)^q \leq e.$$

Since $Q(\mu, \theta) = 0$ under f and $Q(\mu, \theta) = (1/n)qs^2\rho^2$ under g, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c\left(\frac{1}{n}qs^2\rho^2\right)^2 = cn^{2\epsilon+4b-2}(\log n)^2.$$

Proof of Case 2. Let

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^n \psi_0(x_i) \prod_{i=1}^n \psi_0(y_i)$$

For $I \in \ell(n, q)$, let

$$g_I(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^n \psi_{\mu_i}(x_i) \prod_{i=1}^n \psi_{\theta_i}(y_i),$$

where $\mu_i = \theta_i = \rho\mathbb{1}(i \in I)$ with $\rho > 0$, and let

$$g = \frac{1}{\binom{n}{q}} \sum_{I \in \ell(n,q)} g_I.$$

Contrast the choice of f an g here with that used in the proof of Case 1. Rather than fixing μ and mixing nonzero coordinates of θ over the support of μ, in this case mixing is done over all n positions using nonzero coordinates of μ and θ simultaneously.

Similar calculation as that used in the proof of Case 1 yields

$$\int \frac{g^2}{f} \leq \left(1 - \frac{q}{n} + \frac{q}{n}e^{2\rho^2/\sigma^2}\right)^q. \tag{5.4}$$

Now take $\rho = s = n^b$. Since $b < 0$, it follows that when n is sufficiently large,

$$e^{2\rho^2/\sigma^2} \leq n^{1-2\epsilon} = \frac{n}{q^2},$$

hence

$$\int \frac{g^2}{f} \leq \left(1 + \frac{1}{q}\right)^q \leq e.$$

Since $Q(\mu, \theta) = 0$ under f, and $Q(\mu, \theta) = (1/n)q\rho^4$ under g, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c\left(\frac{1}{n}q\rho^4\right)^2 = cn^{2\epsilon+8b-2}.$$

Proof of Case 3. The priors used in this case are very different from that considered in the proofs of Case 1 and Case 2. Let

$$f(x_1,\dots,x_n,y_1,\dots,y_n)=\prod_{i=1}^{q}\psi_s(x_i)\prod_{i=q+1}^{n}\psi_0(x_i)\prod_{i=1}^{q}\psi_s(y_i)\prod_{i=q+1}^{n}\psi_0(y_i),$$
$$g(x_1,\dots,x_n,y_1,\dots,y_n)=\prod_{i=1}^{q}\psi_s(x_i)\prod_{i=q+1}^{n}\psi_0(x_i)\prod_{i=1}^{q}\psi_{s-\delta}(y_i)\prod_{i=q+1}^{n}\psi_0(y_i),$$

where $0<\delta<s$. Note that no mixing is performed in this case. Instead, we fix the sequence $\mu=(s,\dots,s,0,\dots,0)$ in both f and g, and perturb the nonzero entries of θ by a small amount δ in g. This set of priors provides the sharpest rate for the case when the signal is strong, i.e., $s=n^b$ is large. The intuition is that when s is large, estimation of $Q(\mu,\theta)$ is most difficult due to the indistinguishability between $\theta_i=s$ and $\theta_i=s-\delta$, where $\delta\approx 0$.

The chi-square affinity between f and g is given by

$$\int\frac{g^2}{f}=e^{q\delta^2/\sigma^2}.$$

Let $\delta=\sigma/\sqrt{q}=\sigma n^{-\epsilon/2}$. Then we have

$$\int\frac{g^2}{f}=e<\infty.$$

Since $Q(\mu,\theta)=(1/n)qs^4$ under f and $Q(\mu,\theta)=(1/n)qs^2(s-\delta)^2$ under g, it follows from CRI that

$$\begin{aligned}R^*(n,\Omega(\beta,\epsilon,b))&\geq c\bigg(\frac{1}{n}qs^2\big(s^2-(s-\delta)^2\big)\bigg)^2\\&=c\bigg(\frac{1}{n}\sqrt{q}s^3\bigg)^2(1+o(1))=cn^{\epsilon+6b-2}(1+o(1)).\end{aligned}$$

Supplementary Materials

Supplement S1 contains the estimation results for $Q(\mu,\theta)$ when μ and θ have different signal strengths, while supplement S2 contains the proofs for Theorems 3 and 4.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, Volume 1117 of *Lecture Notes in Math.*, pp. 1–198. Springer, Berlin.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli 8*(5), 577–606.

Baraud, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist. 32*(2), 528–551.

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A 50*(3), 381–393.

Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist. 24*(6), 2524–2535.

Cai, T. T. and Low, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist. 32*(2), 552–576.

Cai, T. T. and Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist. 33*(6), 2930–2956.

Cai, T. T. and Low, M. G. (2006a). Adaptive confidence balls. *Ann. Statist. 34*(1), 202–228.

Cai, T. T. and Low, M. G. (2006b). Optimal adaptive estimation of a quadratic functional. *Ann. Statist. 34*(5), 2298–2325.

Collier, O., Comminges, L. and Tsybakov, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv preprint arXiv:1502.00665*.

Consortium, S. P. G.-W. A. S. G. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics 43*(10), 969–976.

Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., Behrens, T., Cho, J. et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics 7*(8), e1002254.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*(3), 962–994.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity 6*(3), 290–323.

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist. 26*(1), 288–314.

Efromovich, S. and Low, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist. 24*(3), 1106–1125.

Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist. 19*(3), 1273–1294.

Genovese, C. R. and Wasserman, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist. 33*(2), 698–729.

Giné, E. and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 47–61.

Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, Volume 169 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

Lepski, O. V. and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli 5*(2), 333–358.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist. 17*(3), 1001–1008.

Li, Y. R., Zhao, S. D., Li, J., Bradfield, J. P., Mohebnasab, M., Steel, L., Kobie, J., Abrams,

D. J., Mentch, F. D., Glessner, J. T. et al. (2015). Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nature communications 6*.

Rankinen, T., Sarzynski, M. A., Ghosh, S. and Bouchard, C. (2015). Are there genetic paths common to obesity, cardiovascular disease outcomes, and cardiovascular risk factors? *Circulation research 116*(5), 909–922.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F. and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics 89*(5), 607–618.

Zhao, S. D., Cai, T. T. and Li, H. (2014). Gene-disease associations via sparse simultaneous signal detection. *Technical Report*.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: tcai@wharton.upenn.edu

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: xtan@wharton.upenn.edu

(Received February 2016; accepted May 2016)