INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Compressed Smooth Sparse Decomposition

Shancong Mou, Jianjun Shi

To cite this article:
Shancong Mou, Jianjun Shi (2023) Compressed Smooth Sparse Decomposition. INFORMS Journal on Data Science 2(1):60-80. https://doi.org/10.1287/ijds.2022.0023

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 2, No. 1, April–June 2023, pp. 60–80
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Compressed Smooth Sparse Decomposition

Shancong Mou,[a] Jianjun Shi[a,*]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
*Corresponding author
Contact: shancong.mou@gatech.edu (SM); jshi33@isye.gatech.edu, https://orcid.org/0000-0002-3774-9176 (JS)

Received: January 11, 2022
Revised: June 20, 2022
Accepted: September 2, 2022
Published Online in Articles in Advance: November 7, 2022

https://doi.org/10.1287/ijds.2022.0023

Abstract. Image-based anomaly detection systems are of vital importance in various manufacturing applications. The resolution and acquisition rate of such systems are increasing significant in recent years under the fast development of image sensing technology. This enables the detection of tiny anomalies in real time. However, such a high resolution and a high acquisition rate of image data not only slow down the speed of image processing algorithms but also, increase data storage and transmission cost. To tackle this problem, we propose a fast and data-efficient method with theoretical performance guarantee that is suitable for sparse anomaly detection in images with a smooth background (smooth plus sparse signal). The proposed method, named compressed smooth sparse decomposition (CSSD), is a one-step method that unifies the compressive image acquisition- and decomposition-based image processing techniques. To further enhance its performance in a high-dimensional scenario, a Kronecker compressed smooth sparse decomposition (KronCSSD) method is proposed. Compared with traditional smooth and sparse decomposition algorithms, significant transmission cost reduction and computational speed boost can be achieved with negligible performance loss. Simulation examples and several case studies in various applications illustrate the effectiveness of the proposed framework.

History: Kwok-Leung Tsui served as the senior editor for this article.
Funding: This work is partially support by the National Science Foundation Division of Engineering Education and Centers [Grant 2052714].
Data Ethics & Reproducibility Note: The code capsule is available on Code Ocean at https://doi.org/10.24433/CO.6352310.v2 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2022.0023).

Keywords: anomaly detection • compressive sensing • image processing • smooth sparse decomposition

1. Introduction

High-quality image sensing systems are widely used in manufacturing processes for product quality monitoring and fault diagnosis. The resolution and acquisition rate of such systems increase significantly benefiting from the rapid development of image sensing technology. For example, in a hot rolling process, an in situ image-based sensor can detect a micrometer-sized seam on a rolling bar at a speed of up to 225 miles per hour (Yan et al. 2018). For another example, to monitor solar activity, satellites can capture high-resolution solar images with a high acquisition rate, producing terabytes of data per day (Wang et al. 2018). To achieve real-time inspection, a large volume of high-resolution images needs to be transmitted and processed in real time. Such a large volume of high-resolution image data poses a big challenge not only on the speed of image processing algorithms but also, for the storage and transmission of the data itself.

Matrix decomposition-based image processing techniques are widely used in image-based process monitoring and anomaly detection. They achieve the goal by integrating the prior for background and anomaly components into the optimization problem. In terms of utilizing the low-rank and sparse property, robust principal component analysis was first proposed by Candès et al. (2011) to decompose a data matrix into low-rank and element-wise sparse components. One of its famous applications is dynamic foreground and static background separation (Bouwmans and Zahzah 2014). Following this approach, numerous algorithm variants have been proposed, including outlier pursuit (Xu et al. 2012), which aims to decompose the data matrix into a low-rank component and a column-wise sparse component, and low-rank plus compressed sparse decomposition (Mardani et al. 2013), which aims to decompose the data matrix into low-rank and compressed sparse components and so on. For utilizing smooth and sparse properties, smooth and sparse decomposition (SSD) methods (Minaee et al. 2015, Yan et al. 2017) are proposed for anomaly detection in images with smooth backgrounds. Following this approach, several explorations have been conducted, including spatiotemporal smooth sparse decomposition (ST-SSD) (Yan et al. 2018), additive tensor decomposition (Mou et al. 2021), and so on. By adopting

the matrix decomposition approach, both the background and anomaly can be captured without detection time delay. However, because of the requirement of storage, transmission, and processing of the whole image signal, it cannot be applied in the scenario with low-transmission bandwidth but high processing speed requirements: for example, the solar flare detection application (Augusto et al. 2011).

To mitigate the data storage and transmission burden and improve sensing efficiency, compressive sensing (CS) (Candes et al. 2006) has been proposed, in which the data are directly collected in a compressed form and then, reconstructed accurately with high probability. More specifically, suppose that the original signal is a *sparse* vector $\boldsymbol{y} \in \mathbb{R}^n$, and the main idea is to store and transmit a small set of compressive measurements $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} \in \mathbb{R}^p$, where $\boldsymbol{A} \in \mathbb{R}^{p \times n}$ is an underdetermined sensing matrix ($p \ll n$) satisfying specific properties. Then, the original signal can be reconstructed from its compressed form $\boldsymbol{y}'$, on which assorted image processing algorithms can be applied for defect detection and so on. For a comprehensive review of CS, please refer to Marques et al. (2018) and Rani et al. (2018). Even though promising, the naïve approach that tries to first reconstruct the image from the compressed measurement and then, apply matrix decomposition algorithms for anomaly detection has two issues.

i. For smooth plus sparse signals, the existence of such a sensing matrix $\boldsymbol{A}$ satisfying specific properties is unknown.

ii. The reconstruction process is usually computationally intensive (Marques et al. 2018), which on the other hand, slows down the overall computational speed of the defect detection algorithms.

Recently, to integrate the CS with matrix decomposition algorithms, Waters et al. (2011) proposed an SpaRCS method to recover low-rank and sparse matrices directly from compressive measurements, and Tanner and Vary (2020) gave a rigorous performance discussion. However, those methods do not consider the smooth plus sparse decomposition problems and are not efficient in dealing with high-order data.

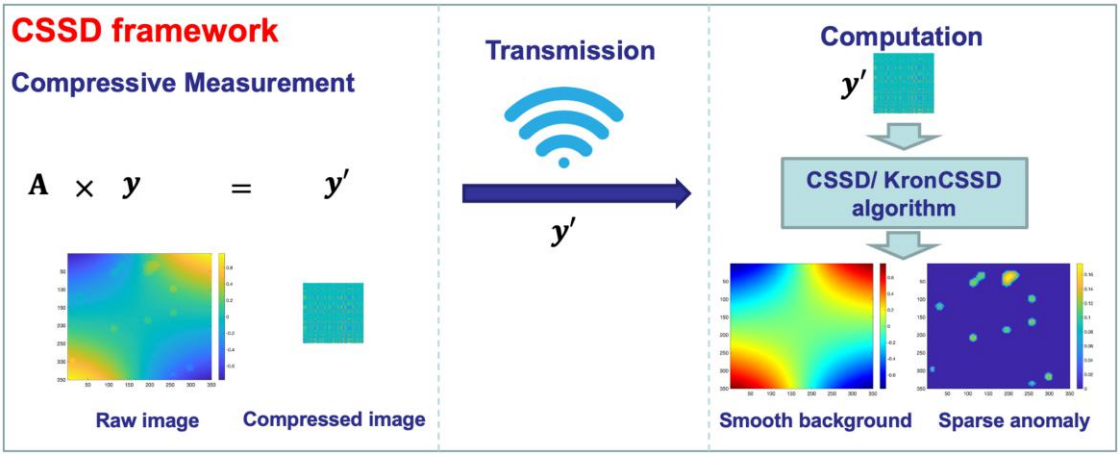
In this paper, we discuss the possibility of adopting compressive data acquisition systems for image-based quality monitoring and fault detection in applications where the background is smooth, and anomalies are sparse. To achieve so, we propose a compressed smooth sparse decomposition (CSSD) framework. In this framework, the signal processing algorithms are directly applied to the compressed data, and no reconstruction step is needed. By doing so, a significant cost reduction in sensing, storage, and transmission as well as a boost in the speed of image processing algorithms can be achieved with negligible performance loss. We also established the theoretical foundation of adopting such a compressive data acquisition system for smooth plus sparse signals as well as the performance guarantee of the proposed algorithm. To further improve its performance in high-order scenarios, a Kronecker compressed smooth sparse sensing (KronCSSD) is proposed.

The remainder of this paper is organized as follows. In Section 2, we present the CSSD framework. In Section 3, we use simulation studies to validate the proposed framework. In Section 4, we demonstrate the proposed framework using several case studies. Finally, Section 5 concludes the paper.

2. Compressed Smooth Sparse Decomposition Framework

In this section, we will present the proposed CSSD framework. As mentioned in Section 1, we aim to design a fast and data-efficient method for sparse anomaly (sparse signal component) detection in signals with a smooth background (smooth signal component). For simplicity, we first discuss the methodology for one-dimensional (1D) signals and then, generalize it to n-dimensional images. Figure 1 provides an overview of the proposed methodology. There are three stages in the proposed methodology. (i) The signal is acquired in its compressed form through compressive measurement. (ii) Then, the compressed data are transmitted to

Figure 1. (Color online) An Overview of the CSSD/KronCSSD Framework



the server. (iii) Finally, a decomposition algorithm will be applied directly to the compressed signal to decompose it into its corresponding smooth and sparse signal components.

Mathematically, let $\boldsymbol{y} \in \mathbb{R}^n$ be a smooth plus sparse signal (which will be defined formally in Section 2.1). We aim to store and transmit a small set of compressive measurements $\boldsymbol{y}'$ (i.e., $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$) and then, reconstruct the smooth and sparse signal components from $\boldsymbol{y}'$. To achieve so, there are three questions to be addressed.

i. What is a smooth plus sparse signal?

ii. How do we compress such a signal?

iii. How do we reconstruct such a signal from compressive measurements?

The remainder of this section is organized as follows to answer those three questions. We start with a formal definition of 1D smooth plus sparse signals in Section 2.1. Based on that, we introduce the compressive measurement method and discuss its theoretical properties for such signals in Section 2.2. In Section 2.3, we present the proposed CSSD framework that can directly reconstruct the smooth and sparse signal components from the compressive measurement by solving the following optimization problem:

$$\min_{\boldsymbol{\theta}, \boldsymbol{\theta}_a} \|\boldsymbol{\theta}_a\|_1$$
$$\text{s.t.} \|\boldsymbol{A}(\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a) - \boldsymbol{y}'\|_2 \le \epsilon_1, \tag{1}$$

where $\boldsymbol{B}$ and $\boldsymbol{B}_a$ are bases, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ are corresponding coefficients, and ϵ_1 is the bound for measurement error. The reconstruction accuracy is also characterized theoretically. Then, we generalize the CSSD algorithm to n dimensions using Kronecker compressive sensing (KCS) (Duarte and Baraniuk 2011) and propose a KronCSSD formulation in Section 2.4. Finally, in Section 2.5, we present the advantage of the proposed framework and give the strategy of selecting the compressive ratio, tuning parameters, and bases in practice.

2.1. The Set of Smooth Plus Sparse Signals

In this section, we define the set of smooth plus sparse signals mathematically. The smooth signals originate from the spline smoothing (De Boor and De Boor 1978, Eilers and Marx 1996), where the raw signal is approximated by a linear combination of a set of spline basis functions for smooth interpolation and denoising, and a spline regression technique is usually utilized. To improve the regression robustness with respect to outliers, outliers are explicitly accounted for in the regression model as a sparse component of the raw signal (Giannakis et al. 2011, Mateos and Giannakis 2011). This idea was further generalized to incorporate the special structure of outliers (Yan et al. 2017, 2018).

As a summary, a 1D signal $\boldsymbol{y} \in \mathbb{R}^n$ is defined as a smooth plus sparse signal if it can be decomposed into two signal components: (i) a smooth signal $\boldsymbol{m} \in \mathbb{R}^n$ in a low-dimensional subspace spanned by a set of smooth bases (i.e., $\boldsymbol{m} = \boldsymbol{B}\boldsymbol{\theta}$, where $\boldsymbol{B} \in \mathbb{R}^{n\times r}$ is a basis matrix with $r \ll n$) and (ii) a sparse signal $\boldsymbol{a} \in \mathbb{R}^n$ in a relatively high-dimensional subspace spanned by a set of predefined bases, of which the coefficients admit sparse property (i.e., $\boldsymbol{a} = \boldsymbol{B}_a\boldsymbol{\theta}_a$, where $\boldsymbol{B}_a \in \mathbb{R}^{n\times q}$ is a basis matrix with $q \le n$ and $\boldsymbol{\theta}_a \in \mathbb{R}^q$ is an s-sparse vector; i.e., $\|\boldsymbol{\theta}_a\|_0 \le s$). Given a smooth plus sparse signal $\boldsymbol{y}$, we define the aforementioned decomposition as SSD (i.e., $\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{a}$).

To ensure the uniqueness of SSD in a nontrivial case when $n \le r + q$, the following definition is introduced.

Definition 2.1. The local support property of $\boldsymbol{B}_a$, $\boldsymbol{B}_a \in \mathbb{R}^{n\times q}$. $\boldsymbol{B}_a$ only has local support such that each column of $\boldsymbol{B}_a$ only has nonzero values inside a specific interval. The length of this interval is defined as $l(\boldsymbol{B}_a)$.

Notice that $l(\boldsymbol{B}_a) \in \{1, \ldots, n\}$. The local support property with a small l ensures the sparsity of $\boldsymbol{a}$. For example, the B-spline basis has a local support property (Unser 1999).

Definition 2.2. The incoherence condition of $\boldsymbol{B}$, $\boldsymbol{B} \in \mathbb{R}^{n\times r}$. Let $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the reduced singular value decomposition (SVD) of $\boldsymbol{B}$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{r\times r}$. Its incoherence condition parameter $\mu(\boldsymbol{B})$ is defined as the smallest value such that

$$\max_{i\in\{1,..,r\}} \|\boldsymbol{U}^T\boldsymbol{e}_i\|_2 \le \sqrt{\frac{\mu(\boldsymbol{B})r}{n}},$$

where $\boldsymbol{e}_i$ is the i th standard basis vector in $\mathbb{R}^n$.

Notice that $\mu(B) \in [1, \sqrt{n/r}]$. The incoherent condition with a small μ ensures that the $\boldsymbol{m}$ is not sparse (Candès et al. 2011).

The following theorem ensures the uniqueness of the SSD decomposition.

Theorem 2.1. *If $\mu(\boldsymbol{B}) < n(2rsl)^{-1}$, then the SSD decomposition is unique with respect to $\boldsymbol{m}$ and $\boldsymbol{a}$.*

Theorem 2.1 gives the condition that the smooth plus sparse signal can be uniquely decomposed into a smooth part and a sparse part. The proof of Theorem 2.1 is in Appendix A.

Formally, we define the set of smooth plus sparse signals as follows.

Definition 2.3. The set of smooth and sparse signals is defined as $MS_{r,s,\mu,l}$:

$$MS_{r,s,\mu,l} = \{\boldsymbol{y} \in \mathbb{R}^n \,|\, \boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a,\ \boldsymbol{B}_a \in \mathbb{R}^{n\times q},$$
$$l(\boldsymbol{B}_a) = l,\ \boldsymbol{\theta}_a \in \mathbb{R}^q, \|\boldsymbol{\theta}_a\|_0 \le s,\ \boldsymbol{B} \in \mathbb{R}^{n\times r},$$
$$\mu(\boldsymbol{B}) = \mu < n(2rsl)^{-1}, \boldsymbol{\theta} \in \mathbb{R}^r\}.$$

For such a smooth plus sparse signal, how to compress it while ensuring the reconstruction performance will be discussed in the next section.

2.2. Compressive Sensing for Smooth Plus Sparse Signals

As mentioned in Section 1, to ensure the reconstruction performance, the sensing matrix $\boldsymbol{A}$ has to satisfy the so-called restricted isometry property (RIP) (Candes 2008). It has been proven that a random matrix can satisfy the RIP property for the sparse signal (Candes 2008), the low-rank signal (Recht et al. 2010), and the rank plus sparse signal (Tanner and Vary 2020). However, the existence of such a matrix for the smooth plus sparse signal, which is the foundation of adopting compressed data acquisition techniques in applications with smooth background and sparse anomalies, is still unknown. In this section, we will discuss the existence of such a sensing matrix. Before stating the result, we will first present the relevant definitions that are necessary to derive the main result.

Definition 2.4. RIP for $MS_{r,s,\mu,l}$. Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be a linear measurement matrix. For every quadruple (r,s,μ,l), define the restricted isometry constant (RIC) $\delta_{r,s,\mu,l}$ to be the smallest positive constant such that

$$(1-\delta_{r,s,\mu,l})\|\boldsymbol{y}\|_2 \le \|\boldsymbol{A}\boldsymbol{y}\|_2 \le (1+\delta_{r,s,\mu,l})\|\boldsymbol{y}\|_2, \quad \forall \boldsymbol{y} \in MS_{r,s,\mu,l}.$$

If such a $\delta_{r,s,\mu,l} \in (0,1)$ exists, we say that $\boldsymbol{A}$ satisfies the RIP.

Theorem 2.2. *Suppose that $\delta_{r,2s,\mu,l} < 1$ for some integer $r,s,l \ge 1$ and positive numbers $\mu < n(2rsl)^{-1}$; then, there is a $\boldsymbol{y}_0$ in the set $MS_{r,s,\mu,l}$, which is the only solution for $\boldsymbol{A}\boldsymbol{y}_0 = \boldsymbol{b}$.*

Theorem 2.2 guarantees the uniqueness of the smooth plus sparse signal that satisfies the sensing equation when $\boldsymbol{A}$ satisfies the RIP. The proof of Theorem 2.2 is in Appendix B.

Next, we prove that for the set of smooth plus sparse signals, $MS_{r,s,\mu,l}$, there exists such a matrix $\boldsymbol{A}$ satisfying the RIP property with RIC = $\delta_{r,s,\mu,l}$ with high probability.

Notice that the RIP for a matrix is difficult to verify. A suitable set of random matrices that obey the RIP for the set of sparse vectors with high probability (Recht et al. 2010, Tanner and Vary 2020) is defined as follows.

Definition 2.5. Nearly isometric matrices (Baraniuk et al. 2008). Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be a random variable that takes values in linear maps from $\mathbb{R}^n$ to $\mathbb{R}^p$; then, for any $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{A}$ is nearly isometrically distributed if

i. $\mathbb{E}[\|\boldsymbol{A}\boldsymbol{y}\|_2^2] = \|\boldsymbol{y}\|_2^2$ and

ii. $Pr(|\|\boldsymbol{A}\boldsymbol{y}\|_2^2 - \|\boldsymbol{y}\|_2^2| \ge \epsilon\|\boldsymbol{y}\|_2^2) \le 2e^{-pc_0(\epsilon)},\ 0 < \epsilon < 1,$

where $c_0(\epsilon)$ is a constant that only depends on ϵ.

The $p \times n$ matrix with independent, identically distributed (i.i.d.) Gaussian entries satisfies those two properties (Baraniuk et al. 2008) $\left(\text{i.e., } \boldsymbol{A}_{ij} \sim \mathcal{N}\left(0, \frac{1}{p}\right),\right.$ $\left.\text{with } c_0(\epsilon) = \epsilon^2/4 - \epsilon^3/6\right)$. There are also other distributions satisfying the nearly isometric property, such as the $p \times n$ matrix with i.i.d. Bernoulli entries and their related distribution (Baraniuk et al. 2008).

Then, the following theorem states that the nearly isometric matrices can also serve as the sensing matrix for smooth plus sparse signals and gives the magnitude of the number of linear measurements.

Theorem 2.3. *Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be a matrix from the families described in Definition 2.5. Furthermore, assume that $\mu < n(2rsl)^{-1}$ and the basis matrix $\boldsymbol{B}_a$ for the sparse signal component satisfies the RIP with RIC $\delta_{\boldsymbol{B}_a,s} \in (0,1)$: that is, $\delta_{\boldsymbol{B}_a,s}$ to be the smallest positive constant such that*

$$(1-\delta_{\boldsymbol{B}_a,s})\|\boldsymbol{\theta}_a\|_2 \le \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2 \le (1+\delta_{\boldsymbol{B}_a,s})\|\boldsymbol{\theta}_a\|_2,$$
$$\forall \boldsymbol{\theta}_a \in \{\boldsymbol{\theta}_a \in \mathbb{R}^q, \|\boldsymbol{\theta}_a\|_0 \le s\}.$$

For a given $\delta \in (0,1)$, there exists constants $c_1, c_2 > 0$ depending only on δ, such that the RIC for $MS_{r,s,\mu,l}$ is upper bounded by δ, with the probability of at least $1 - \exp(-c_1 p)$, whenever

$$p \ge c_2\left(\ln 2 + r\ln\frac{24}{\delta}\tau_1 + s\left(1 + \ln\frac{24}{\delta}\tau_0 + \ln\frac{n}{s}\right)\right),$$

where $\eta = \sqrt{\frac{\mu rsl}{n}}$, $\tau_0 = \frac{1}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})(1-\eta^2)}}$, $\tau_1 = \|\boldsymbol{B}^\dagger\|_2\left(1 + \frac{1}{\sqrt{1-\eta^2}}\right)$, *and* $\boldsymbol{B}^\dagger = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$.

Theorem 2.3 states that if the nearly isometric matrix is selected as the sensing matrix, the RIC for $MS_{r,s,\mu,l}$ is upper bounded with high probability. In practice, $\boldsymbol{B}$ and $\boldsymbol{B}_a$ are prespecified based on the understanding/engineering knowledge of the process (please refer to Section 2.5.4 for more detail). Theorem 2.3 also provides a guidance on the magnitude of linear measurements p, which determines the compressive measurement matrix $\boldsymbol{A}$.

The proof of Theorem 2.3 is in Appendix C.

In this section, we answered two fundamental questions. (i) How do we compress the smooth plus sparse signal (design the compressive measurement matrix $\boldsymbol{A}$)?

(ii) How many linear measurements are needed to preserve the information in smooth plus sparse signals with high probability?

In the next section, we will discuss the problem formulation to recover the smooth and sparse signal components simultaneously from the compressed signal using the CSSD framework.

2.3. Compressed Smooth Sparse Decomposition

As mentioned in Section 1, one way to recover the smooth component and sparse component is first reconstructing the compressed image and then, using the SSD algorithm. However, this will slow down the speed of the defect detection algorithm. Instead, we propose to solve the one-step convex relaxation Problem (1). Natural questions are if we can recover the smooth and sparse signals from the compressed measurement $\boldsymbol{y}'$ by solving Problem (1) and what the accuracy is. The following theorem guarantees the recovery performance of the proposed convex relaxation Problem (1).

Theorem 2.4. Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ *be a matrix from the families described in Definition 2.5. Let the signal* $\boldsymbol{y} = \boldsymbol{m}_0 + \boldsymbol{a}_0 \in MS_{r,s,\mu,l}$, *where* $\boldsymbol{m}_0 = \boldsymbol{B}\boldsymbol{\theta}_0$ *and* $\boldsymbol{a}_0 = \boldsymbol{B}_a\boldsymbol{\theta}_{a0}$. *Assume that Problem* (1) *is feasible, and let the optimal solution be* $\boldsymbol{m}^* = \boldsymbol{B}\boldsymbol{\theta}^*$ *and* $\boldsymbol{a}^* = \boldsymbol{B}_a\boldsymbol{\theta}_a^*$. *Assume that the basis matrix* B_a *for sparse signal component satisfies the RIP with RIC* $\delta_{B_a,2s} \in (0,1)$. *Let* $a = (1+\alpha_1+\alpha_2)\gamma^2 + 2\alpha_2 + 2$ *and* $c = 1 - \gamma^2\alpha_1\alpha_2 - {\alpha_2}^2$, *where* $\alpha_1 = \frac{\eta}{1-\eta^2}$, $\alpha_2 = \frac{\sqrt{2}\eta}{1-2\eta^2}$, $\gamma = \sqrt{\frac{1+\delta_{B_a,2s}}{1-\delta_{B_a,2s}}}$, *and* $\eta = \sqrt{\frac{\mu rsl}{n}}$. *Suppose that* $r,s,l \in \mathbb{N}$ *and* $\mu < n(2rsl)^{-1}$, *such that* $c > 0$ *and* $\delta_{r,3s,\mu,l} \in (0, c/a)$; *then,*

$$\|\boldsymbol{a}_0 - \boldsymbol{a}^*\|_2 = \|\boldsymbol{B}_a\boldsymbol{\theta}_{a0} - \boldsymbol{B}_a\boldsymbol{\theta}_a^*\|_2 \le C_a\epsilon_1$$

and

$$\|\boldsymbol{m}_0 - \boldsymbol{m}^*\|_2 = \|\boldsymbol{B}\boldsymbol{\theta}_0 - \boldsymbol{B}\boldsymbol{\theta}^*\|_2 \le C_m\epsilon_1,$$

where

$$C_a = \frac{(1+\gamma^2)(1+\alpha_2)\sqrt{1+\delta_{r,3s,\mu,l}}}{c - a\delta_{r,3s,\mu,l}}$$

and

$$C_m = \frac{\sqrt{1+\delta_{r,3s,\mu,l}} + \left(\delta_{r,3s,\mu,l} + \frac{\gamma^2}{1+\gamma^2}\alpha_1 + \frac{1}{1+\gamma^2}\alpha_2\right)C_a}{\left(1-\delta_{r,3s,\mu,l}\right)}.$$

Theorem 2.4 gives the conditions that the proposed convex relaxation Problem (1) can recover the true smooth and sparse signal up to a constant times the noise bound.

The proof of Theorem 2.4 is in Appendix D.

Notice that one advantage of the proposed CSSD framework is that it is compatible with existing decomposition algorithms. For example, for the SSD algorithm proposed by Yan et al. (2017), the problem formulation becomes

$$\min_{\theta,\theta_a} \|\boldsymbol{y}' - \boldsymbol{A}(\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a)\|_2^2 + \lambda\|\boldsymbol{\theta}_a\|_1, \tag{2}$$

which can be solved efficiently by using the algorithm proposed by Yan et al. (2017).

2.4. Kronecker Compressed Smooth Sparse Decomposition

In the previous section, we discussed the CSSD framework for the 1D signal. In this section, we will generalize the proposed CSSD to KronCSSD framework for high-order tensor data.

Let $\mathcal{Y} \in \mathbb{R}^{n_1\times\cdots\times n_d}$ be the original signal and $\boldsymbol{y} \in \mathbb{R}^N$ be its corresponding vectorized signal (i.e., $\boldsymbol{y} = \mathrm{vec}(\mathcal{Y})$ and $N = \prod_{i=1}^d n_i$). Let $\boldsymbol{A} \in \mathbb{R}^{p\times N}$ be a measurement matrix satisfying the RIP. Let $\boldsymbol{y}' \in \mathbb{R}^p$ be the compressed data, such that $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$. $\boldsymbol{B} = \otimes_{i=1}^d \boldsymbol{B}_i$ and $\boldsymbol{B}_a = \otimes_{i=1}^d \boldsymbol{B}_{ai}$ are the known bases for smooth and sparse components, respectively. Notice that problem formulation (1) can still be used for recovering the high-order smooth and sparse signal components from the compressive measurement. However, there are two issues. (i) In practice, the global CS measurements matrix $\boldsymbol{A}$ is hard to realize using the CS device (Duarte and Baraniuk 2011). (ii) The resulting bases $\boldsymbol{B}$ and $\boldsymbol{B}_a$ can be extremely large as the dimension of the data increases, which will not only cause a big challenge in the storage of such a large matrix but also, result in the computational issue in handling such large matrices.

In reality, the high-dimensional tensor data usually have low-rank properties along each mode, which have been extensively exploited in tensor low-rank modeling techniques, such as CANDECOMP/PARAFAC (CP)/Tucker decompositions (Kolda and Bader 2009). This makes it possible for designing a sensing matrix for each mode, which is called Kronecker CS (Duarte and Baraniuk 2011). Inspired by the KCS method, we propose the KronCSSD framework formulation as follows.

Let $\boldsymbol{A}_i \in \mathbb{R}^{p_i\times n_i}$ be a measurement matrix with RIP for each mode of the tensor; then, we have the following formulation,

$$\begin{aligned}
&\min_{\boldsymbol{\Theta},\boldsymbol{\Theta}_a} \|\mathrm{vec}(\boldsymbol{\Theta}_a)\|_1 \\
&\text{s.t.} \|\mathrm{vec}(\mathcal{Y}_1 - \mathcal{Y}')\|_2 \le \epsilon_1, \\
&\quad \mathcal{Y}_1 = \boldsymbol{\Theta}\times_1(\boldsymbol{A}_1\boldsymbol{B}_1)\times_2\cdots\times_d(\boldsymbol{A}_d\boldsymbol{B}_d) \\
&\quad + \boldsymbol{\Theta}_a\times_1(\boldsymbol{A}_1\boldsymbol{B}_{a1})\times_2\cdots\times_d(\boldsymbol{A}_d\boldsymbol{B}_{ad}),
\end{aligned} \tag{3}$$

where $\mathcal{Y}' = \mathcal{Y}\times_1\boldsymbol{A}_1\times_2\cdots\times_d\boldsymbol{A}_d$ is the compressive measurement. $\boldsymbol{\Theta} \in \mathbb{R}^{r_1\times\cdots\times r_d}$ and $\boldsymbol{\Theta}_a \in \mathbb{R}^{q_1\times\cdots\times q_d}$ are the basis coefficients for the smooth and sparse signal components, respectively.

The proposed KronCSSD framework is a nontrivial generalization of the CSSD framework for 1D signals. Its performance will be shown empirically in simulation and case studies. The theoretical discussion of the KronSSD framework is left for future work.

For example, for a two-dimensional (2D) image $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$, when adopting the SSD algorithm (Yan et al. 2017), the problem formulation becomes

$$\min_{\boldsymbol{\Theta}, \boldsymbol{\Theta}_a} \|\boldsymbol{Y}' - \boldsymbol{A}_1\boldsymbol{B}_1\boldsymbol{\Theta}\boldsymbol{B}_2^T\boldsymbol{A}_2^T - \boldsymbol{A}_1\boldsymbol{B}_{a1}\boldsymbol{\Theta}_a\boldsymbol{B}_{a2}^T\boldsymbol{A}_2^T\|_2^2 + \lambda\|\mathrm{vec}(\boldsymbol{\Theta}_a)\|_1, \tag{4}$$

where $\boldsymbol{Y}'$ is the compressed image (i.e., $\boldsymbol{Y}' = \boldsymbol{A}_1\boldsymbol{Y}\boldsymbol{A}_2^T$). It can be solved efficiently by using the algorithm proposed by Yan et al. (2017).

2.5. Discussion

2.5.1. Advantages of the Proposed CSSD/KronCSSD Methods. In this section, we will give a brief discussion about the advantages of the proposed CSSD/KronCSSD methods. We define the compressive ratio as $c = \prod_{i=1}^{d} p_i/n_i$. The smaller the compressive ratio, the fewer data will be transmitted. The proposed CSSD/KronCSSD methods have the following characteristics.

i. We propose to directly acquire the compressed image, which not only reduces the sensing cost but also, reduces the data transmission and storage cost by $\prod_{i=1}^{d} p_i/n_i$ times.

ii. The smooth and sparse signal components can be recovered by solving a smaller-scale convex optimization problem with input data $\prod_{i=1}^{d} p_i/n_i$ times smaller than that of the original problem, which significantly boosts the computation.

2.5.2. Compressive Ratio Selection in Practice. Theorems 2.3 and 2.4 show that the 1D smooth plus sparse signal can be recovered with high probability from a compressive measurement if the compressive ratio is above a specific threshold. However, there are several parameters (such as c_1, c_2) that are difficult to obtain when calculating the threshold in practice. We propose a practical procedure of selecting the compressive ratio utilizing the historical data. Suppose some training signals with their real background and anomalies are available; the compressive ratio can be selected with the following guidelines.

i. If there is a requirement for reconstruction accuracy for smooth and sparse signal components, then $\hat{p}$ is chosen as the smallest value that satisfies such a requirement.

ii. If there is no such requirement, we recommend choosing the compressive ratio corresponding to the sharp change point of the slope of the loss function-compressive ratio curve. This point exists because of the existence of such a threshold, after which the reconstruction with high probability is guaranteed by Theorems 2.3 and 2.4. We will demonstrate this in the simulation study in Section 3.1.

For high-order tensor data, the selection of the p_i for each mode can be challenging. We provide some empirical guidelines as follows.

i. If the smoothness of the background is similar along different modes, a unified compressive ratio is recommended.

ii. If the smoothness of the background is different, more sensing budget should be allocated to the mode along which the background is less smooth.

iii. We recommend fixing the ratio among p_i along different modes and adopting steps (i) and (ii) for 1D signal to determine the compressive ratio.

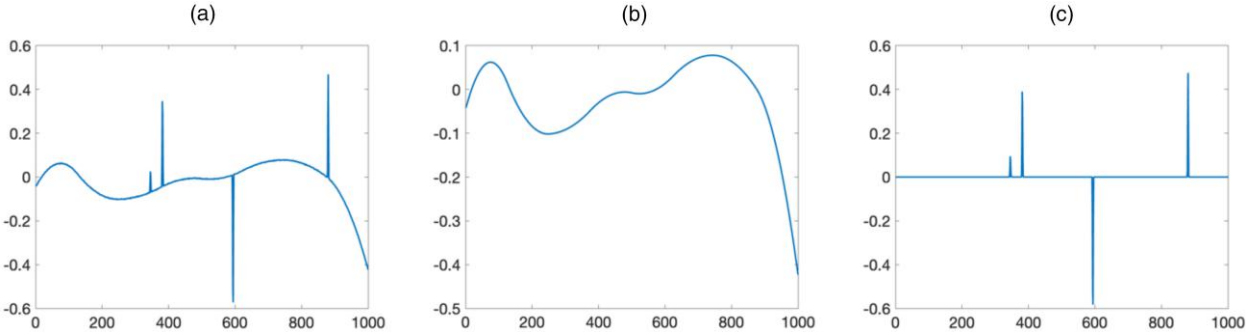
2.5.3. Tuning Parameter Selection in Practice. Notice that in problem formulations (1) and (3), there is a hyperparameter ϵ_1, indicating the bound for measurement noise. If the measurement error bound is known from the accuracy of the measurement device, it can be directly used here. Otherwise, a crossvalidation step using historical data is recommended. Similarly, there is a tuning parameter λ in their corresponding Lagrangian form Equations (2) and (4), which controls the sparsity of the decomposed anomaly. Crossvalidation can be used to determine this parameter. For more detail, please refer to Yan et al. (2017).

2.5.4. Bases Selection in Practice. The bases $\boldsymbol{B}$ and $\boldsymbol{B}_a$ are prespecified based on the understanding/engineering knowledge of the process. The selection of such bases is discussed in detail in section 3.4 of Yan et al. (2017). In general, any smooth basis, such as splines or kernels, can be used for the background. For sparse anomalies, such as small regions scattered over the background or in the form of thin lines, an identity basis is recommended. Linear (quadratic) B splines are recommended for anomalous regions with sharp corners (curved boundaries). However, to ensure the uniqueness of the smooth sparse decomposition, we do require the bases $\boldsymbol{B}$ and $\boldsymbol{B}_a$ to satisfy specific properties, such as those mentioned in Definitions 2.1 and 2.2 and Theorem 2.1, which should be checked for selected bases. We will demonstrate this in the simulation study in Section 3.1.

3. Simulation Study

In this section, we will demonstrate the proposed CSSD and KronCSSD framework with simulation studies. First, the CSSD method is applied on 1D signals in Section 3.1, and then, we apply the KronCSSD method on 2D images in Section 3.2.

Figure 2. (Color online) A Sample Raw Signal and Its Smooth and Sparse Components



Notes. (a) Raw signal. (b) Smooth signal component $\boldsymbol{m}$. (c) Sparse signal component $\boldsymbol{a}$.

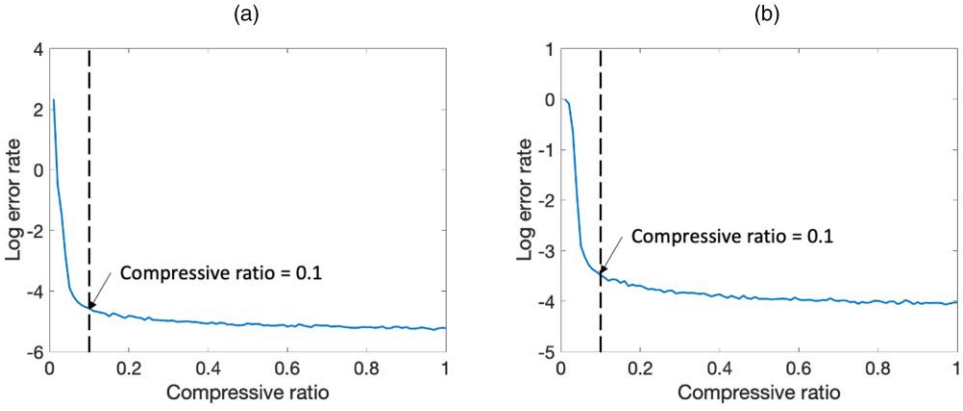
3.1. CSSD on a 1D Signal

A 1D signal $(\boldsymbol{y} \in \mathbb{R}^n)$ is assumed to be a superposition of the smooth signal component ($\boldsymbol{m} = \boldsymbol{B}\boldsymbol{\theta}$), the sparse signal component ($\boldsymbol{a} = \boldsymbol{B}_a\boldsymbol{\theta}_a$), and noise $\boldsymbol{e}$ (i.e., $\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{a} + \boldsymbol{e}$). To demonstrate the proposed CSSD framework, we will first conduct compressive data acquisition on the raw signal (i.e., $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$). Then, the data reconstruction and decomposition are achieved in one step. The performance is evaluated by the relative error between the true signal $\boldsymbol{a}$ and the reconstructed one $\hat{\boldsymbol{a}}$ (i.e., $\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|_2/\|\boldsymbol{a}\|_2$ and $\|\boldsymbol{m} - \hat{\boldsymbol{m}}\|_2/\|\boldsymbol{m}\|_2$ for the smooth background).

In the simulation study, we generate a 1D signal from $MS_{r,s,\mu,l}$ and let $n = 1{,}000$. The smooth background is generated from a random linear combination of B-spline bases with three knots ($r = 10$) (i.e., $\boldsymbol{m} = \boldsymbol{B}\boldsymbol{\theta}$, where $\boldsymbol{B} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\theta} \in \mathbb{R}^r$ is a random vector such that $\boldsymbol{\theta}_i \sim \boldsymbol{N}(0,1),\ i \in \{1,\ldots,4\}$). The incoherence condition parameter $\mu = \mu(\boldsymbol{B}) = 0.82$. The sparse signal component is generated by a sparse random linear combination of degree 2 B-spline bases with 500 knots $(q = 500,$ $l = 4)$: that is, $\boldsymbol{a} = \boldsymbol{B}_a\boldsymbol{\theta}_a$, where $\boldsymbol{B}_a \in \mathbb{R}^{n\times q}$ and $\boldsymbol{\theta}_a \in \mathbb{R}^q$ is a four-sparse random vector ($s = 4$) such that its nonzero elements follow i.i.d. standard normal distribution. Notice that the RIC, $\delta_{\boldsymbol{B}_a,s}$ for matrix $\boldsymbol{B}_a$, is hard to calculate in general. However, there is a loose upper bound that can be used, which is $\delta_{\boldsymbol{B}_a,s} \le \delta_{\boldsymbol{B}_a,p} = \max\{\lambda_{max} - 1,$ $1 - \lambda_{min}\} = 0.60$, where λ_{max} and λ_{min} are the maximum and minimum singular values of $\boldsymbol{B}_a$, respectively. The noise signal generated as a random vector $\boldsymbol{e} \in \mathbb{R}^n$ is a random vector such that $\boldsymbol{e}_i \sim \boldsymbol{N}(0,\ 0.001^2),\ i \in \{1,\ldots,n\}$. Figure 2 shows a sample raw signal and its smooth, sparse components.

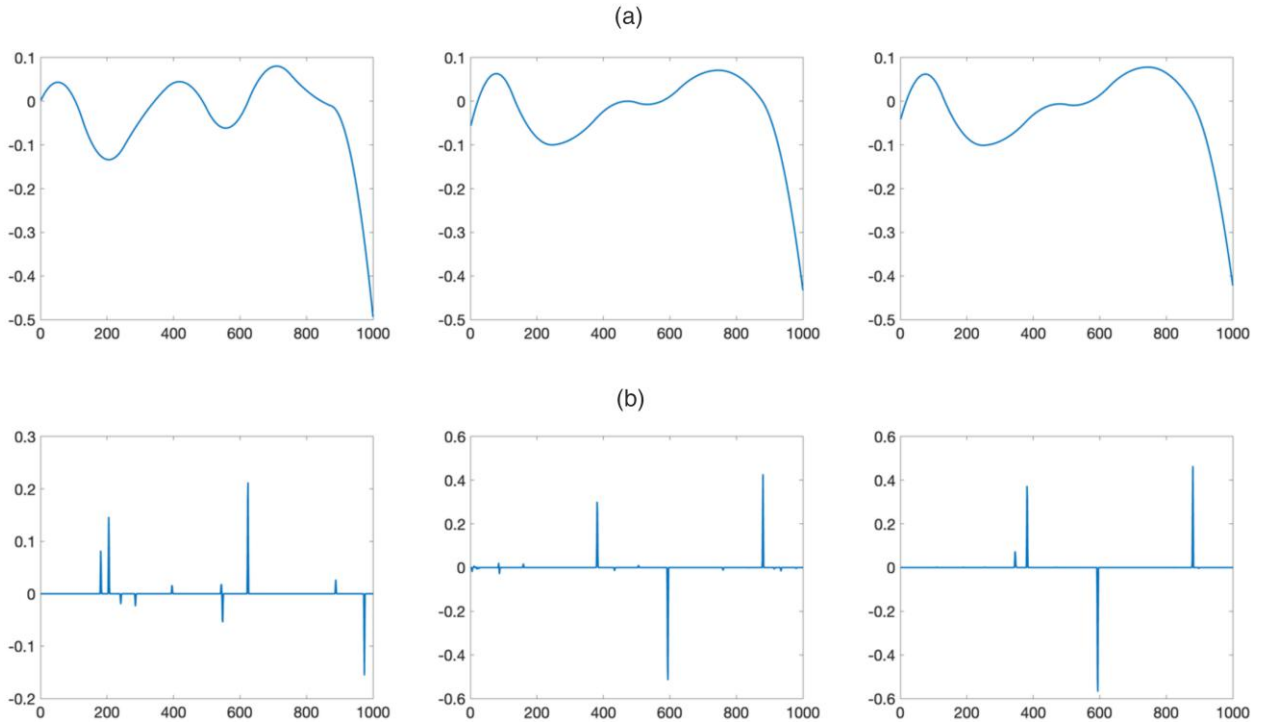
Before we state the result, we first check the assumption of Theorems 2.3 and 2.4. For Theorem 2.3, it is easy to check that $0.82 = \mu \le \frac{n}{2rsl} = \frac{1{,}000}{2\times10\times4\times4} = 6.25$ and also, that $\delta_{\boldsymbol{B}_a,s} \le 0.60 \in (0,1)$. For Theorem 2.4, $\delta_{\boldsymbol{B}_a,2s} \le 0.60 \in (0,1)$, $c = 0.37 > 0$, and $\delta_{r,3s,\mu,l} \in (0, 0.04)$. Notice that the range for $\delta_{r,3s,\mu,l}$ is small in this case because of the loose upper bound for $\delta_{\boldsymbol{B}_a,2s}$, which can be further improved. This indicates that the smooth and sparse signal can be recovered by the proposed algorithm with high probability provided that the compressive ratio is above a specific threshold, which is demonstrated by the following observation.

Figure 3. (Color online) Average Log Relative Reconstruction Error



Notes. (a) Log relative error for the smooth component. (b) Log relative error for the sparse component.

Figure 4. (Color online) Reconstructed Smooth and Sparse Signal Components



Notes. (a) Smooth signal component of compressive ratios 0.02 (left panel), 0.05 (center panel), and 0.1 (right panel). (b) Sparse signal component of compressive ratios 0.02 (left panel), 0.05 (center panel), and 0.1 (right panel).

The simulation is repeated 100 times, and the average log relative error for the background and sparse signal components with respect to compressive ratio are shown in Figure 3. We can observe a large error at the beginning, and it drops very fast with an increase of the compressive ratio. Then, above a threshold (0.1 approximately) of compressive ratio, the error becomes small (below 3%), and the decrease of error becomes less significant, which demonstrates the effectiveness of the reconstruction algorithm. The threshold of 0.1 can be chosen as the compressive ratio mentioned in Section 2.5.2.

The reconstructed signal components in 1 of the 100 simulations are shown in Figure 4. We can observe that when adopting the compressive ratio of 0.1, both the smooth and sparse signal components can be reconstructed with high accuracy.

We also vary the magnitude of sparse signals to examine the reconstruction performance of the proposed method. The result and analysis are provided in Appendix H.

3.2. CSSD on a 2D Image

In this simulation study, we aim to decompose an image into the smooth background, sparse anomalies,

Figure 5. (Color online) True Background and Anomalies

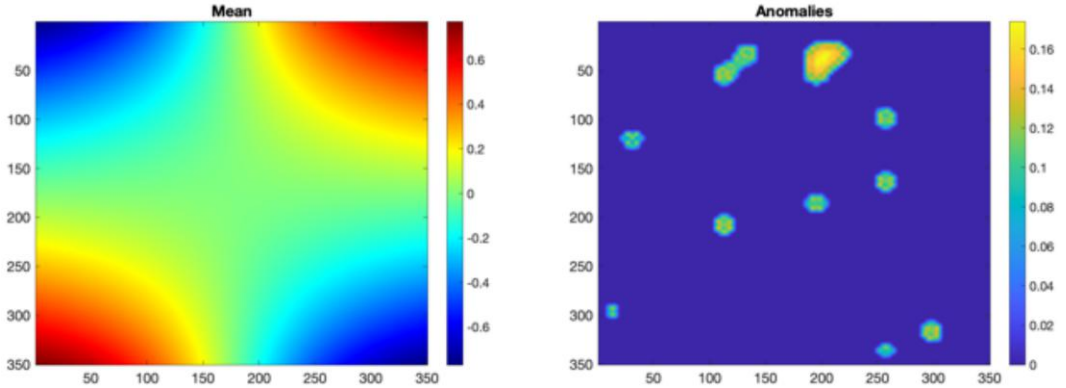
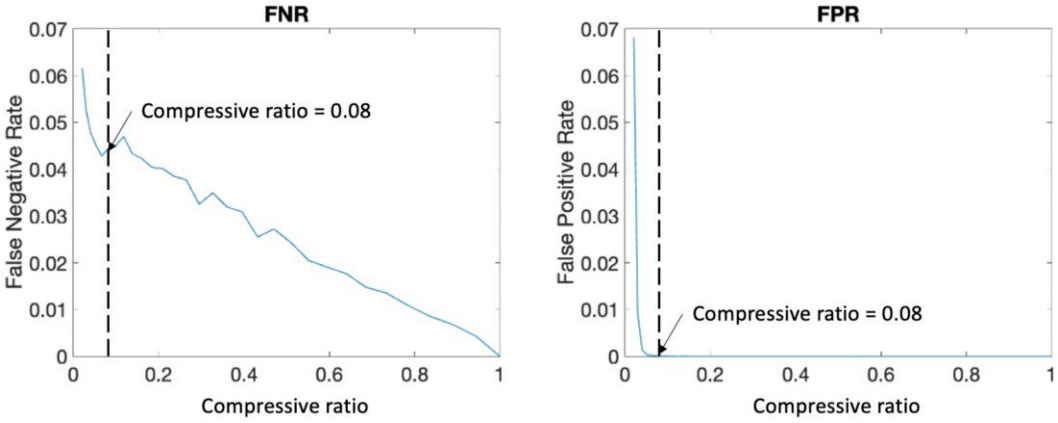


Figure 6. (Color online) Average FNR and FPR



and noise. A 350×350 image with smooth background and the sparse anomaly is generated similar to Yan et al. (2017) (i.e., $\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{A} + \boldsymbol{E}$, where $\boldsymbol{M}$ is the smooth background, $\boldsymbol{A}$ is the sparse anomalies, and $\boldsymbol{E}$ is i.i.d. Gaussian noise such that $\boldsymbol{E}_i \sim NID(0, \sigma^2)$). The smooth background is generated from a linear combination of B-spline bases with 3×3 knots, and the anomalies are generated from a sparse linear combination of B-spline bases with 88×88 knots. The background and anomalies are shown in Figure 5.

We can see that the anomaly size covers a large range. The mean absolute value of the background plus anomaly is $\mu = 0.21$. In this simulation, we first study the reconstruction performance under a noise-free scenario. Then, we increase the noise level to test the robustness of the algorithm.

3.2.1. Noise-Free Case. In this section, we vary the compressive ratio from 4% to 100%. For each compressive ratio, we simulate 100 times and the average false-negative rate (FNR) and false-positive rate (FPR) are reported in Figure 6. The FPR is defined as the portion of normal pixels predicted as anomaly,

$$FPR = \frac{\sum_{i,j}\left(1 - I_{\boldsymbol{A}\neq 0}\{\boldsymbol{A}(i,j)\}\right) I_{\hat{\boldsymbol{A}}\neq 0}\{\hat{\boldsymbol{A}}(i,j)\}}{\sum_{i,j}\left(1 - I_{\boldsymbol{A}\neq 0}\{\boldsymbol{A}(i,j)\}\right)},$$

and the FNR is defined as the portion of anomalous pixels predicted as normal background,

$$FNR = \frac{\sum_{i,j} I_{\boldsymbol{A}\neq 0}\{\boldsymbol{A}(i,j)\}\left(1 - I_{\hat{\boldsymbol{A}}\neq 0}\{\hat{\boldsymbol{A}}(i,j)\}\right)}{\sum_{i,j} I_{\boldsymbol{A}\neq 0}\{\boldsymbol{A}(i,j)\}},$$

where $\boldsymbol{A}$ is the true anomaly, $\hat{\boldsymbol{A}}$ is the predicted anomaly, and $I_\Omega(\cdot)$ is the indicator function: that is,

$$I_\Omega(x) = \begin{cases} 1, & \text{if } x \in \Omega \\ 0, & \text{otherwise.} \end{cases}$$

From Figure 6, we can see that both the FPR and FNR ratios decrease as the compressive ratio increases. The FPR drops so fast that when the compressive ratio achieves 8%, there is no false alarm, which is desired for the anomaly detection algorithm. The FNR drops slower, and less than 5% of the anomaly pixels are ignored when the compressive ratio achieves 8%. However, it does not miss any cluster of the anomalies, even though some of them are small.

The recovered sparse signal components are shown in Figure 7 for compressive ratios 4% (Figure 7(a)), 8% (Figure 7(b)), 33% (Figure 7(c)), and 73% (Figure 7(d)). For comparison, we apply the SSD algorithm (Yan et al. 2017), of which the FNR and the FPR are zero and computation time is 0.16 seconds.

We record the computation time in Figure 8. A significant boosting of the computation is observed when the compressive ratio is 8%, which speeds up the SSD algorithm by 4.3 times.

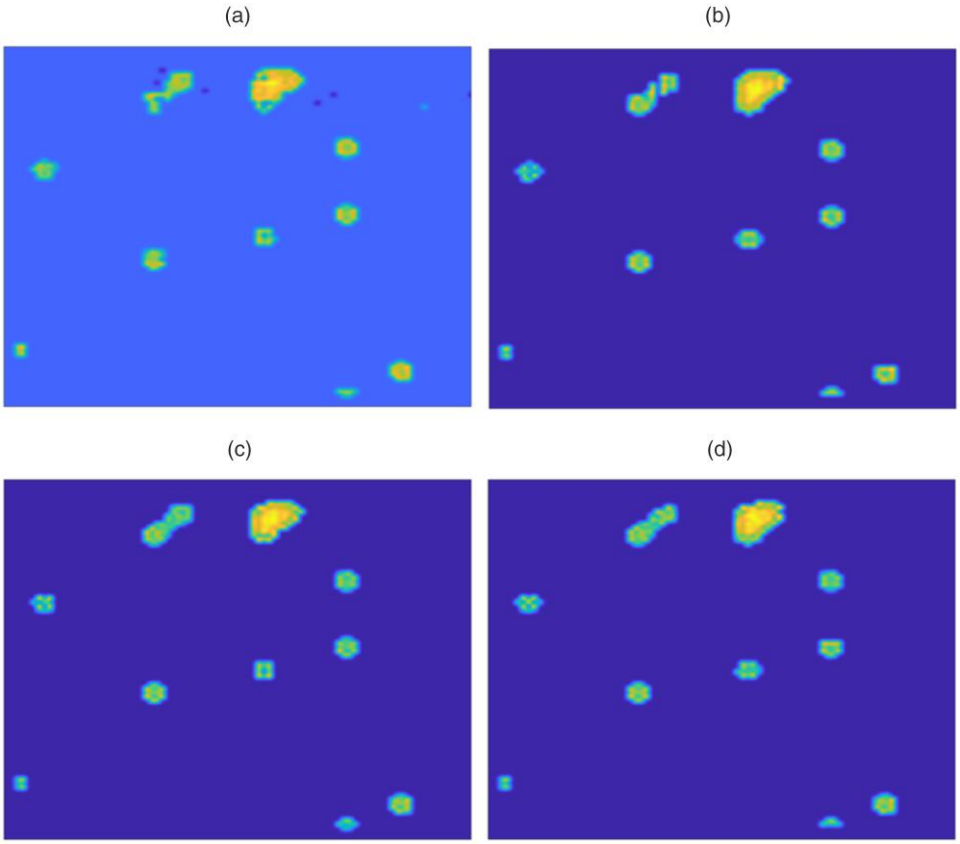
3.2.2. Noisy Case. The signal-to-noise ratio $\mu/\epsilon \in [4, 40]$ is studied to evaluate the robustness of the algorithm. The three-dimensional plot of FPR, the signal-to-noise ratio, and the compressive ratio are shown in Figure 9.

The purple line indicates the equipotential line of FPR = 0.01. We can see that with the increase of the signal-to-noise ratio, we are allowed to use a less compressive ratio to achieve satisfactory decomposition results. The majority of the area lies below the equipotential line of FPR = 0.01, which means that the proposed algorithm is robust.

4. Case Study

In this section, we use three real cases to demonstrate the effectiveness of the proposed CSSD/KronCSSD framework. For comparison, we also apply the SSD method in each case study. The compressive ratio (c) and the average computational time (t) for a single image are reported in Table 1. A significant transmission bandwidth reduction and computation boost can be observed with negligible performance degradation (Figures 10–12), compared with the vanilla SSD algorithm.

Figure 7. (Color online) The Recovered Sparse Signal Components



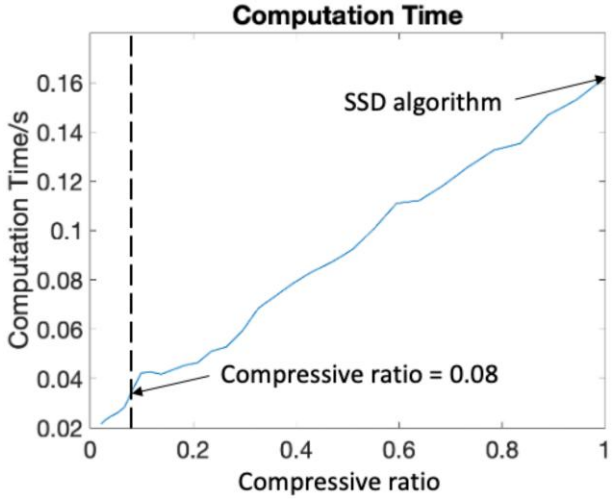
Notes. (a) Compressive ratio 4%. (b) Compressive ratio 8%. (c) Compressive ratio 33%. (d) Compressive ratio 73%.

4.1. Surface Defect Detection in a Steel Rolling Process

As mentioned in Section 1, vision sensors collect high-resolution images of the product surface with a high data acquisition rate in the rolling processes. This poses a challenge in data storage, transmission, and processing. One sample image of size 128×512 with typical anomalies is shown in Figure 10(a). The black scratches shown in the red block are surface anomalies. For a detailed description, the readers are encouraged to refer to Yan et al. (2018). The detected anomalies are shown in Figure 10, (b) and (c) by using the proposed KronCSSD algorithm and the SSD algorithm, respectively. The data set has 100 images, and more example results can be found in Appendix I.

The KronCSSD method achieves a similar anomaly detection performance with 54% bandwidth and is 1.8 times faster. By adopting the KronCSSD method, (i) we can achieve a faster anomaly detection and thus, reduce the loss through a timely intervention of the manufacturing process, and (ii) we can keep the manufacturing inspection information for a longer time period with the same storage capability, which is important for root cause analysis.

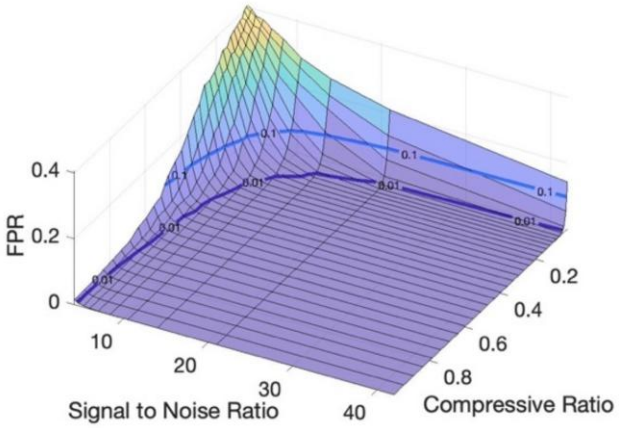
Figure 8. (Color online) The Computation Time



4.2. Solar Flare Detection

Another important application is solar flare detection from satellite images. A solar flare is defined as a sudden, transient, and intense variation in brightness over

Figure 9. (Color online) The Contour Plot of FPR, Signal-to-Noise Ratio, and Compressive Ratio



the sun's surface. It has a significant influence on radio communication on Earth. Each second, thousands of high-resolution images are captured by a satellite, which poses a big challenge to real-time data transmission and processing (Yan et al. 2018). The data set has 300 images, and more example images can be found in Appendix I.

One sample image of size 232×292 with a typical solar flare is shown in Figure 11(a), where the yellowish bright region is the solar flare. The detected anomalies are shown in Figure 11, (b) and (c) by using the proposed KronCSSD and SSD algorithms, respectively. The KronCSSD method achieves a similar solar flare detection performance as the SSD method but with 22% bandwidth, and it is 2.5 times faster. Notice that the decomposed images can also be used for downstream tasks, such as control charts and so on, which are beyond the scope of this paper.

By adopting the KronCSSD method, we can improve the transmission rate under the same transmission bandwidth and thus, achieve almost five times faster solar flare detection, which is of vital importance for protecting radio communications, power grids, and navigation systems.

Table 1. Comparison Between KronCSSD and SSD

	Surface defect		Solar flare		Indentation	
	c	t/s	c	t/s	c	t/s
KronCSSD	54%	0.073	22%	0.034	48%	0.053
SSD	1	0.135	1	0.086	1	0.094

4.3. Silicon Surface Indentation Detection

The stress map of size 90×550 of a silicon surface laminate with surface indentation is shown in Figure 12, where clusters of high-stress areas indicate the surface indentation (Yan et al. 2017). We aim to detect those high-stress areas. The detected anomalies are shown in Figure 12, (b) and (c). We can see that the KronCSSD method achieves a similar detection performance with the SSD method but with 40% bandwidth, and it is 1.8 times faster, which significantly reduces the storage and transmission cost and improves the processing speed.

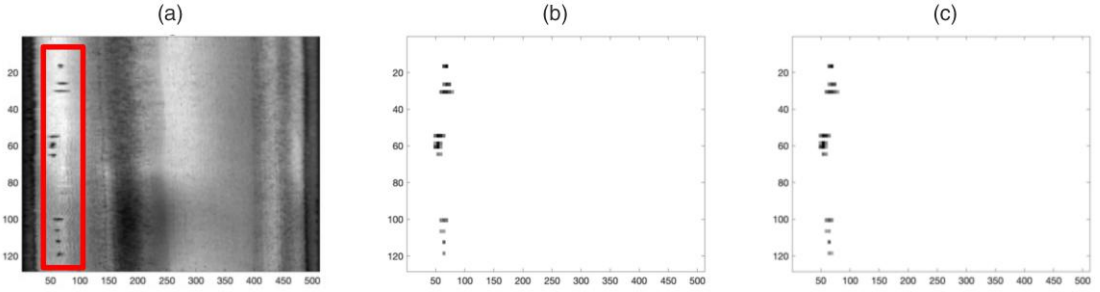
5. Conclusion

In this paper, we proposed a CSSD framework for efficient data acquisition, transmission, and processing for sparse anomaly detection in smooth backgrounds. To further enhance its computational efficiency, a KronCSSD framework is proposed for tensor data.

The contributions of this work are twofold. (i) Theoretically, we showed the feasibility of combining compressive sensing and smooth sparse decomposition. This enables the adoption of a compressive data acquisition approach. (ii) Practically, the proposed framework is compatible with many existing decomposition-based anomaly detection algorithms, such as SSD, ST-SSD, and so on, which achieve both a significant cost reduction in sensing, storage, and transmission and a boost in speed but with negligible loss in their performance.

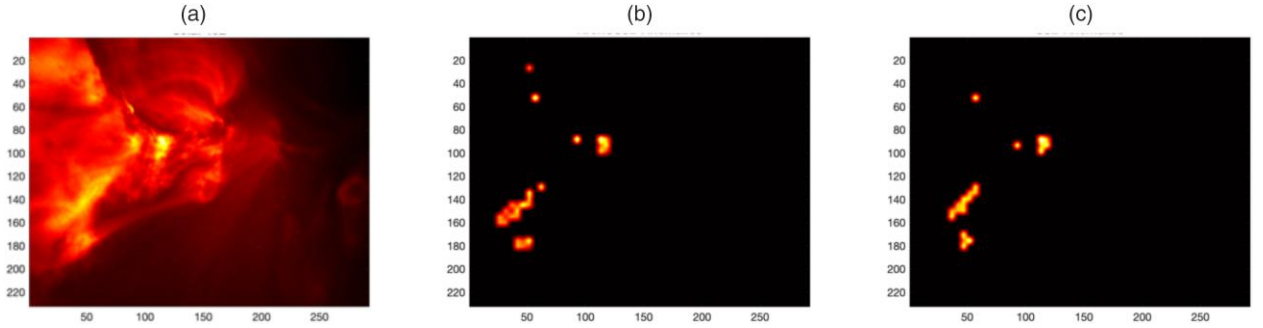
In this article, we use a simulation study to demonstrate the effectiveness and robustness of the proposed CSSD/KronCSSD framework. Three case studies across different applications demonstrate the versatility of the proposed framework. The authors believe that

Figure 10. (Color online) Steel Rolling Images and Detected Anomalies



Notes. (a) Raw image. (b) KronCSSD result. (c) SSD result.

Figure 11. (Color online) Solar Activity Images and Detected Solar Flare



Notes. (a) Raw image. (b) KronCSSD result. (c) SSD result.

the CSSD/KronCSSD framework can be applied in a wider range of applications toward more efficient data acquisition, transmission, and processing.

Further studies on the theoretical properties of the proposed KronCSSD can be a future direction.

Appendix A

Proof of Theorem 2.1. We prove Theorem 2.1 by contradiction. Assume there exists two different decompositions for the same smooth plus sparse signal $\boldsymbol{y}$ (i.e., $\boldsymbol{y} = \boldsymbol{m}_1 + \boldsymbol{a}_1 = \boldsymbol{m}_2 + \boldsymbol{a}_2$, where $\boldsymbol{m}_1 \neq \boldsymbol{m}_2$ and $\boldsymbol{a}_1 \neq \boldsymbol{a}_2$). Then,

$$\boldsymbol{m}_1 - \boldsymbol{m}_2 = -\boldsymbol{a}_1 + \boldsymbol{a}_2. \tag{A.1}$$

Because $\boldsymbol{m}_1 \neq \boldsymbol{m}_2$, we can normalize both side by $\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_2$ and denote $\tilde{\boldsymbol{m}} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)/\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_2$ and $\tilde{\boldsymbol{a}} = (-\boldsymbol{a}_1 + \boldsymbol{a}_2)/\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_2$. Notice that $\tilde{\boldsymbol{m}}$ is in the column space of $\boldsymbol{B}$, which is spanned by columns of the $\boldsymbol{U}$ (recall that $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$; i.e., $\tilde{\boldsymbol{m}} = \boldsymbol{U}\boldsymbol{x}$, where $\boldsymbol{x}$ is the coefficient vector, $\boldsymbol{x} \in \mathbb{R}^r$). Because $\|\tilde{\boldsymbol{m}}\|_2 = 1$, we have $\|\boldsymbol{x}\|_2 = 1$. We can bound each element in $\tilde{\boldsymbol{m}}$ as follows:

$$|\tilde{m}_i| = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{U}\boldsymbol{x} \le \|\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{U}\|_2\|\boldsymbol{x}\|_2 \le \max_{i\in\{1,..,r\}} \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\|_2 \le \sqrt{\frac{\mu(\boldsymbol{B})r}{n}} < \sqrt{\frac{1}{2ls}}, \quad \forall i \in \{1,\ldots,n\}.$$

According to Definition 2.1, $\|\boldsymbol{a}_1\|_0 \le ls$ and $\|\boldsymbol{a}_2\|_0 \le ls$. Therefore, $\|\tilde{\boldsymbol{a}}\|_0 \le 2ls$. Moreover, according to Equation (A.1), we conclude that $\|\tilde{\boldsymbol{m}}\|_0 \le 2ls$. Denote the support of $\tilde{\boldsymbol{m}}$ as $\mathbf{T}_{\tilde{\boldsymbol{m}}}$; we have $\|\tilde{\boldsymbol{m}}\|_2 = \sqrt{\sum_{i\in\mathbf{T}_{\tilde{\boldsymbol{m}}}} |\tilde{m}_i|^2} < 1$, which is a contradiction. □

Appendix B

Proof of Theorem 2.2. We prove Theorem 2.2 with contradiction. Assume there exists another vector $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a \in MS_{r,s,\mu,l}$ such that $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$ and $\boldsymbol{y} \neq \boldsymbol{y}_0$. Then, $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{y}_0 = \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \boldsymbol{B}_a(\boldsymbol{\theta}_a - \boldsymbol{\theta}_{a0})$ is a nonzero vector. Because $\|\boldsymbol{\theta}_a - \boldsymbol{\theta}_{a0}\|_1 \le \|\boldsymbol{\theta}_a\|_1 + \|\boldsymbol{\theta}_{a0}\|_1 \le 2s$, by Definition 2.3, we have that $\boldsymbol{z} \in MS_{r,2s,\mu,l}$. Therefore, $0 = \|\boldsymbol{A}\boldsymbol{z}\|_2 \ge (1 - \delta_{r,2s,\mu,l})\|\boldsymbol{z}\|_2 > 0$, which is a contradiction. Notice that the proof is inspired by the proof of lemma 3.1 in Candes and Tao (2005).

Appendix C

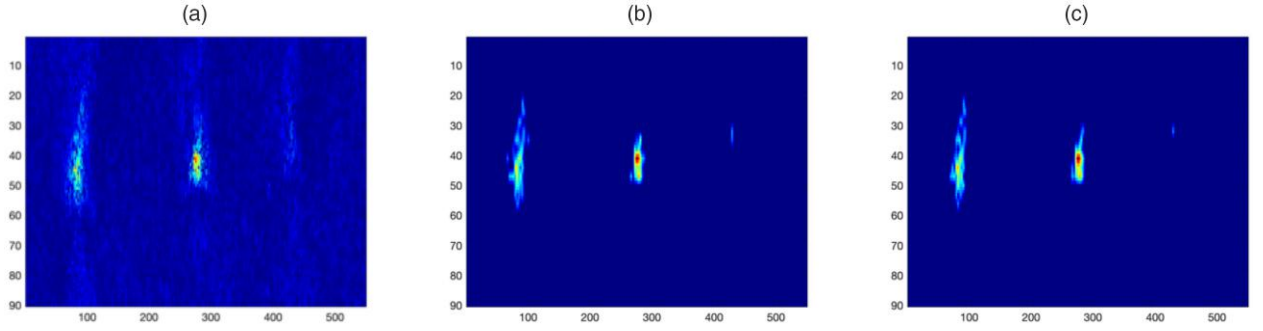
Proof of Theorem 2.3. This proof is inspired by Baraniuk et al. (2008) and Tanner and Vary (2020).

We will first derive the RIC for a fixed subspace $MS_{r,\mathbf{T},\mu,l}$ of $MS_{r,s,\mu,l}$ when $\boldsymbol{\theta}_a$ is restricted in a fixed subspace $\mathbf{T}$ with the fixed support such that the number of nonzero elements is s:

$$MS_{r,\mathbf{T},\mu,l} = \{\boldsymbol{y} \in \mathbb{R}^n \,|\, \boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a, \boldsymbol{B}_a \in \mathbb{R}^{n\times q},\ l(\boldsymbol{B}_a) = l,\ \boldsymbol{\theta}_a \in \mathbf{T},\ \boldsymbol{B} \in \mathbb{R}^{n\times r}, \mu(\boldsymbol{B}) = \mu < n(2rsl)^{-1},\ \boldsymbol{\theta} \in \mathbb{R}^r\}.$$

Then, we use a covering argument that counts over all possible sparse subspaces $\mathbf{T}$ with support less than or equal to s. Finally, we can derive the RIC for $MS_{r,s,\mu,l}$.

Figure 12. (Color online) Silicon Stress Map and Detected Indentation



Notes. (a) Raw image. (b) KronCSSD result. (c) SSD result.

The following lemma describes the RIC for a fixed subspace $MS_{r,\mathbf{T},\mu,l}$ and is proved in Appendix E.

Lemma C.1. *RIC for a fixed subspace $MS_{r,\mathbf{T},\mu,l}$. Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be a matrix from the families described in Definition 2.5. Furthermore, assume that $\mu < n(2rsl)^{-1}$ and that the basis matrix $\boldsymbol{B}_a$ for the sparse signal component satisfies the RIP with RIC $\delta_{\boldsymbol{B}_a,s} \in (0,1)$: that is, $\delta_{\boldsymbol{B}_a,s}$ is the smallest positive constant such that*

$$\left(1-\delta_{\boldsymbol{B}_a,s}\right)\|\boldsymbol{\theta}_a\|_2 \le \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2 \le \left(1+\delta_{\boldsymbol{B}_a,s}\right)\|\boldsymbol{\theta}_a\|_2,$$
$$\forall \boldsymbol{\theta}_a \in \{\boldsymbol{\theta}_a \in \mathbb{R}^q \,|\, \|\boldsymbol{\theta}_a\|_0 \le s\}.$$

For a given $\delta \in (0,1)$, there exists a constant $c_0 > 0$ depending only on δ, such that the RIC for $MS_{r,\mathbf{T},\mu,l}$ is upper bounded by δ with the probability of at least $1-2\left(\frac{24}{\delta}\tau_1\right)^r\left(\frac{24}{\delta}\tau_0\right)^s e^{-pc_0(\delta/2)}$, where $\eta = \sqrt{\frac{\mu rsl}{n}}$, $\tau_0 = \frac{1}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})(1-\eta^2)}}$, $\tau_1 = \|\boldsymbol{B}^\dagger\|_2\left(1+\frac{1}{\sqrt{1-\eta^2}}\right)$, and $\boldsymbol{B}^\dagger = \left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T$.

Notice that for a fixed subspace $MS_{r,\mathbf{T},\mu,l}$, the RIP will fail with probability less than or equal to $2\left(\frac{24}{\delta}\tau_1\right)^r\left(\frac{24}{\delta}\tau_0\right)^s e^{-pc_0(\delta/2)}$. Because there are $\binom{n}{s} \le \left(\frac{en}{s}\right)^s$ such subspaces, the probability to fail for $MS_{r,s,\mu,l}$, which is a combination of those $\binom{n}{s}$ subspaces, will be less than or equal to

$$\binom{n}{s}2\left(\frac{24}{\delta}\tau_1\right)^r\left(\frac{24}{\delta}\tau_0\right)^s e^{-pc_0\left(\frac{\delta}{2}\right)}$$
$$\le \exp\left(-c_0\left(\frac{\delta}{2}\right)p + \ln 2 + r\ln\frac{24}{\delta}\tau_1 + s\left(1+\ln\frac{24}{\delta}\tau_0 + \ln\frac{n}{s}\right)\right).$$

Then, for any give δ, there exist $c_1, c_2 > 0$, such that the probability to fail for $MS_{r,s,\mu,l}$ is less than or equal to $\exp(-c_1 p)$, provided that $p \ge c_2\left(\ln 2 + r\ln\frac{24}{\delta}\tau_1 + s\left(1+\ln\frac{24}{\delta}\tau_0 + \ln\frac{n}{s}\right)\right)$, where $c_2 = \left[c_0\left(\frac{\delta}{2}\right) - c_1\right]^{-1}$. This finishes the proof.

Appendix D

Proof of Theorem 2.4. The proof of Theorem 2.4 is inspired by Candes et al. (2006) and Tanner and Vary (2020). Assume that in Problem (1), ϵ_1 is properly chosen such that Problem (1) is feasible. In the following discussion, we will use $(\cdot)^*$ to denote the optimal solution of Problem (1) and $(\cdot)_0$ to denote the signal we wish to recover. Let $\boldsymbol{R} = \boldsymbol{X}^* - \boldsymbol{X}_0 = \boldsymbol{R}^m + \boldsymbol{R}^a$, where $\boldsymbol{R}^m = \boldsymbol{m} - \boldsymbol{m}_0 = \boldsymbol{B}\boldsymbol{\theta}^* - \boldsymbol{B}\boldsymbol{\theta}_0$ and $\boldsymbol{R}^a = \boldsymbol{B}_a\boldsymbol{\theta}_a^* - \boldsymbol{B}_a\boldsymbol{\theta}_{a0}$ are the residuals of the smooth and sparse signal components, respectively.

Let $\boldsymbol{h} = \boldsymbol{\theta}_a^* - \boldsymbol{\theta}_{a0} = \boldsymbol{h}_{\mathbf{T}_0} + \boldsymbol{h}_{\mathbf{T}_0^c}$, where $\mathbf{T}_0$ is the support of $\boldsymbol{\theta}_{a0}$, $\boldsymbol{h}_{\mathbf{T}_0}$ denotes the projection of $\boldsymbol{h}$ onto $\mathbf{T}_0$ such that

$$\boldsymbol{h}_{\mathbf{T}_0}(t) = \begin{cases} t, & \text{if } t \in \mathbf{T}_0 \\ 0, & \text{otherwise,} \end{cases}$$

and $\mathbf{T}_0^c$ denotes the complementary set of $\mathbf{T}_0$.

Because $\boldsymbol{\theta}_{a0}$ is feasible and $\boldsymbol{\theta}_a^*$ is the optimal solution of Problem (1), we must have $\|\boldsymbol{\theta}_a^*\|_1 \le \|\boldsymbol{\theta}_{a0}\|_1$, which is equivalent to $\|\boldsymbol{\theta}_{a0} + \boldsymbol{h}_{\mathbf{T}_0} + \boldsymbol{h}_{\mathbf{T}_0^c}\|_1 \le \|\boldsymbol{\theta}_{a0}\|_1$. Because $\mathbf{T}_0$ and $\mathbf{T}_0^c$ are complementary to each other, we have $\|\boldsymbol{\theta}_{a0} + \boldsymbol{h}_{\mathbf{T}_0}\|_1 + \|\boldsymbol{h}_{\mathbf{T}_0^c}\|_1 \le \|\boldsymbol{\theta}_{a0}\|_1$. Because $\|\boldsymbol{\theta}_{a0} + \boldsymbol{h}_{\mathbf{T}_0}\|_1 \ge \|\boldsymbol{\theta}_{a0}\|_1 - \|\boldsymbol{h}_{\mathbf{T}_0}\|_1$, we have $\|\boldsymbol{\theta}_{a0}\|_1 - \|\boldsymbol{h}_{\mathbf{T}_0}\|_1 + \|\boldsymbol{h}_{\mathbf{T}_0^c}\|_1 \le \|\boldsymbol{\theta}_{a0}\|_1$. Hence, $\|\boldsymbol{h}_{\mathbf{T}_0^c}\|_1 \le \|\boldsymbol{h}_{\mathbf{T}_0}\|_1$. Because $\|\boldsymbol{h}_{\mathbf{T}_0}\|_1 \le \sqrt{s}\|\boldsymbol{h}_{\mathbf{T}_0}\|_2$, we have

$$\|\boldsymbol{h}_{\mathbf{T}_0^c}\|_1 \le \sqrt{s}\|\boldsymbol{h}_{\mathbf{T}_0}\|_2. \tag{D.1}$$

Similar to Candes et al. (2006), we order the elements of $\mathbf{T}_0^c$ in decreasing order of their magnitude and enumerate $\mathbf{T}_0^c$ as $v_1, \ldots, v_{n-|\mathbf{T}_0|}$. Then, $\mathbf{T}_0^c$ is divided into subsets $\mathbf{T}_i^c$ of size M, where

$$\mathbf{T}_i^c = \left\{v_j : (i-1)M \le j \le iM\right\}.$$

Let $\boldsymbol{h}_{\mathbf{T}_i^c}$ be the projection of $\boldsymbol{h}$ onto $\mathbf{T}_i^c$; we have

$$\begin{aligned} &\|\boldsymbol{h}_{\mathbf{T}_i^c}\|_0 \le M, \quad \forall i \ge 1 \\ &\mathbf{T}_i^c \cap \mathbf{T}_j^c = \emptyset, \quad \forall i \ne j \\ &\|\boldsymbol{h}_{\mathbf{T}_{i+1}^c}\|_2 \le \frac{1}{\sqrt{M}}\|\boldsymbol{h}_{\mathbf{T}_i^c}\|_1, \quad \forall i \ge 1, \end{aligned} \tag{D.2}$$

where the last inequality comes from the fact that $\mathbf{T}_0^c$ is in decreasing order, such that

$$\left|\boldsymbol{h}_{\mathbf{T}_{i+1}^c}\right|_{(v)} \le \frac{1}{M}\sum_{j\in\mathbf{T}_i^c}\left|\boldsymbol{h}_{\mathbf{T}_i^c}\right|_{(j)}, \ \forall v \in \mathbf{T}_{i+1}^c.$$

Define $\boldsymbol{R}^a_{\mathbf{T}_i^c} = \boldsymbol{B}_a\boldsymbol{h}_{\mathbf{T}_i^c}$ and $\boldsymbol{R}^a_{\mathbf{T}_0} = \boldsymbol{B}_a\boldsymbol{h}_{\mathbf{T}_0}$, and combine Equations (D.1) and (D.2). Then, we have

$$\begin{aligned} \sum_{j\ge2}\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2 &\le \sum_{j\ge2}\sqrt{1+\delta_{\boldsymbol{B}_a,M}}\|\boldsymbol{h}_{\mathbf{T}_j^c}\|_2 \\ &\le_{(a)} \sum_{j\ge1}\frac{\sqrt{1+\delta_{\boldsymbol{B}_a,M}}\|\boldsymbol{h}_{\mathbf{T}_j^c}\|_1}{\sqrt{M}} = \frac{\sqrt{1+\delta_{\boldsymbol{B}_a,M}}\|\boldsymbol{h}_{\mathbf{T}_0^c}\|_1}{\sqrt{M}} \\ &\le_{(b)} \frac{\sqrt{s}\sqrt{1+\delta_{\boldsymbol{B}_a,M}}\|\boldsymbol{h}_{\mathbf{T}_0}\|_2}{\sqrt{M}} \le \frac{\sqrt{s}\sqrt{1+\delta_{\boldsymbol{B}_a,M}}\|\boldsymbol{R}^a_{\mathbf{T}_0}\|_2}{\sqrt{M}\sqrt{1-\delta_{\boldsymbol{B}_a,M}}}, \end{aligned}$$

where (a) follows Equation (D.2), (b) follows Equation (D.1), and the RIP property of $\boldsymbol{B}_a$ is used because $\|\boldsymbol{h}_{\mathbf{T}_j^c}\|_0 \le M$.

Denote $\gamma = \sqrt{\frac{1+\delta_{\boldsymbol{B}_a,M+s}}{1-\delta_{\boldsymbol{B}_a,M+s}}}$. (A tighter bound can be achieved by using $\sqrt{\frac{1+\delta_{\boldsymbol{B}_a,M}}{1-\delta_{\boldsymbol{B}_a,M}}}$. However, we adopt $\delta_{\boldsymbol{B}_a,M+s}$ instead of $\delta_{\boldsymbol{B}_a,M}$ for simplicity in the following proof.) Then, we have

$$\sum_{j\ge2}\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2 \le \sqrt{\frac{s}{M}}\gamma\|\boldsymbol{R}^a_{\mathbf{T}_0}\|_2. \tag{D.3}$$

Next, we derive the bounds for $\boldsymbol{R}^m$ and $\boldsymbol{R}^a$.

Bound for $\boldsymbol{R}^m$:

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{R}^m\|_2^2 &= |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}(\boldsymbol{R}-\boldsymbol{R}^a)\rangle| \\ &= |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}(\boldsymbol{R}-\boldsymbol{R}^a)\rangle| \\ &= |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}\rangle + \langle \boldsymbol{A}\boldsymbol{R}^m, -\boldsymbol{A}\boldsymbol{R}^a\rangle| \\ &\le |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}\rangle| + |\langle \boldsymbol{A}\boldsymbol{R}^m, -\boldsymbol{A}\boldsymbol{R}^a\rangle| \\ &= |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}\rangle| + \left|\left\langle \boldsymbol{A}\boldsymbol{R}^m, -\boldsymbol{A}\left(\boldsymbol{R}^a_{\mathbf{T}_0} + \boldsymbol{R}^a_{\mathbf{T}_1^c} + \sum_{j\ge2}\boldsymbol{R}^a_{\mathbf{T}_j^c}\right)\right\rangle\right| \\ &\le |\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}\rangle| + \left|\left\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\left(\boldsymbol{R}^a_{\mathbf{T}_0} + \boldsymbol{R}^a_{\mathbf{T}_1^c}\right)\right\rangle\right| \\ &\quad + \sum_{j\ge2}\left|\left\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}^a_{\mathbf{T}_j^c}\right\rangle\right|. \end{aligned} \tag{D.4}$$

In the following discussion, we will bound those terms. According to the Cauchy–Schwarz inequality, the first term

can be bounded as

$$|\langle \boldsymbol{AR}^{\boldsymbol{m}}, \boldsymbol{AR}\rangle| \le \|\boldsymbol{AR}^{\boldsymbol{m}}\|_2\|\boldsymbol{AR}\|_2 \le \sqrt{1+\delta_{r,s,\mu,I}}\epsilon_1\|\boldsymbol{R}^{\boldsymbol{m}}\|_2, \quad \text{(D.5)}$$

where the last inequality comes from the RIP property and the first constraint in Problem (1).

The third term can be bounded as follows. Denoting $z_1 = \boldsymbol{R}^{\boldsymbol{m}}/\|\boldsymbol{R}^{\boldsymbol{m}}\|_2$ and $z_2 = \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}/\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\|_2$, we have

$$\begin{aligned}
&\frac{|\langle \boldsymbol{AR}^{\boldsymbol{m}}, \boldsymbol{AR}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\rangle|}{\|\boldsymbol{R}^{\boldsymbol{m}}\|_2\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\|_2}\\
&= |\langle \boldsymbol{Az}_1, \boldsymbol{Az}_2\rangle| = \frac{1}{4}|\|\boldsymbol{A}(z_1+z_2)\|_2^2 - \|\boldsymbol{A}(z_1-z_2)\|_2^2|\\
&\le_{(a)} \frac{1}{4}\max\left\{\begin{array}{l}|(1+\delta_{r,M,\mu,I})\|z_1+z_2\|_2^2 - (1-\delta_{r,M,\mu,I})\|z_1-z_2\|_2^2|,\\ |(1+\delta_{r,M,\mu,I})\|z_1-z_2\|_2^2 - (1-\delta_{r,M,\mu,I})\|z_1+z_2\|_2^2|\end{array}\right\}\\
&= |\delta_{r,M,\mu,I} + \langle z_1, z_2\rangle| \le_{(b)} \delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2},
\end{aligned}$$

where $\eta_1 = \sqrt{\frac{\mu r M I}{n}}$; inequality (*a*) follows the RIP property because $z_1 + z_2 \in MS_{r,M,\mu,I}$ and $z_1 - z_2 \in MS_{r,M,\mu,I}$; and inequality (*b*) follows Equation (F.1) in the proof of Lemma E.1 and $\|z_1\|_2 = \|z_2\|_2 = 1$, $\langle z_1, z_2\rangle \le \frac{\eta_1}{1-\eta_1^2}\|z_1\|_2\|z_2\|_2 = \frac{\eta_1}{1-\eta_1^2}$, provided that $M \le 2s$.

Therefore,

$$|\langle \boldsymbol{AR}^{\boldsymbol{m}}, \boldsymbol{AR}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\rangle| \le \left(\delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2}\right)\|\boldsymbol{R}^{\boldsymbol{m}}\|_2\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\|_2. \quad \text{(D.6)}$$

Similarly, the second term can be bounded as

$$\begin{aligned}
&|\langle \boldsymbol{AR}^{\boldsymbol{m}}, \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c})\rangle|\\
&\le \left(\delta_{r,M+s,\mu,I} + \frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^{\boldsymbol{m}}\|_2\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2, \quad \text{(D.7)}
\end{aligned}$$

where $\eta_2 = \sqrt{\frac{\mu r(M+s)I}{n}}$, provided that $M \le s$.

Plugging Equations (D.5)–(D.7) into Equation (D.4), we have

$$\begin{aligned}
&\|\boldsymbol{AR}^{\boldsymbol{m}}\|_2^2\\
&\le \|\boldsymbol{R}^{\boldsymbol{m}}\|_2\left(\sqrt{1+\delta_{r,s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I} + \frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2\right.\\
&\quad \left. + \left(\delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2}\right)\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\|_2\right)\\
&\le_{(a)} \|\boldsymbol{R}^{\boldsymbol{m}}\|_2\left(\sqrt{1+\delta_{r,s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I} + \frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2\right.\\
&\quad \left. + \left(\delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2}\right)\sqrt{\frac{s}{M}}\gamma\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0}\|_2\right),
\end{aligned}$$

where inequality (*a*) follows Equation (D.3).

According to the RIP property, we have

$$\begin{aligned}
&(1-\delta_{r,s,\mu,I})\|\boldsymbol{R}^{\boldsymbol{m}}\|_2^2\\
&\le \|\boldsymbol{R}^{\boldsymbol{m}}\|_2\left(\sqrt{1+\delta_{r,s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I} + \frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2\right.\\
&\quad \left. + \left(\delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2}\right)\sqrt{\frac{s}{M}}\gamma\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0}\|_2\right).
\end{aligned}$$

Consequently, we have

$$\|\boldsymbol{R}^{\boldsymbol{m}}\|_2 \le \frac{\begin{array}{l}\sqrt{1+\delta_{r,s,\mu,I}}\epsilon_1 + \left(\left(\delta_{r,M,\mu,I} + \frac{\eta_1}{1-\eta_1^2}\right)\sqrt{\frac{s}{M}}\gamma^2\right.\\ \left. + \left(\delta_{r,M+s,\mu,I} + \frac{\eta_2}{1-\eta_2^2}\right)\right)\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2\end{array}}{(1-\delta_{r,s,\mu,I})}, \quad \text{(D.8)}$$

where the inequality follows from

$$\begin{aligned}
\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0}\|_2 &= \|\boldsymbol{B}_a\boldsymbol{h}_{\mathbf{T}_0}\|_2 \le_{(a)} \sqrt{1+\delta_{B_a,s}}\|\boldsymbol{h}_{\mathbf{T}_0}\|_2 \le_{(b)} \sqrt{1+\delta_{B_a,s}}\|\boldsymbol{h}_{\mathbf{T}_0} + \boldsymbol{h}_{\mathbf{T}_1^c}\|_2\\
&\le_{(c)} \sqrt{\frac{1+\delta_{B_a,s}}{1-\delta_{B_a,M+s}}}\|\boldsymbol{B}_a(\boldsymbol{h}_{\mathbf{T}_0} + \boldsymbol{h}_{\mathbf{T}_1^c})\|_2 \le \gamma\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2, \quad \text{(D.9)}
\end{aligned}$$

inequalities (*a*) and (*c*) follow the RIP property of $\boldsymbol{B}_a$, and inequality (*b*) follows that $\mathbf{T}_0 \cap \mathbf{T}_1^c = \emptyset$.

Bound for $\boldsymbol{R}^{\boldsymbol{a}}$:

$$\begin{aligned}
&\|\boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c})\|_2^2 = |\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c} - \boldsymbol{R} + \boldsymbol{R})\rangle|\\
&= |\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), \boldsymbol{AR}\rangle| + \left|\left\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), -\boldsymbol{A}\left(\boldsymbol{R}^{\boldsymbol{m}} + \sum_{j\ge 2}\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\right)\right\rangle\right|\\
&\le |\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), \boldsymbol{AR}\rangle| + |\langle \boldsymbol{AR}^{\boldsymbol{m}}, \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c})\rangle|\\
&\quad + \sum_{j\ge 2}|\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), \boldsymbol{AR}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\rangle|. \quad \text{(D.10)}
\end{aligned}$$

In the following discussion, we will bound those terms. According to the Cauchy–Schwarz inequality, the first term can be bounded as

$$\begin{aligned}
|\langle \boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}), \boldsymbol{AR}\rangle| &\le \|\boldsymbol{A}(\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c})\|_2\|\boldsymbol{AR}\|_2\\
&\le \sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_0} + \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_1^c}\|_2, \quad \text{(D.11)}
\end{aligned}$$

where the last inequality comes from the RIP property and the first constraint in Problem (1).

The third term can be bounded as follows. Denoting $z_2 = \boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}/\|\boldsymbol{R}^{\boldsymbol{a}}_{\mathbf{T}_j^c}\|_2$, $j \ge 2$ and $z_3 = (\boldsymbol{R}^{\boldsymbol{a}}_{(\mathbf{T}_0)} + \boldsymbol{R}^{\boldsymbol{a}}_{(\mathbf{T}_1^c)})/\|\boldsymbol{R}^{\boldsymbol{a}}_{(\mathbf{T}_0)} + \boldsymbol{R}^{\boldsymbol{a}}_{(\mathbf{T}_1^c)}\|_2$, we

have

$$\begin{aligned}
&\frac{\left|\left\langle \boldsymbol{A}(\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}), \boldsymbol{A}\boldsymbol{R}^a_{\mathbf{T}_j^c}\right\rangle\right|}{\|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2} = |\langle \boldsymbol{A}z_3, \boldsymbol{A}z_2\rangle| \\
&= \frac{1}{4}\left|\|\boldsymbol{A}(z_3+z_2)\|_2^2 - \|\boldsymbol{A}(z_3-z_2)\|_2^2\right| \\
&\le_{(a)} \frac{1}{4}\max\left\{\begin{array}{l}|(1+\delta_{r,2M+s,\mu,I})\|z_3+z_2\|_2^2 - (1-\delta_{r,2M+s,\mu,I})\|z_3-z_2\|_2^2|, \\ |(1+\delta_{r,M+2s,\mu,I})\|z_3-z_2\|_2^2 - (1-\delta_{r,2M+s,\mu,I})\|z_3+z_2\|_2^2|\end{array}\right\} \\
&= |\delta_{r,2M+s,\mu,I} + \langle z_3, z_2\rangle| \\
&\le_{(b)} \delta_{r,2M+s,\mu,I},
\end{aligned}$$

where inequality (*a*) follows the RIP property because $z_3+z_2 \in MS_{r,2M+s,\mu,I}$ and $z_3-z_2 \in MS_{r,2M+s,\mu,I}$. Inequality (*b*) comes from that $\mathbf{T}_i^c \cap \mathbf{T}_j^c = \emptyset,\ \forall i \ne j$ and $\mathbf{T}_0 \cap \mathbf{T}_i^c = \emptyset,\ \forall i$.

Therefore,

$$\left|\left\langle \boldsymbol{A}\boldsymbol{R}^m, \boldsymbol{A}\boldsymbol{R}^a_{\mathbf{T}_j^c}\right\rangle\right| \le \delta_{r,2M+s,\mu,I}\|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2. \tag{D.12}$$

Plugging Equations (D.7), (D.11), and (D.12) into Equation (D.10), we have

$$\begin{aligned}
&\|\boldsymbol{A}(\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c})\|_2^2 \\
&\le \|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\left(\sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^m\|_2 \right. \\
&\qquad \left. + \delta_{r,2M+s,\mu,I}\sum_{j\ge 2}\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2\right) \\
&\le_{(a)} \|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\left(\sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^m\|_2 \right. \\
&\qquad \left. + \delta_{r,2M+s,\mu,I}\sqrt{\frac{s}{M}}\gamma\|\boldsymbol{R}^a_{\mathbf{T}_0}\|_2\right) \\
&\le_{(b)} \|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\left(\sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^m\|_2 \right. \\
&\qquad \left. + \delta_{r,2M+s,\mu,I}\sqrt{\frac{s}{M}}\gamma^2\left(\|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|\right)\right),
\end{aligned}$$

where inequalities (*a*) and (*b*) follows the same argument as deriving Equation (D.8).

According to the RIP property, we have

$$\begin{aligned}
&(1-\delta_{r,M+s,\mu,I})\|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2^2 \\
&\le \|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2\left(\sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^m\|_2 \right. \\
&\qquad \left. + \delta_{r,2M+s,\mu,I}\sqrt{\frac{s}{M}}\gamma^2\|\boldsymbol{R}^a_{\mathbf{T}_0}\|_2\right).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
&\|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2 \\
&\left(\sqrt{1+\delta_{r,M+s,\mu,I}}\epsilon_1 + \left(\delta_{r,M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right)\|\boldsymbol{R}^m\|_2 \right. \\
&\left. + \delta_{r,2M+s,\mu,I}\sqrt{\frac{s}{M}}\gamma^2\|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|\right) \\
&\le \frac{}{(1-\delta_{r,M+s,\mu,I})}.
\end{aligned} \tag{D.13}$$

Notice that Equations (D.8) and (D.13) still hold if we relax $\delta_{r,M+s,\mu,I}$, $\delta_{r,s,\mu,I}$ to $\delta_{r,2M+s,\mu,I}$. For simplicity, here we replace $\delta_{r,M+s,\mu,I}$, $\delta_{r,s,\mu,I}$ with $\delta_{r,2M+s,\mu,I}$ in the following derivation.

Plugging Equation (D.8) into Equation (D.13) and letting $x \equiv \|\boldsymbol{R}^a_{\mathbf{T}_0}+\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2$, $y \equiv \|\boldsymbol{R}^m\|_2$, we have

$$\left(D_1 - \frac{B_1B_2}{D_2} - C_1\right)x \le A_1\epsilon_1 + \frac{B_1}{D_2}A_2\epsilon_1, \tag{D.14}$$

where

$$A_1 = \sqrt{1+\delta_{r,2M+s,\mu,I}},\ B_1 = \left(\delta_{r,2M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right),$$
$$C_1 = \delta_{r,2M+s,\mu,I}\sqrt{\frac{s}{M}}\gamma^2,\ D_1 = \left(1-\delta_{r,2M+s,\mu,I}\right),$$

and

$$\begin{aligned}
A_2 &= \sqrt{1+\delta_{r,2M+s,\mu,I}}, \\
B_2 &= \left(\delta_{r,2M+s,\mu,I}+\frac{\eta_1}{1-\eta_1^2}\right)\sqrt{\frac{s}{M}}\gamma^2 + \left(\delta_{r,2M+s,\mu,I}+\frac{\eta_2}{1-\eta_2^2}\right), \\
D_2 &= \left(1-\delta_{r,2M+s,\mu,I}\right).
\end{aligned}$$

Here, we require $D_1 - \frac{B_1B_2}{D_2} - C_1 > 0$ in Equation (D.14) and let $M=s$, which is

$$1-\gamma^2\alpha_1\alpha_2 - {\alpha_2}^2 - \left((1+\alpha_1+\alpha_2)\gamma^2 + 2\alpha_2 + 2\right)\delta_{r,3s,\mu,I} > 0. \tag{D.15}$$

Let $a = (1+\alpha_1+\alpha_2)\gamma^2 + 2\alpha_2 + 2$ and $c = 1-\gamma^2\alpha_1\alpha_2 - {\alpha_2}^2$, where $\alpha_1 = \frac{\eta}{1-\eta^2}$, $\alpha_2 = \frac{\sqrt{2}\eta}{1-2\eta^2}$, $\gamma = \sqrt{\frac{1+\delta_{B_a,2s}}{1-\delta_{B_a,2s}}}$, and $\eta = \sqrt{\frac{\mu r s l}{n}}$. If $c>0$, then there exist a $\delta_{r,3s,\mu,I} > 0$ such that Equation (D.15) is valid. Then, the denominator $-a\delta_{r,3s,\mu,I} + c > 0\ \ \forall \delta_{r,3s,\mu,I} \in (0, c/a)$. Consequently,

$$\|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2 \le \frac{(1+\alpha_2)\sqrt{1+\delta_{r,3s,\mu,I}}}{c - a\delta_{r,3s,\mu,I}}\epsilon_1.$$

Notice that from Equations (D.1) and (D.9), we have

$$\sum_{j\ge 2}\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2 \le \gamma\|\boldsymbol{R}^a_{\mathbf{T}_0}\|_2 \le \gamma^2\|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2.$$

Therefore, $\|\boldsymbol{R}^a\|_2 \le \|\boldsymbol{R}^a_{\mathbf{T}_0}\| + \|\boldsymbol{R}^a_{\mathbf{T}_1^c}\|_2 + \sum_{j\ge 2}\|\boldsymbol{R}^a_{\mathbf{T}_j^c}\|_2 \le C_a\epsilon_1$, where

$$C_a = \frac{(1+\gamma^2)(1+\alpha_2)\sqrt{1+\delta_{r,3s,\mu,I}}}{c - a\delta_{r,3s,\mu,I}}.$$

Similarly, we can bound $\|\boldsymbol{R}^m\|_2$ as $\|\boldsymbol{R}^m\|_2 \le C_m\epsilon_1$, where

$$C_m = \frac{\sqrt{1+\delta_{r,3s,\mu,I}} + \left(\delta_{r,3s,\mu,I} + \frac{\gamma^2}{1+\gamma^2}\alpha_1 + \frac{1}{1+\gamma^2}\alpha_2\right)C_a}{\left(1-\delta_{r,3s,\mu,I}\right)}. \quad \square$$

Appendix E

Proof of Lemma C.1. In this section, we will provide the proof for Lemma C.1. By linearity of the measurement matrix $\boldsymbol{A}$, without loss of generality, it is enough to prove this lemma when $\|\boldsymbol{y}_2\| = 1$. The proof mainly has two steps. First, the bounds for $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}$ are derived, and a finite set of points to approximate the set $MS_{r,\mathbf{T},\mu,I}$ to any accuracy in

norm 2 sense can be found. Then, the concentration inequality can be applied through a union bound. This is a common approach in compressive sensing literature (Baraniuk et al. 2008, Tanner and Vary 2020).

To derive the upper bounds for $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}$, we first derive the upper bounds for the signals $\boldsymbol{m} = \boldsymbol{B}\boldsymbol{\theta}$ and $\boldsymbol{a} = \boldsymbol{B}_a\boldsymbol{\theta}_a$, which are given in the following lemma.

Lemma E.1. *The smooth signal component $\boldsymbol{m}$ and sparse signal component $\boldsymbol{a}$ of the signal $\boldsymbol{y}$ in $MS_{r,\boldsymbol{T},\mu,l}$ with $\mu < \frac{n}{2rsl}$ can be bounded as follows*:

$$\|\boldsymbol{m}\|_2 = \|\boldsymbol{B}\boldsymbol{\theta}\|_2 \le \frac{\|\boldsymbol{y}\|_2}{\sqrt{1-\eta^2}}, \tag{E.1}$$

$$\|\boldsymbol{a}\|_2 = \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2 \le \frac{\|\boldsymbol{y}\|_2}{\sqrt{1-\eta^2}}, \tag{E.2}$$

where $\eta = \sqrt{\frac{\mu rsl}{n}}$.

The proof is presented in Appendix F.

According to the RIP for $\boldsymbol{B}_a$, we have

$$\sqrt{(1-\delta_{\boldsymbol{B}_a,s})}\|\boldsymbol{\theta}_a\|_2 \le \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2 \le \sqrt{(1+\delta_{\boldsymbol{B}_a,s})}\|\boldsymbol{\theta}_a\|_2. \tag{E.3}$$

Combining Equations (E.2) and (E.3), we have

$$\|\boldsymbol{\theta}_a\|_2 \le \frac{1}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})}}\|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2 \le \frac{\|\boldsymbol{y}\|_2}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})(1-\eta^2)}} = \frac{1}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})(1-\eta^2)}}.$$

Denoting $\tau_0 = \frac{1}{\sqrt{(1-\delta_{\boldsymbol{B}_a,s})(1-\eta^2)}}$, we have $\|\boldsymbol{\theta}_a\|_2 \le \tau_0$. Recall that $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a$; we have $\boldsymbol{\theta} = \boldsymbol{B}^{\dagger}(\boldsymbol{y} - \boldsymbol{B}_a\boldsymbol{\theta}_a)$, where $\boldsymbol{B}^{\dagger} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}$. Therefore, according to the triangle inequality and the Cauchy–Schwarz inequality, we have

$$\|\boldsymbol{\theta}\|_2 \le \|\boldsymbol{B}^{\dagger}\|_2(\|\boldsymbol{y}\|_2 + \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2) \le \|\boldsymbol{B}^{\dagger}\|_2\left(1 + \frac{1}{\sqrt{1-\eta^2}}\right).$$

Denoting $\tau_1 = \|\boldsymbol{B}^{\dagger}\|_2\left(1 + \frac{1}{\sqrt{1-\eta^2}}\right)$, we have $\|\boldsymbol{\theta}\|_2 \le \tau_1$.

Because we have derived the bounds for $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}$, the covering number of $MS_{r,\boldsymbol{T},\mu,l}$ is given by the following lemma whose proof is in Appendix G.

Lemma E.2. *There exists a set $\boldsymbol{Q} \in MS_{r,\boldsymbol{T},\mu,l}$, such that for all $\boldsymbol{y} \in MS_{r,\boldsymbol{T},\mu,l}$, with $\|\boldsymbol{y}\|_2 = 1$ we have $\min_{\boldsymbol{q}\in\boldsymbol{Q}}\|\boldsymbol{q} - \boldsymbol{y}\|_2 \le \frac{\delta}{4}$ and $|\boldsymbol{Q}| \le \left(\frac{24}{\delta}\tau_1\right)^r\left(\frac{24}{\delta}\tau_0\right)^s$, where $|\boldsymbol{Q}|$ is its cardinality.*

Next, we will prove the main result by applying the concentration inequality (Definition 2.5(ii)) with union bound. Let $\epsilon = \delta/2$,

$$\left(1 - \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2^2 \le \|\boldsymbol{A}\boldsymbol{q}\|_2^2 \le \left(1 + \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2^2 \quad \forall \boldsymbol{q} \in \boldsymbol{Q}, \tag{E.4}$$

with probability greater than $1 - 2|\boldsymbol{Q}|e^{-pc_0(\delta/2)}$.

Because $\delta \in (0,1)$, we have $1 - \frac{\delta}{2} \le \sqrt{1-\frac{\delta}{2}}$ and $\sqrt{1+\frac{\delta}{2}} \le 1 + \frac{\delta}{2}$. Then, Equation (E.4) can be written as

$$\left(1 - \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2 \le \|\boldsymbol{A}\boldsymbol{q}\|_2 \le \left(1 + \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2 \quad \forall \boldsymbol{q} \in \boldsymbol{Q}, \tag{E.5}$$

with probability greater than $1 - 2|\boldsymbol{Q}|e^{-pc_0(\delta/2)}$.

By the triangle inequality, we have

$$\|\boldsymbol{A}\boldsymbol{y}\|_2 \le \|\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{q})\|_2 + \|\boldsymbol{A}\boldsymbol{q}\|_2. \tag{E.6}$$

Define

$$U = \max_{\boldsymbol{y}\in MS_{r,\boldsymbol{T},\mu,l},\|\boldsymbol{y}\|_2=1}\|\boldsymbol{A}\boldsymbol{y}\|_2, \tag{E.7}$$

which is attainable because $MS_{r,\boldsymbol{T},\mu,l}$ is closed.

Combining Equations (E.5) and (E.6), we have $\forall \boldsymbol{y} \in MS_{r,\boldsymbol{T},\mu,l}$, with $\|\boldsymbol{y}\|_2 = 1$; there exists a $\boldsymbol{q} \in MS_{r,\boldsymbol{T},\mu,l}$, such that $\|\boldsymbol{A}\boldsymbol{y}\|_2 \le \|\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{y})\|_2 + \left(1 + \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2$, with probability greater than $1 - 2|\boldsymbol{Q}|e^{-pc_0(\delta/2)}$.

Because $\boldsymbol{q} - \boldsymbol{y} \in MS_{r,\boldsymbol{T},\mu,l}$, if $\boldsymbol{q} - \boldsymbol{y} = 0$, we have $\|\boldsymbol{A}\boldsymbol{y}\|_2 \le \left(1 + \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2 = 1 + \frac{\delta}{2}$.

If $\boldsymbol{q} - \boldsymbol{y} \ne 0$, we have $\|\boldsymbol{A}\boldsymbol{y}\|_2 \le \left\|\boldsymbol{A}\frac{(\boldsymbol{q}-\boldsymbol{y})}{\|\boldsymbol{q}-\boldsymbol{y}\|_2}\right\|_2\|\boldsymbol{q} - \boldsymbol{y}\|_2 + \left(1 + \frac{\delta}{2}\right)\|\boldsymbol{q}\|_2 \le \frac{\delta}{4}U + 1 + \frac{\delta}{2}$.

Notice that the second inequality comes from Equation (E.7) combined with Q being a $\frac{\delta}{4}$ covering of $MS_{r,\boldsymbol{T},\mu,l}$. In summary, we have $\|\boldsymbol{A}\boldsymbol{y}\|_2 \le \frac{\delta}{4}U + 1 + \frac{\delta}{2}$.

Because U is attainable, according to Equation (E.7), we have $U \le \frac{\delta}{4}U + 1 + \frac{\delta}{2}$. Consequently, we have $U \le 1 + \frac{3}{4-\delta}\delta \le 1 + \delta$ because $\delta < 1$. Therefore, $\|\boldsymbol{A}\boldsymbol{y}\|_2 \le 1 + \delta$, with probability greater than $1 - 2|\boldsymbol{Q}|e^{-pc_0(\delta/2)}$.

Similarly, we can prove that $\|\boldsymbol{A}\boldsymbol{y}\|_2 \ge 1 - \delta$ with probability greater than $1 - 2|\boldsymbol{Q}|e^{-pc_0(\delta/2)}$.

Finally, according to Lemma E.2, we have that

$$1 - 2|\boldsymbol{Q}|e^{-pc_0\left(\frac{\delta}{2}\right)} \ge 1 - 2\left(\frac{24}{\delta}\tau_1\right)^r\left(\frac{24}{\delta}\tau_0\right)^s e^{-pc_0\left(\frac{\delta}{2}\right)}.$$

This finishes the proof.

Appendix F

Proof of Lemma E.1. To prove the result, we first derive a nontrivial upper bound for the inner produce between $\boldsymbol{m}$ and $\boldsymbol{a}$. Let $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be the reduced SVD of $\boldsymbol{B}$; then,

$$\begin{aligned}
|\boldsymbol{m}^{\mathrm{T}}\boldsymbol{a}| = |\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{a}| = |\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{T}\boldsymbol{a}| &= \left|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{T}\sum_i^n a_i\boldsymbol{e}_i\right| \\
&= \left|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\sum_i^n a_i\boldsymbol{U}^{T}\boldsymbol{e}_i\right| \\
&\le \|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\|_2\left\|\sum_i^n a_i\boldsymbol{U}^{T}\boldsymbol{e}_i\right\|_2 \\
&\le \|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\|_2\sum_i^n|a_i|\,\|\boldsymbol{U}^{T}\boldsymbol{e}_i\|_2 \\
&\le \|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\|_2\|\sum_i^n|a_i|\max_{j\in\{1,\dots r\}}\|\boldsymbol{U}^{T}\boldsymbol{e}_j\|_2 \\
&\le_{(a)} \|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\|_2\,\|\boldsymbol{a}\|_1\sqrt{\frac{\mu r}{n}} \\
&\le_{(b)} \sqrt{\frac{\mu rsl}{n}}\|\boldsymbol{m}\|_2\,\|\boldsymbol{a}\|_2,
\end{aligned} \tag{F.1}$$

where inequality (a) follows $\|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\|_2 = \|\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\|_2$ because $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$ and Definition 2.2. Inequality (b) follows from $\|\boldsymbol{a}\|_1 \le \sqrt{ls}\|\boldsymbol{a}\|_2$.

Let $\eta = \sqrt{\frac{\mu r s l}{n}}$ because $\mu < \frac{n}{rsl}$; we have $\eta < 1$ and

$$|\boldsymbol{m}^{\mathrm{T}}\boldsymbol{a}| = \frac{\left|\|\boldsymbol{y}\|_2^2 - \|\boldsymbol{m}\|_2^2 - \|\boldsymbol{a}\|_2^2\right|}{2} \le \eta \|\boldsymbol{m}\|_2 \boldsymbol{a}_2.$$

Therefore, we have

$$\|\boldsymbol{m}\|_2^2 + \|\boldsymbol{a}\|_2^2 - \|\boldsymbol{y}\|_2^2 \le 2\eta \|\boldsymbol{m}\|_2 \|\boldsymbol{a}\|_2.$$

By completing the square, we have

$$\left(\|\boldsymbol{m}\|_2 + \eta \|\boldsymbol{a}\|_2\right)^2 + \left(1 - \eta^2\right)\|\boldsymbol{a}\|_2^2 - \|\boldsymbol{y}\|_2^2 \le 0.$$

Because $\left(\|\boldsymbol{m}\|_2 + \eta\|\boldsymbol{a}\|_2\right)^2 \ge 0$, we have

$$\|\boldsymbol{a}\|_2 \le \frac{1}{\sqrt{\left(1-\eta^2\right)}}\|\boldsymbol{y}\|_2.$$

Similarly, we can derive that

$$\|\boldsymbol{m}\|_2 \le \frac{1}{\sqrt{\left(1-\eta^2\right)}}\|\boldsymbol{y}\|_2.$$

This finishes the proof.

Appendix G

Proof of Lemma E.2. We first state results for the covering number of a set (Vershynin 2018). The covering number of a smallest ϵ net for a unit l_2 norm ball in d-dimensional space is $(3/\epsilon)^d$.

Let $\mathbf{M} = \left\{\boldsymbol{m} \in \mathbb{R}^n \,|\, \boldsymbol{m} = \boldsymbol{B}\boldsymbol{\theta},\ \boldsymbol{\theta} \in \mathbb{R}^r,\ \|\boldsymbol{\theta}\|_2 \le \tau_1, \mu(\boldsymbol{B}) = \mu\right\}$ and $\mathbf{S} = \{\boldsymbol{a} \in \mathbb{R}^n \,|\, \boldsymbol{a} = \boldsymbol{B}_a\boldsymbol{\theta}_a, \boldsymbol{\theta}_a \in \mathbf{T}, \|\boldsymbol{\theta}_a\|_2 \le \tau_0, l(\boldsymbol{B}_a) = l\}$. There exist two finite $\frac{\delta}{8}$ covering sets of $\mathbf{M}$ and $\mathbf{S}$, which are $\mathbf{Q_M} \subseteq \mathbf{M}$ and $\mathbf{Q_S} \subseteq \mathbf{S}$.

For all $\boldsymbol{q}_\mathbf{M} \in \mathbf{Q_M}$ and for all $\boldsymbol{m} \in \mathbf{M}$, we have

$$\min_{\boldsymbol{q}_\mathbf{M} \in \mathbf{Q_M}} \|\boldsymbol{m} - \boldsymbol{q}_\mathbf{M}\|_2 \le \frac{\delta}{8};$$

for all $\boldsymbol{q}_\mathbf{S} \in \mathbf{Q_s}$ and for all $\boldsymbol{a} \in \mathbf{S}$, we have

$$\min_{\boldsymbol{q}_\mathbf{S} \in \mathbf{Q}_\mathbf{s}} \|\boldsymbol{a} - \boldsymbol{q}_\mathbf{S}\|_2 \le \frac{\delta}{8}.$$

Therefore, we have $|\mathbf{Q_M}| \le \left(\frac{24}{\delta}\tau_1\right)^r$ and $|\mathbf{Q_s}| \le \left(\frac{24}{\delta}\tau_0\right)^s$.

Define $\mathbf{Q_{MS}} = \left\{\boldsymbol{q}_\mathbf{M} + \boldsymbol{q}_\mathbf{S} \,|\, \boldsymbol{q}_\mathbf{M} \in \mathbf{Q_M},\ \boldsymbol{q}_\mathbf{S} \in \mathbf{Q_S}\right\} \subseteq MS_{r,\mathbf{T},\mu,l}$. Then, $\forall \boldsymbol{y} \in MS_{r,\mathbf{T},\mu,l}$, there exists a pair $\boldsymbol{q}_\mathbf{MS} = \boldsymbol{q}_\mathbf{M} + \boldsymbol{q}_\mathbf{S} \in MS_{r,\mathbf{T},\mu,l}$, such that

$$\|\boldsymbol{q}_\mathbf{MS} - \boldsymbol{y}\|_2 = \|\boldsymbol{q}_\mathbf{M} - \boldsymbol{m} + \boldsymbol{q}_\mathbf{S} - \boldsymbol{a}\|_2 \le \|\boldsymbol{q}_\mathbf{M} - \boldsymbol{m}\|_2 + \|\boldsymbol{q}_\mathbf{S} - \boldsymbol{a}\|_2 \le \frac{\delta}{4}.$$

Therefore, $\mathbf{Q_{MS}}$ is a $\delta/4$ covering of $MS_{r,\mathbf{T},\mu,l}$ and $|\mathbf{Q_{MS}}| \le \left(\frac{24}{\delta}\tau_1\right)^r \left(\frac{24}{\delta}\tau_0\right)^s$. This finishes the proof.

Appendix H. Simulation Study by Varying the Magnitude of Sparse Signals

We adopt the same simulation data generation procedure as in Section 3.1 while changing the distribution of elements in $\boldsymbol{\theta}_a \in \mathbb{R}^q$ such that the nonzero elements follow i.i.d. normal distribution with mean zero and standard deviation σ_s in the range of {0.065, 0.125, 0.25, 0.5}. The example signals and reconstruction performance are shown in Figure H.1.

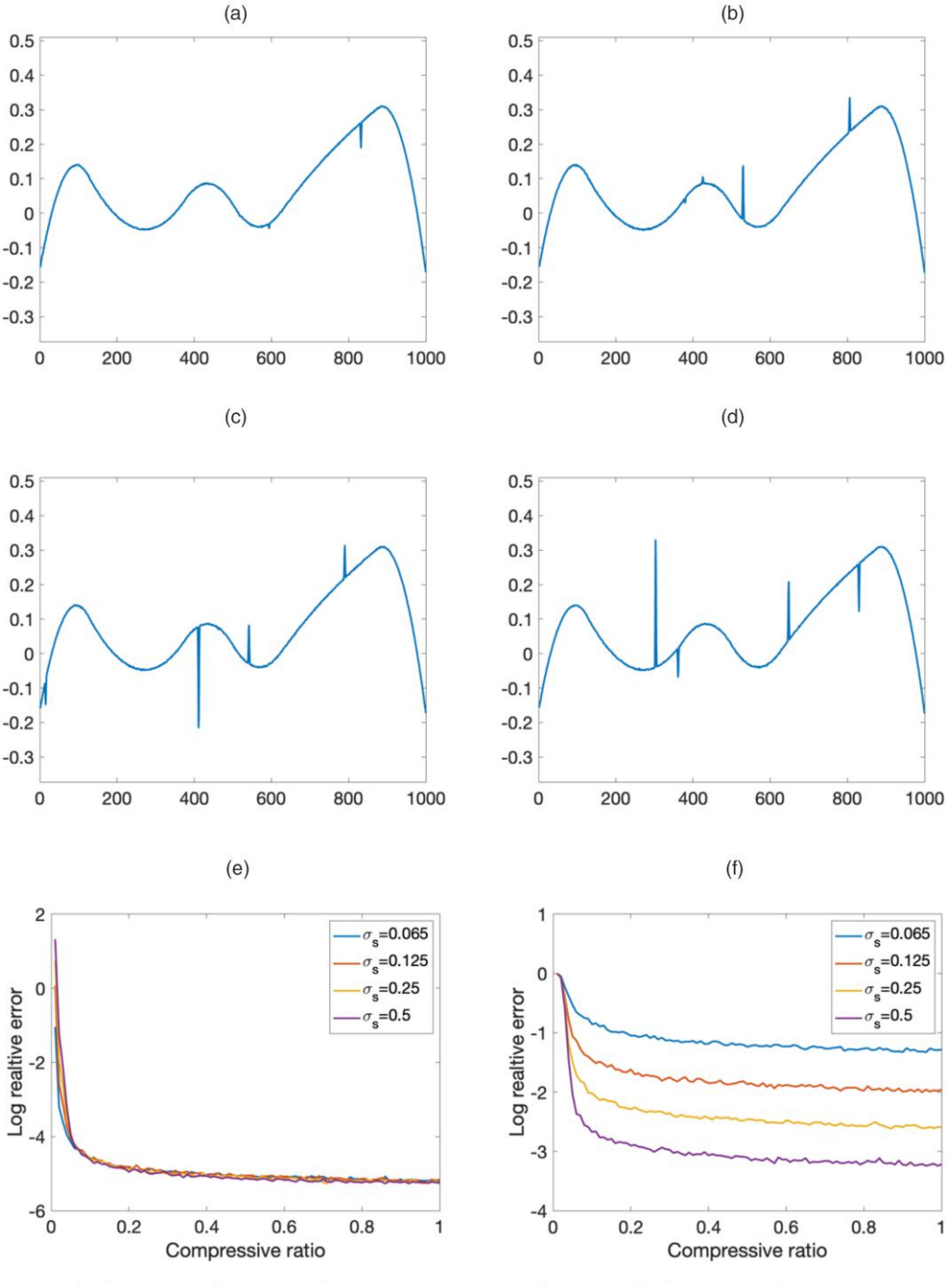
There are several observations.

1. The log relative error of the smooth component does not change with different magnitudes of σ_s. This agrees with the theoretical result in Theorem 2.4, where the reconstruction error is bounded by a constant times the noise bound because the noise level is kept the same in all simulations.

2. The log relative error of the sparse component decreases as σ_s increases. This also agrees with the theoretical result in Theorem 2.4. Because the log relative error of the sparse component is defined as $\log\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|_2/\|\boldsymbol{a}\|_2$, according to Theorem 2.4, the reconstruction error term can be approximated by $C_a\epsilon_1$ (i.e., $\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|_2 \sim C_a\epsilon_1$, where C_a is independent of $\boldsymbol{\theta}_a$). $\|\boldsymbol{a}\|_2 = \|\boldsymbol{B}_a\boldsymbol{\theta}_a\|_2$ increases as the magnitude of elements in $\boldsymbol{\theta}_a$ increases. Therefore, $\log\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|_2/\|\boldsymbol{a}\|_2$ will decrease as σ_s increases.

3. The 0.1 threshold (approximately) of the compressive ratio still holds with different magnitudes of sparse signal component.

Figure H.1. (Color online) Simulation Study by Varying σ_s



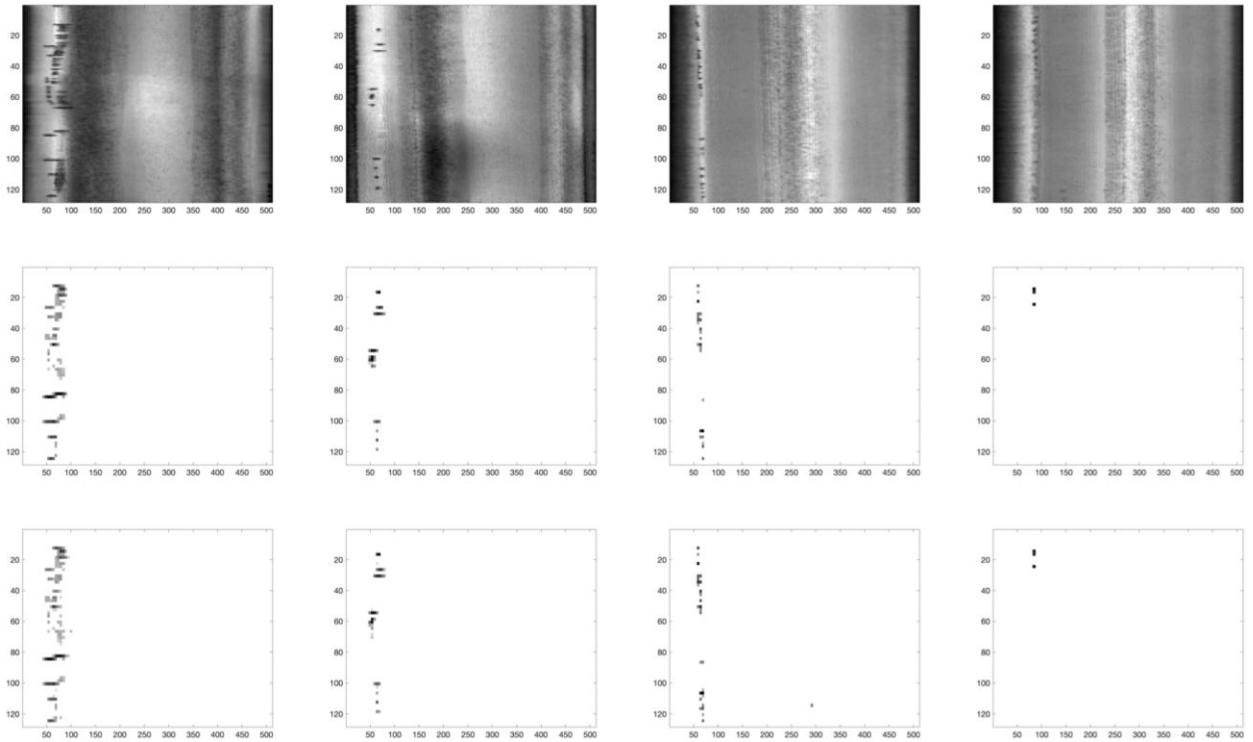
Notes. (a)–(d) Example signals corresponding to different σ_s, (e) log relative error for the smooth component, and (f) log relative error for the sparse component. (a) $\sigma_s = 0.065$. (b) $\sigma_s = 0.125$. (c) $\sigma_s = 0.25$. (d) $\sigma_s = 0.5$. (e) Log relative error for the smooth component. (f) Log relative error for the sparse component

Appendix I. Sample Case Study Images

Appendix I.1. Sample Images of Case Study 4.1

See Figure I.1.

Figure I.1. Sample Images of Surface Defect Detection in the Steel Rolling Process

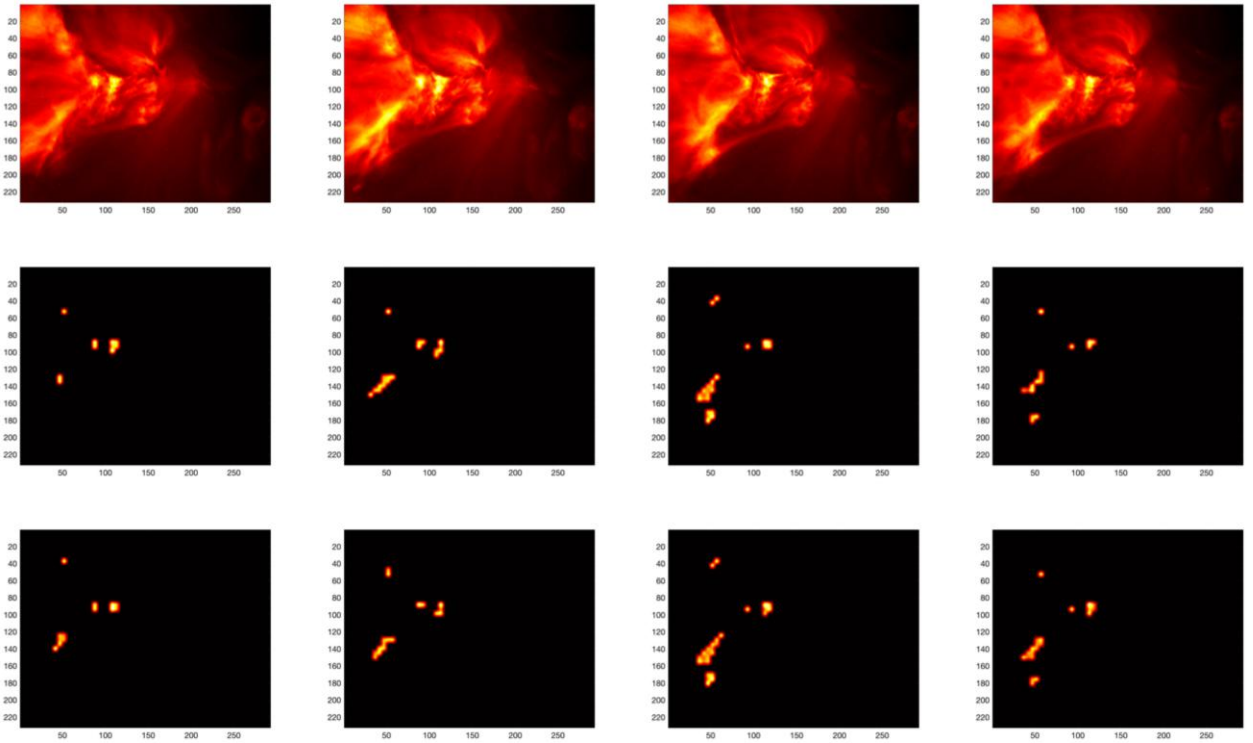


Notes. The top row shows raw images. The middle row shows corresponding SSD results. The bottom row shows KronCSSD results.

Appendix I.2. Sample Images of Case Study 4.2

See Figure I.2.

Figure I.2. (Color online) Sample Images of Solar Flare Detection



Notes. The top row shows raw images. The middle row shows corresponding SSD results. The bottom row shows KronCSSD results.

References

Augusto CRA, Fauth AC, Navia CE, Shigeouka H, Tsui KH (2011) Connection among spacecrafts and ground level observations of small solar transient events. *Experiment. Astronomy* 31(2):177–197.

Baraniuk R, Davenport M, DeVore R, Wakin M (2008) A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* 28(3):253–263.

Bouwmans T, Zahzah EH (2014) Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. *Comput. Vision Image Understanding* 122:22–34.

Candes EJ (2008) The restricted isometry property and its implications for compressed sensing. *Competus Rendus Math.* 346(9–10):589–592.

Candes EJ, Tao T (2005) Decoding by linear programming. *IEEE Trans. Inform. Theory* 51(12):4203–4215.

Candes EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* 59(8):1207–1223.

Candès EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? *J. ACM* 58(3):1–37.

De Boor C, De Boor C (1978) *A Practical Guide to Splines*, vol. 27 (Springer-Verlag, New York).

Duarte MF, Baraniuk RG (2011) Kronecker compressive sensing. *IEEE Trans. Image Processing* 21(2):494–504.

Eilers PH, Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.* 11(2):89–121.

Giannakis GB, Mateos G, Farahmand S, Kekatos V, Zhu H (2011) USPACOR: Universal sparsity-controlling outlier rejection. Tichavsky P, Cernocky H, Prochazka A, eds. *2011 IEEE Internat. Conf. Acoustics Speech Signal Processing* (IEEE, New York), 1952–1955.

Kolda TG, Bader BW (2009) Tensor decompositions and applications. *SIAM Rev.* 51(3):455–500.

Mardani M, Mateos G, Giannakis GB (2013) Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies. *IEEE Trans. Inform. Theory* 59(8):5186–5205.

Marques EC, Maciel N, Naviner L, Cai H, Yang J (2018) A review of sparse recovery algorithms. *IEEE Access* 7:1300–1322.

Mateos G, Giannakis GB (2011) Robust nonparametric regression by controlling sparsity. *2011 IEEE Internat. Conf. Acoustics Speech Signal Processing (ICASSP)*.

Minaee S, Abdolrashidi A, Wang Y (2015) Screen content image segmentation using sparse-smooth decomposition. *2015 49th Asilomar Conf. Signals Systems Comput*.

Mou S, Wang A, Zhang C, Shi J (2021) Additive tensor decomposition considering structural data information. *IEEE Trans. Automation Sci. Engrg.* 19(4):2904–2917.

Rani M, Dhok SB, Deshmukh RB (2018) A systematic review of compressive sensing: Concepts, implementations and applications. *IEEE Access* 6:4875–4894.

Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* 52(3):471–501.

Tanner J, Vary S (2020) Compressed sensing of low-rank plus sparse matrices. Preprint, submitted July 18, https://arxiv.org/abs/2007.09457v1.

Unser M (1999) Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine* 16(6):22–38.

Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47 (Cambridge University Press, Cambridge, United Kingdom).

Wang A, Xian X, Tsung F, Liu K (2018) A spatial-adaptive sampling procedure for online monitoring of big data streams. *J. Quality Tech.* 50(4):329–343.

Waters AE, Sankaranarayanan AC, Baraniuk RG (2011) SpaRCS: Recovering low-rank and sparse matrices from compressive measurements. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 1089–1097.

Xu H, Caramanis C, Sanghavi S (2012) Robust PCA via outlier pursuit. *IEEE Trans. Inform. Theory* 58(5):3047–3064.

Yan H, Paynabar K, Shi J (2017) Anomaly detection in images with smooth background via smooth-sparse decomposition. *Technometrics* 59(1):102–114.

Yan H, Paynabar K, Shi J (2018) Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics* 60(2):181–197.