Electronic Journal of Statistics
Vol. 7 (2013) 2098–2130
ISSN: 1935-7524
DOI: 10.1214/13-EJS840

Computationally efficient estimators for sequential and resolution-limited inverse problems

Darren Homrighausen

Department of Statistics
Colorado State University
Fort Collins, CO 80523
e-mail: darrenho@stat.colostate.edu

and

Christopher R. Genovese

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15223
e-mail: genovese@stat.cmu.edu

Abstract: A common problem in the sciences is that a signal of interest is observed only indirectly, through smooth functionals of the signal whose values are then obscured by noise. In such *inverse problems*, the functionals dampen or entirely eliminate some of the signal's interesting features. This makes it difficult or even impossible to fully reconstruct the signal, even without noise. In this paper, we develop methods for handling *sequences* of related inverse problems, with the problems varying either systematically or randomly over time. Such sequences often arise with automated data collection systems, like the data pipelines of large astronomical instruments such as the Large Synoptic Survey Telescope (LSST). The LSST will observe each patch of the sky many times over its lifetime under varying conditions. A possible additional complication in these problems is that the observational resolution is limited by the instrument, so that even with many repeated observations, only an approximation of the underlying signal can be reconstructed. We propose an efficient estimator for reconstructing a signal of interest given a sequence of related, resolution-limited inverse problems. We demonstrate our method's effectiveness in some representative examples and provide theoretical support for its adoption.

Keywords and phrases: Deconvolution, signal processing, complex Gaussian.

Received June 2012.

1. Introduction

In many applications, a signal of interest can only be indirectly observed. Examples of such *inverse problems* include astronomical imaging from ground-based telescopes, where atmospheric turbulence and instrument effects blur the images; Positron Emission Tomography, where measured photon intensities are averages over lines; and seismic reconstruction, where the measured quake effects, observed at the surface, represent the integrated properties of the seismic waves along their path through the Earth. As in these examples, the basic measurements in an inverse problem are smooth functionals of the parameter of interest that dampen or entirely eliminate many interesting features, making it difficult to estimate those features from noisy data.

There is a long and rich literature on estimation methods for inverse problems. We cannot hope to provide a comprehensive list, but see O'Sullivan (1986); Wahba (1990); Donoho (1995); Tenorio (2001); Candés and Donoho (2002); Cavalier et al. (2002) and the references contained therein for an introduction and Cavalier (2008) for a modern review of the state of the field. In addition, many specialized techniques have been developed for particular classes of inverse problems that arise frequently in specific applications, such as astronomy (Starck, Pantin and Murtagh, 2002; van Dyk et al., 2006), geophysics (Backus and Gilbert, 1968), and tomography (Ólafsson and Quinto, 2005).

One implicit assumption of most existing work on inverse problems is that the effective resolution of the observational device increases as more data is gathered. In this paper, we consider a different regime in which we seek to recover a signal given a sequence of related, but varying, inverse problems. Recent technological advances allowing automated data-collection have revealed situations in which this scenario occurs. All of the data is collected by the same device and hence the resolution of the entire sequence of observations is fixed. However, the conditions under which the data is collected varies.

Our motivating example is image reconstruction by the Large Synoptic Survey Telescope (LSST), a multi-year, Earth-based astronomical survey of the entire sky. The LSST will produce images of the sky at unprecedented depth, eventually cataloging billions of astronomical objects. Each patch of sky will be repeatedly observed over the lifetime of the instrument, roughly once every 3–4 days. The goal is to obtain an accurate estimate of the underlying scene, as a baseline for detecting transient phenomena and answering other scientific questions. While these repeated observations can improve accuracy, there are potentially significant variations across images in exact position, orientation, and atmospheric turbulence ("seeing"). This variation in the underlying inverse problem complicates estimation, for instance, making simple averaging tend to perform poorly. There are two other important features of LSST data. First, the spatial resolution of the images remains fixed throughout, which fundamentally limits the resolution of any reconstruction even with an arbitrarily long period of data collection. Second, the data are collected and must be processed in near real-time with limited access to past observations. (All the data are stored but are not all available in a timely fashion.) Thus an important consideration for an estimation procedure here is that it be computable in an on-line manner, using a small store to quickly update the estimate with each new image obtained.

More generally, we consider the following problem. We want to recover an unknown signal/parameter $\theta \in \mathbb{R}^p$ from indirect measurements of the form

$$\mathbf{Y}_i = K_i\theta + \sigma\mathbf{W}_i, \quad \text{for} \quad i = 1, 2, \ldots. \tag{1}$$

Here, each $\mathbf{Y}_i$ is a measured signal, such as an audio recording or a (vectorized) image. Each *forward operator* K_i embodies the indirectness in the measurement process, and typically acts as a smoothing operator in some sense. In addition, we assume that K_i also captures the discretization (e.g., pixelization) due to the resolution-limited nature of the instruments, which is assumed fixed however many repeated observations are obtained. Finally, the $\mathbf{W}_i$'s represent stochastic noise in the observations. We take them here to be independent, mean zero Gaussian p-vectors with variance-covariance matrix I_p, the order p identity. We begin by assuming a known noise level. In this case, if the noise level σ_i varies across observations, we can simply rescale the problem to a constant noise level, so without loss of generality we take σ to be constant. For some applications, like astronomical imaging, the assumption of known noise level can be practically reasonable as there are often reliable noise level estimates available, but we consider the problem of estimating an unknown noise variance in Section 3.2.

The goal of this paper is to develop an effective estimator of θ from a sequence of varying, resolution-limited inverse problems of the form in equation (1). We develop an estimator that has many favorable features, such as (i) all tuning parameters must be selected in a data-dependent way, (ii) the resulting procedure (including the choice of tuning parameters) has good statistical performance, and (iii) the estimator $\hat{\theta}_n$ based on an n-sequence can be efficiently updated to produce the estimator $\hat{\theta}_{n+1}$ after observing $\mathbf{Y}_{n+1}$. Such an estimator is novel in the literature. Additionally, we show that this estimator has good theoretical properties and relies on weaker assumptions than the previous literature. Lastly, in analyzing our estimator, we develop some theory for estimation of the mean of a complex Gaussian random variable, which, to our knowledge, is new in that literature.

To clarify the key ideas, we begin with a brief example highlighting the sequential and resolution-limited aspects of the inverse problem. During satellite imaging operations, a location on Earth is imaged many times over the life span of the satellite. The quality of the recorded observations can be low and variable due to changing atmospheric and/or weather conditions giving a sequential problem. Additionally, the pixelization induced by the instrument is fixed, no matter how many repeated observations are observed. Hence, for each i, K_i is a mapping represents both the effects of the atmosphere and pixelization.

The left column of Figure 1 offers a representative panel of four such images taken of the White House and surrounding buildings. Note that the amount of blurring, corresponding to the forward operators K_i, can vary quite widely. Our proposed estimator $\hat{\theta}_n$ builds an estimate of the unknown scene θ via sequential updating. After each observation in the left-hand column of Figure 1, we update the previous estimate, which is given in the right-hand column in the same row. Within a few observations the reconstruction has become quite sharp, but the gain from each observation depends strongly on its level of blurring. We emphasize that all tuning parameters are chosen in an automatic, data-dependent way throughout this procedure.

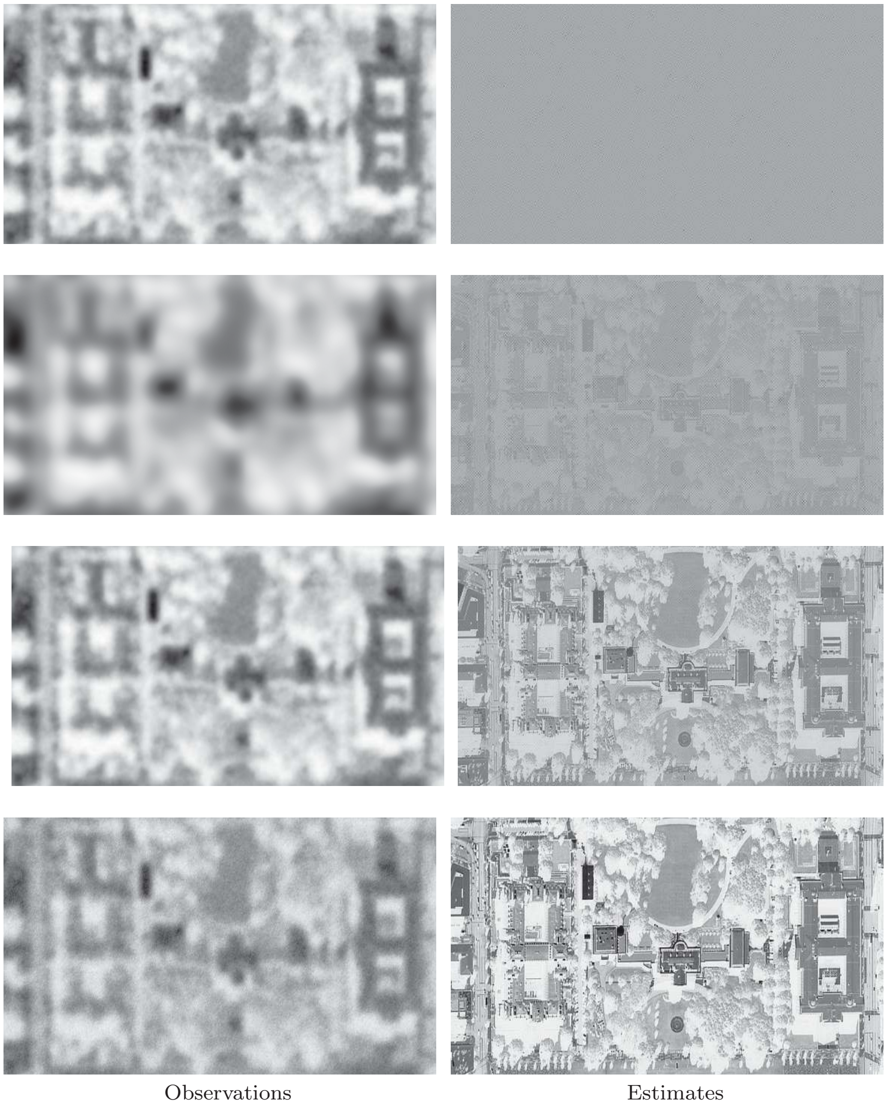


FIG 1. *Example of images of the White House from a satellite and associated recovery of the unknown signal θ using our proposed estimator. In the left column (Observations) the different amounts of blurring are due to varying atmospheric conditions and correspond to the forward operators K_i in equation* (1). *In the right column (Estimates) we report the output of our estimator using the data in the left column. Each row corresponds to making another observation Y_i and updating our estimator with this new data.*

Sequential inverse problems have been analyzed in the literature. In work on analog-to-digital conversion, the recovery of the original analog signal is an inverse problem as there is not a unique analog signal corresponding to each digital signal. If the signal is instead sampled multiple times at different, carefully chosen sampling rates, Berenstein and Patrick (1990) and Casey and Walnut (1994) find conditions under which the original signal can be reconstructed in a loss-less way. Note that, as opposed to our paper, these approaches deal with only the case where $\sigma = 0$ and the K_i, which correspond to the sampling rate, can be chosen by the experimenter. Subsequent work on sequential inverse problems, beginning with Piana and Bertero (1996), develop two methods. The first adapts Tikonov-Phillips regularization (known in statistics as ridge regression) to the sequential context. The second is a method based on Landwieber iterations (see Bertero and Boccacci (1998) for an overview of Landwieber iterations). These methods have been implemented in the software package AIRY (Bertero and Boccacci, 2000a,b; Correia et al., 2002).

Our method has two major advantages over these previous proposals. First, the diagonalization condition we make (assumption (A3) in Section 2) contains as a special case the convolutional assumption which is standard in the field. More importantly, our method comes with a data-driven method for choosing the tuning parameter. This is extremely important in many applications, where completely automated estimation is crucial.

This paper is organized as follows. We outline our method in Section 2, discuss computational considerations and explore strategies for estimating the noise level in Section 3. Lastly, we describe the results of a simulation study in Section 4. We defer all proofs to the appendix.

Notation. In the development of our technique, we use complex-valued vectors and random variables, so it may be useful to clarify the notation we use. For $a \in \mathbb{C}$ and $A \in \mathbb{C}^{p \times q}$, define a^* and A^* to be the Hermitian adjoint of a and A, respectively. Correspondingly, define $|a|^2 = a^*a$ and $|A|^2 = A^*A$ to be the squared complex modulus of a scalar and matrix, respectively. Likewise, for any vector $x \in \mathbb{C}^p$, $||x||^2 = x^*x$. If $AA^* = I_p = A^*A$, then we say that A is unitary. We utilize a bold faced font for vectors, $\mathbf{b}_n \in \mathbb{C}^p$, and denote its j^{th} entry as b_{nj}, where the subscript n indicates dependence on the sample size. Similarly, A_{nj} is the j^{th} element of the main diagonal of the matrix A_n. We abuse notation slightly by using $\boldsymbol{\lambda}$ as both a vector in $\mathbb{C}^p$ and as a function from $\mathbb{C}^p$ to $\mathbb{C}^p$ given by component-wise multiplication.

2. Methodology

The model introduced in equation (1) reflects that the observations we gather from modern scientific devices are often sequential, noisy, and blurred from different forward operators. Our methodology leverages an equivalence between certain families of these forward operators that allows for the data to be transformed to a common coordinate system whereby an estimator can be developed via minimizing an estimate of the risk.

We begin this section by stating the following assumptions under which our method can be derived.

(A1) The noise parameter $\sigma > 0$ is known.
(A2) The $(K_i)_{i=1}^n$ are known smoothing matrices.

(A3) There exists a unitary matrix $\Psi \in \mathbb{C}^{p \times p}$ and diagonal matrices D_i such that $K_i = \Psi D_i \Psi^*$ for all $i = 1, \dots, n, \dots$

(A4) There exists an $N < \infty$ such that for all j there exists an $1 \leq i_* \leq N$ such that $|D_{i_* j}| > 0$.

(A5) Define $\Delta_{nj} := \sum_{i=1}^{n} |D_{ij}|^2$. Then the (D_i) are such that

$$\lim_{n \to \infty} \frac{\max_j \Delta_{nj}}{\min_j \Delta_{nj}} < \infty.$$

Assumptions (A1) and (A2) are very standard in the statistical inverse problem literature. We discuss a strategy for estimating σ in Section 3.2. Assumption (A4) is also commonly made and it ensures that, at some point, the entire signal θ is identified and loosely corresponds to the intersection of the null spaces of the $(K_i)_{i=1}^{n}$ eventually only containing the zero vector. Assumption (A5) merely prevents a pathological case where the K_i are becoming more ill-conditioned without bound as $n \to \infty$. Assumption (A3) is crucial to our method and while the reason for it will become clear, the following theorem provides a general family of matrices that satisfy it.

Theorem 1. *If the $(K_i)_{i=1}^{n}$ all correspond to the discrete convolution operation, then there exists a unitary matrix Ψ and a sequence of diagonal matrices $(D_i)_{i=1}^{n}$, all of which could have complex entries, such that (A3) holds. If θ is a one (two)-dimensional signal, then the K_i are (block) circulant and the entries of the matrix Ψ are the discrete one (two)-dimensional Fourier basis and the entries of D_i are the corresponding discrete one (two)-dimensional Fourier coefficients.*

Hence, we see that assumption (A3) is more general than the convolutional assumption made in Piana and Bertero (1996) and many other works concerning statistical inverse problems. See Appendix A for a proof of Theorem 1.

2.1. Overview and main results

An overview of our procedure is as follows. The parameter θ and each observation $\mathbf{Y}_i$ is rotated by Ψ^*. The rotated $\mathbf{Y}_i$'s are combined together to form a sufficient statistic $\mathbf{B}_n$. The estimators we consider are of the form $\hat{\theta} = \Psi \boldsymbol{\lambda}(\mathbf{B}_n) := \Psi(\lambda_j B_{nj})_{j=1}^{p}$. Define this set of estimators to be

$$\mathcal{E} = \{\hat{\theta} = \Psi \boldsymbol{\lambda}(\mathbf{B}_n) : \boldsymbol{\lambda} \in \mathbb{C}^p\}. \tag{2}$$

We choose from the estimators in $\mathcal{E}$ using a combination of minimizing an empirical estimator of the risk and adding some additional regularization. We notate our estimator as $\hat{\theta}_n = \Psi \hat{\boldsymbol{\lambda}}(B_n)$ (distinguishing it from a generic estimator $\hat{\theta}$ by the subscript n), where the weights $\hat{\boldsymbol{\lambda}}$ are defined in the text containing and preceding equation (12). Additionally, we define our loss function to be the l^2 norm with associated risk

$$R(\hat{\theta}, \theta) := \mathbb{E}||\hat{\theta} - \theta||^2 \tag{3}$$

and set $\Theta := \{\theta : ||\theta||_2^2 \leq T^2\}$ for any $0 < T^2 < \infty$ as the parameter space. Then

Theorem 2. *Under assumptions (A1)–(A5),*

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \gamma_n^{-1} R\left(\hat{\theta}_n, \theta\right) < \infty$$

where

$$\gamma_n = \min_j \frac{\sigma^2}{\Delta_{nj}}.$$

Remark 2.1. If all $D_{ij} \equiv D_j$ for some $D_j \in \mathbb{C}$, then $\gamma_n \asymp 1/n$; that is the parametric rate. However, the forward operators (K_i) in effect ensure that each observation doesn't decrease the risk equally. The quantity Δ_{nj} relates to how much information is present in the first n observations about the j^{th} component of $\Psi^*\theta$.

Additionally, we compare our estimator to the $\mathcal{E}$-oracle with the following result.

Theorem 3. *Suppose assumptions (A1)–(A5) and let*

$$R_*(\theta) := \min_{\hat{\theta}\in\mathcal{E}} R(\hat{\theta}, \theta)$$

be the risk of the $\mathcal{E}$-oracle. Then

$$R\left(\hat{\theta}_n, \theta\right) \leq R_*(\theta)(1 + O(1)),$$

where the term $O(1)$ does not depend on θ.

An interesting extension of this model is to when the (K_i) are considered random. We answer this question in an interesting case.

Random eigenvalues Suppose that the (K_i) are random operators such that $K_i = \Psi D_i \Psi^*$ for all $i = 1, 2, \ldots$ and $\text{diag}(D_i) \overset{i.i.d}{\sim} \mathcal{D}$, where $\mathcal{D}$ is any p-variate complex distribution that doesn't have too much mass near zero. Specifically,

(B4) The distribution $\mathcal{D}$ is such that there exists an a, τ, and $\rho > 1$, where for $0 \leq \tau \leq a$
$$\mathbb{P}_{\mathcal{D}}\left(|D_{1j}|^2 < \tau\right) = (\tau)^{\rho}.$$

This is a stochastic extension of assumption (A4) as it allows the random eigenvalues to be arbitrarily close to zero in magnitude but with the probability of them being small going to zero. Lastly, let (W_i) and (D_i) be mutually independent.

Theorem 4. *Suppose assumption (B4) holds. Then*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{E}_{(D_i),(Y_{ij})} \left|\left|\hat{\theta}_n - \theta\right|\right|^2 = 0,$$

where $\mathbb{E}_{(D_i),(Y_{ij})}$ corresponds to integration with respect to the joint distribution of (D_i) and (Y_{ij}).

2.2. Rotations and degenerate complex Gaussians

Returning to equation (1) and using assumption (A3), for $i = 1, 2, \dots$ we define $\mathbf{X}_i := \Psi^* \mathbf{Y}_i$, $\beta := \Psi^* \theta$, and $\mathbf{U}_i := \Psi^* \mathbf{W}_i$. Then it follows that

$$\mathbf{X}_i = D_i \beta + \sigma \mathbf{U}_i. \tag{4}$$

Note that in this case $\mathbf{U}_i \overset{i.i.d}{\sim} CN(0, I_p, \Psi^* \overline{\Psi})$[1]. It is also convenient to look at equation (4) component-wise,

$$X_{ij} = D_{ij} \beta_j + \sigma U_{ij} \tag{5}$$

for $j = 1, \dots, p$. Note that for these multiplications to be defined, we have to think about $\mathbb{R}^p$ being embedded in $\mathbb{C}^p$ by having imaginary part equal to zero. We follow this convention without comment in what follows.

Remark 2.2. Commonly, the sequence space formulations found in equations (4) and (5) are accomplished by a real, orthogonal matrix instead of a complex, unitary one. Allowing for the sequence $(K_i)_{i=1}^n$ to share the same eigenvectors necessitates permitting Ψ to be complex.

We can rearrange equation (5) to define

$$B_{nj} := \sum_{i=1}^{n} \frac{D_{ij}^* X_{ij}}{\Delta_{nj}}, \tag{6}$$

where $\Delta_{nj} := \sum_{i=1}^n |D_{ij}|^2$. Then $\mathbf{B}_n := (B_{nj})_{j=1}^p$ is distributed

$$\mathbf{B}_n \sim CN\left(\beta, \sigma^2 \Delta_n^{-1}, \sigma^2 \Delta_n^{-2} \sum_{i=1}^{n} D_i^* \Psi^* \overline{\Psi} D_i^* \right). \tag{7}$$

To develop an automatic procedure for signal estimation in sequential inverse problems, we begin by regularizing the unbiased estimator $\mathbf{B}_n$ of β through the use of a tuning parameter vector. We choose this tuning parameter by minimizing an estimator of the risk. This type of procedure, known generally as unbiased risk estimation, has been revisited regularly in many fields for solving various problems related to denoising (Stein, 1981; Donoho and Johnstone, 1995, for example). However, as inverse problems generally result in unstable estimators of both the parameter β and the risk R, we compensate by including additional regularization.

The formulation in equation (7) is related to the (real-valued) normal means problem. In particular, a p-dimensional complex normal can be thought of $2p$-dimensional real Gaussian with a complicated covariance matrix. This real-valued Gaussian vector actually only has values in a p-dimensional subspace of

[1] A complex normal has an extra parameter compared with a real normal. For a zero mean complex normal random variable $\mathbf{U}$, this is denoted $CN(0, \mathbb{E}\mathbf{U}\mathbf{U}^*, \mathbb{E}\mathbf{U}\mathbf{U}^\top)$.

$\mathbb{R}^{2p}$ and hence is degenerate. Therefore, we choose to analyze the sequence $(\mathbf{B}_n)$ as complex random Gaussian.

Complex Gaussian random variables have been studied in statistics and related fields for many years (see Wooding (1956); Goodman (1963) for early papers). However, most of the literature, e.g. Gallager (2008), focuses on the *circularly symmetric* case; that is, when $\Psi^*\overline{\Psi}$ is diagonal. In this case, the real-valued version is no longer degenerate and is hence much easier to manipulate. However, this assumption does not hold for our purposes. More modern literature, such as Schreier and Scharf (2003); Schreier, Scharf and Mullis (2005), consider the more general case of *improper* complex Gaussians; that is when $\Psi^*\overline{\Psi}$ is not diagonal. However, these works consider forming minimum variance unbiased estimators, which is not appropriate for our purposes.

Additionally, owing to the presence of the $D_i's$, this model is heteroskedastic, with complex variance parameter. In general, considering the heteroskedastic model leads to a much more involved theory than in the homoscedastic case, such as in Brown (1975), and is still the topic of contemporary research (Brown, Nie and Xie, 2012).

Remark 2.3. Note that our approach is fundamentally different from attempting to estimate β using $D_i^{-1}X_i$. First, assuming D_i^{-1} exists for all i necessitates a stronger assumption than (A4). More importantly, if Y_i is an extremely low quality observation, then $|D_{ij}|$ is very close to zero for some j. In this case, suppose we are estimating β with the linear estimator $\sum_i a_i D_i^{-1}X_i$, for constants a_i. Then the variance of this estimator is extremely high (due to inverting the very small elements of D_i). Compare this to estimators based on B_{nj}, where these low quality observations do not have any negative effect.

2.3. Estimators and tuning parameter selection

The specifics of our approach are related to the procedure found in Beran (2000). However, the goal in Beran (2000), unlike our paper, is the estimation of the regression function in an assumed linear model instead of the coefficients themselves. That is, referring to the notation in equation (1), the estimation of $K_i\theta$ instead of the estimation of θ. This is an important distinction as both estimating θ is intrinsically harder than estimating $K_i\theta$ and θ is the object of actual interest. Also, the theoretical justification that appears in Beran (2000) is essentially entirely asymptotic in p. This is a regime we do not consider relevant for the problem at hand.

As Ψ is unitary, we can define an equivalent risk to the one defined in equation (3) in terms of β

$$R(\hat{\theta},\theta) := \mathbb{E}||\hat{\theta}-\theta||^2 = \mathbb{E}||\Psi^*(\hat{\theta}-\theta)||^2 = \mathbb{E}||\hat{\beta}-\beta||^2 =: R(\hat{\beta},\beta). \tag{8}$$

Under the model introduced in equation (1) and assumptions (A1)–(A4), the random vector $\mathbf{B}_n$ is sufficient for β in equation (4). This claim can be seen by noting that the map Ψ^* is measure preserving.

Any risk computations made under the data, which is $(X_i)_{i=1}^n$ in our notation, are equivalent to those made under a sufficient statistic (Bahadur, 1954, Theorem 7.1). Hence, the expectations in equation (8) are equal whether under $(X_i)_{i=1}^n$ or $\mathbf{B}_n$, which implies that for each n, we can treat $\mathbf{B}_n$ as the data.

To begin to formulate an estimator of β, and therefore θ, we use the following result.

Proposition 5. *Define $\hat{\psi}_j := (|B_{nj}|^2 - \sigma^2/\Delta_{nj})/|B_{nj}|^2$. Then the random function*

$$\hat{R}_n(\boldsymbol{\lambda}) := \sum_{j=1}^{p} (\lambda_j - \hat{\psi}_j)^2 |B_{nj}|^2 \tag{9}$$

provides, up to a constant independent of $\boldsymbol{\lambda}$, an unbiased estimate of $R(\boldsymbol{\lambda})$. Additionally,

$$\min_{\lambda \in \mathbb{C}^p} R(\boldsymbol{\lambda}) = \min_{\lambda \in \mathcal{L}} R(\boldsymbol{\lambda})$$

where $\mathcal{L} = [0,1]^p$ is the p-dimensional hypersquare.

The first part of the proposition provides an unbiased estimate of the risk while the second part implies that we gain no improvement in risk by allowing $\boldsymbol{\lambda}$ to have values outside of $\mathcal{L}$.

Using $\hat{R}_n$ from equation (9), define for any $\mathcal{G} \subseteq \mathcal{L}$

$$\hat{\boldsymbol{\lambda}}^{\mathcal{G}} := \operatorname*{argmin}_{\mathcal{G}} \hat{R}_n(\boldsymbol{\lambda}), \tag{10}$$

which produces an estimator of β, $\hat{\beta}^{\mathcal{G}} := \hat{\boldsymbol{\lambda}}^{\mathcal{G}}(\mathbf{B}_n)$, and likewise an estimator of θ, $\hat{\theta}^{\mathcal{G}} := \Psi\hat{\beta}^{\mathcal{G}}$.

There are many possible choices for $\mathcal{G}$. We focus on $\mathcal{G} = \mathcal{L}$, which by inspection of equation (9), results in

$$\hat{\lambda}_j^{\mathcal{L}} = \left(1 - \frac{\sigma^2}{\Delta_{nj}|B_{nj}|^2}\right)_+ \tag{11}$$

where as usual $(\cdot)_+ = \max(\cdot, 0)$ is the soft thresholding function. We modify these weights further to help stabilize its behavior for smaller sample sizes

$$\hat{\lambda}_j = \left(1 - \frac{\Omega_n^2\sigma^2}{\Delta_{nj}|B_{nj}|^2}\right)_+, \tag{12}$$

where $\Omega_n^2 := (p-2)(1+\frac{\max_j \Delta_{nj}}{\min_j \Delta_{nj}})$. Note this choice of Ω_n^2 is motivated by Brown, Nie and Xie (2012) in which it is shown that in heteroscedastic case, the soft thresholded James-Stein type estimator is ensemble minimax with this added term. Lastly, define our estimator of θ to be

$$\hat{\theta}_n := \Psi\hat{\boldsymbol{\lambda}}(\mathbf{B}_n). \tag{13}$$

Other choices of $\mathcal{G}$ can and should be explored in further research into estimation in sequential inverse problems such as $\mathcal{M} := \{\boldsymbol{\lambda} \in \mathcal{L} : \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p\}$, which induces a monotonicity constraint on the estimated coefficients, or block methods of piecewise constant weights (Cavalier and Tsybakov, 2002). Alternatively, the aforementioned Tikonov-Phillips and Landwieber estimators correspond to specific subsets of $\mathcal{L}$. The Tikonov-Phillips estimator takes the form

$$\hat{\beta}_j^{\gamma} := \sum_{i=1}^{n} \frac{D_{ij}^* X_{ij}}{|D_{ij}|^2 + \gamma} \tag{14}$$

which can be rewritten as an element of $\mathcal{E}$ by defining

$$\lambda_j^{\gamma} := \frac{\Delta_{nj}}{\Delta_{nj} + \gamma} \tag{15}$$

with associated estimator $\hat{\beta}^{\gamma} = \boldsymbol{\lambda}^{\gamma}(\mathbf{B_n})$. The Landwieber estimator can be reformulated in the following, non-iterative form

$$\lambda_j^{(\gamma,\tau)} = 1 - [1 - \tau \Delta_{nj}]^{\gamma}, \tag{16}$$

where γ corresponds to the number of iterations and τ is a relaxation parameter. The associated estimator is $\hat{\beta}^{(\gamma,\tau)} = \boldsymbol{\lambda}^{(\gamma,\tau)}(\mathbf{B}_n)$.

It can be shown that, for example, $\sup_{\gamma} |\hat{R}_n(\boldsymbol{\lambda}^{\gamma}) - R(\boldsymbol{\lambda}^{\gamma})| \to 0$ in probability and therefore, by van der Vaart (1998), $\operatorname{argmin}_{\gamma} \hat{R}_n(\boldsymbol{\lambda}^{\gamma}) \to \operatorname{argmin}_{\gamma} R_n(\boldsymbol{\lambda}^{\gamma})$ in probability as well. A similar discussion holds for the Landwieber estimator as well. Hence, Proposition 5 in principle could provide a data-driven method for choosing the tuning parameters in these estimators.

3. Computational properties, variance estimation, and alternate methods

3.1. Computations

The specifics of the computation of an estimator $\hat{\theta}^{\mathcal{G}}$ depend upon the subset $\mathcal{G}$. However, as $\hat{R}_n$ is strictly convex and $\mathcal{G}$ is compact, then if $\mathcal{G}$ is convex, the solution can be found uniquely. Of the estimators mentioned above, all except $\mathcal{M}$ have a closed form solution and therefore trivial computation. The minimization of $\hat{R}_n$ over $\mathcal{M}$ can be accomplished by a well known algorithm called Pooled Adjacent Violators (PAV) (Robertson, Wright and Dykstra, 1988).

Given a sequence of data $(Y_i)_{i=1}^n$, our proposed estimator $\hat{\theta}_n$ can be computed via equations (12) and (13). After obtaining an additional observation Y_{n+1}, $\hat{\theta}_{n+1}$ can be formed using only $\mathbf{B}_n$, Δ_n, D_{n+1}, and $\mathbf{X}_{n+1}$ as input. As Δ_n can be readily updated to Δ_{n+1} and $\hat{\theta}_{n+1}$ can be computed from Δ_{n+1} and $\mathbf{B}_{n+1}$, it suffices to show that $\mathbf{B}_n$ can be updated. Observe,

$$\begin{aligned} B_{n+1,j} &= \frac{\sum_{i=1}^{n} D_{ij}^* X_{ij}}{\Delta_{n+1,j}} + \frac{D_{n+1,j}^* X_{n+1,j}}{\Delta_{n+1,j}} \\ &= \left(\frac{\Delta_{n,j}}{\Delta_{n+1,j}} \right) B_{nj} + \frac{D_{n+1,j}^* X_{n+1,j}}{\Delta_{n+1,j}}. \end{aligned}$$

In this way, our estimator $\hat{\theta}_n$ can updated to $\hat{\theta}_{n+1}$ efficiently in a storage sense by maintaining summary statistics $\mathbf{\Delta}_n$ and $\mathbf{B}_n$ instead of accessing the entire data-stream.

For computational complexity, there are two possible situations. First, if the (K_i) are convolution operators, then by Theorem 1, Ψ and D_i are the Fourier basis and Fourier coefficients, respectively, of K_i. In this case, $\hat{\theta}_n$ can be computed via the Fast Fourier Transform, which implies $O(p \log p)$ computations, which is highly efficient. However, there are K_i that do not satisfy Theorem 1 but do satisfy assumption (A3). For these K_i, the eigenvectors and eigenvalues must be computed via a conventional eigenvector solver, which necessarily has computational complexity $O(p^3)$. Though this could potentially become prohibitive for large scale problems, there do exist modern approximation methods for eigenvalues and eigenvectors that could be used instead, such as in Halko, Martinsson and Tropp (2009). However, we do not explore this idea further in this paper.

3.2. Estimating the variance parameter

To derive the theoretical results of this paper we assume that σ is known. In practice, this not usually the case (though, in the case of the LSST, the noise properties of the telescope are known to a certain extent due to the physics of the device). There are two main properties of the model in equation (1) that are of interest for variance estimation. First, we have access to a long sequence of observations $(Y_i)_{i=1}^n$ and second, each Y_i is gathered after begin corrupted by a forward operator K_i that is ill-conditioned.

Due to the first property, σ can be consistently estimated by identifying a subsequence $\mathcal{N}$ of $\mathbb{N}$ and using the observations $(Y_i)_{i\in\mathcal{N}}$ for σ estimation and $(Y_i)_{i\in\mathbb{N}\setminus\mathcal{N}}$ for θ estimation. If we define $\mathcal{N}'$ to be the set comprised of the first n' entries in $\mathcal{N}$ then we have the following estimator of σ

$$\hat{\sigma}^2_{\text{con}} := \frac{1}{pn'} \sum_{i\in\mathcal{N}} \sum_{j=1}^{p} \left(Y_{ij}^2 - \overline{Y}_j^2 \right).$$

By the (strong) law of large numbers, $1/n' \sum_{i\in\mathcal{N}'} Y_{ij}^2$ converges almost surely to $\sigma^2 + \lim_{n'\to\infty} 1/n' \sum_{i\in\mathcal{N}'} (\mathbb{E}Y_{ij})^2$ and $\overline{Y}_j^2$ converges to almost surely to $\lim_{n'\to\infty} 1/n' \sum_{i\in\mathcal{N}'} (\mathbb{E}Y_{ij})^2$. Hence, $\hat{\sigma}^2_{\text{con}}$ converges almost surely to σ^2.

This approach is unsatisfying as we are potentially using high-quality observations (low amounts of smoothing) for doing σ estimation and low quality observations (high amount of smoothing) for estimating θ. As an alternative, we can use an adaptation of the usual variance estimator from least-squares linear regression (Seber and Lee, 2003, Chapter 3.3) to estimate σ. Due to the second property, there is usually no null-space of the forward operator, which is where the variance estimation usually occurs. However, as the forward operators are ill-conditioned, there is ‘almost’ a null-space.

Suppose for some i_* that Y_{i_*} is an extremely low quality observation in the sense that the signal θ is highly smoothed. This is equivalent to the forward operator K_{i_*} being more ill-conditioned, which in turn implies the existence of a p' such that for $j = p', \ldots, p$, $|D_{i_* j}|^2$ is nearly zero. Suppose now that $\mathcal{N}$ is the set of all such indices i_*. Assume for simplicity that p' is the same for all observations in $\mathcal{N}$ and form the following statistic

$$\hat{\sigma}_i^2 := \frac{1}{p - p'} \sum_{j=p'}^{p} |X_{ij}|^2.$$

Then $\mathbb{E}\hat{\sigma}_i^2 = \sigma^2 + \frac{1}{p-p'} \sum_{j=p'}^{p} |D_{ij}|^2 |\beta_j|^2$ and we report $\hat{\sigma}^2 := 1/n' \sum_{i \in \mathcal{N}'} \hat{\sigma}_i^2$ as our estimator of σ^2. This is in general a biased estimator of the variance, and the bias doesn't disappear asymptotically. This is the price that must be paid for there not existing an exact null space for the forward operators K_i. Nevertheless, $\hat{\sigma}^2$ is still useful. First, it is conservative as it has a positive bias. Perhaps more importantly, this estimator provides an interesting situation where the lowest quality parts (those with index $p' \leq j \leq p$) of the lowest quality observations (those with index in $\mathcal{N}$) provide the best performance ($\hat{\sigma}^2$ has a small bias).

As alluded to after the introduction of the model in equation (1), the variance parameter σ could change between observations. In this case, there is no straightforward σ to estimate. However, suppose the (σ_i) are generally centered on some value $\overline{\sigma}$. Then, $Y_i = K_i\theta + \overline{\sigma} W_i + (\sigma_i - \overline{\sigma}) W_i$ and, by the same derivation that produces equation (6),

$$\mathbf{B}_n = \beta + \overline{\sigma} \Delta_n^{-1} \sum_{i=1}^{n} D_i^* \Psi^* W_i + \Delta_n^{-1} \sum_{i=1}^{n} (\sigma_i - \overline{\sigma}) D_i^* \Psi^* W_i. \tag{17}$$

Under assumption (A5), it is possible to show that with high probability that the remainder term goes to zero if $1/n \sum_{i=1}^{n} |\sigma_i - \overline{\sigma}| \to 0$; that is, if σ_i are suitably centered around $\overline{\sigma}$ in an asymptotic way. Hence, equations (6) and (17) are asymptotically the same and we should expect that the variance estimators introduced in this section will behave reasonably well in the varying σ situation as well.

3.3. Overview of alternate approaches

In equation (1), it is tempting to average the observations (Y_i) directly. This leads to the following model

$$\overline{Y}_n = \overline{K}_n \theta + \frac{\sigma}{\sqrt{n}} W \tag{18}$$

where, under assumption (A3), $\overline{K}_n := 1/n \sum_{i=1}^{n} K_i = \Psi \overline{D}_n \Psi^*$, $\overline{D}_n := 1/n \sum_{i=1}^{n} D_i$, $\overline{Y}_n := 1/n \sum_{i=1}^{n} Y_i$, and $W \sim N(0, I_p)$. This can also be equivalently expressed as

$$\overline{\mathbf{B}}_n = |\overline{D}_n|^{-2} \overline{D}_n^* \overline{X}_n = \beta + \frac{\sigma}{\sqrt{n}} |\overline{D}_n|^{-2} \overline{D}_n^* \Psi^* W. \tag{19}$$

Here, $\overline{X}_n = \Psi^* \overline{Y}_n$. We define the corresponding set of linear estimators to be $\overline{\mathcal{E}} := \{\hat{\theta} = \Psi \boldsymbol{\lambda}(\overline{\mathbf{B}}_n) : \boldsymbol{\lambda} \in \mathbb{C}^p\}$.

Note that we can write equation (18) without any assumptions about the eigenvectors of the forward operators; that is, without assumption (A3). However, under assumption (A3), the following theorem supports forming estimators based on equation (6) instead of equation (18).

Theorem 6. *Suppose for any fixed θ,*

$$R_1 = \inf_{\hat{\theta} \in \mathcal{E}} \mathbb{E}||\hat{\theta} - \theta||_2^2 \quad \text{and} \quad R_2 = \inf_{\hat{\theta} \in \overline{\mathcal{E}}} \mathbb{E}||\hat{\theta} - \theta||_2^2,$$

where the expectations in R_1 and R_2 are under $\mathbf{B}_n$ and $\overline{\mathbf{B}}_n$, respectively. Then

$$R_1 <^* R_2$$

where '$<^*$' means 'strictly less than except when $D_i \equiv D$ for all i and some D.' That is, the oracle linear risk based on equation (6) is strictly less than the oracle linear risk based on equation (18).

Remark 3.1. Note that the classic Tikonov-Phillips estimator based on the $\overline{Y}_n$ is of the form

$$\hat{\theta}_{\text{ridge}} = (\overline{K}_n^\top \overline{K}_n + \tau I)^{-1} \overline{K}_n^\top \overline{Y}_n.$$

This is equivalent to

$$\hat{\theta}_{\text{ridge}} = \Psi(|\overline{D}_n|^2 + \tau I)^{-1} |\overline{D}_n|^2 |\overline{D}_n|^{-2} \overline{D}_n^* \overline{X}_n = \Psi(|\overline{D}_n|^2 + \tau I)^{-1} |\overline{D}_n|^2 \overline{\mathbf{B}}_n, \quad (20)$$

and hence the Tikonov-Phillips estimator is in $\overline{\mathcal{E}}$, among many others.

Alternatively, we could form $\mathcal{K}_n := [K_1^\top, \ldots, K_n^\top]^\top$, $\mathcal{Y}_n := [Y_1^\top, \ldots, Y_n^\top]^\top$, and $\mathcal{W}_n \sim N(0, I_{n \cdot p})$. Then it follows that

$$\mathcal{Y}_n = \mathcal{K}_n \theta + \sigma \mathcal{W}_n.$$

However, estimators based on this approach, such as spline-type estimators, rely on accessing the entire history of observations (Y_i) and forward operators (K_i). This is computationally infeasible as this means both keeping and repeatedly accessing the entire sequence of observations. Hence, this approach doesn't satisfy our requirement of an estimate at time n being efficiently updatable to a new estimate after recording Y_{n+1}.

4. Supporting simulations

4.1. Description

In this section, we present visual results of using our estimator $\hat{\theta}_n$ to reconstruct various signals given access only to smoothed and noisy, but repeated, observations of that signal. For a quantitative comparison, we use the normalized relative risk (RR) given by

$$RR(\hat{\theta},\theta) := \sqrt{\frac{R(\hat{\theta},\theta)}{||\theta||^2}}. \tag{21}$$

We estimate RR by averaging 100 runs of our simulations.

In the one-dimensional cases, we compare our estimator, $\hat{\theta}_n$ to $\hat{\theta}_{\text{ridge}}$ from equation (20), with the smoothing parameter τ chosen by minimizing generalized cross validation (GCV). This is a natural comparison as $\hat{\theta}_{\text{ridge}}$ represents a well-understood type of estimator that a new estimator should outperform.

For the two-dimensional case, we compare our estimator to $\hat{\theta}_{\text{AIRY}}$, which appears in the software package AIRY (Bertero and Boccacci, 2000a,b; Correia et al., 2002) and is well-known in the Astronomy community for processing sequences of low-quality images. There does not exist an established method for choosing the smoothing parameters in $\hat{\theta}_{\text{AIRY}}$. Therefore, we set the tuning parameters interactively to the level that minimizes equation (21). Note that this implies that our estimator need only perform comparably to $\hat{\theta}_{\text{AIRY}}$ as it is using oracle information not available in practice.

For each of the signals introduced below, we fix the noise parameter σ to be such that the signal-to-noise $:= ||\theta||_1/(p\sigma) = 1$. For the one-dimensional examples, we admit K_i that are an equally weighted mixture of three Gaussians, normalized to have l_1 mass equal to 1, with means $\mu_1 = -0.75$, $\mu_2 = 0.00$, and $\mu_3 = 0.50$, along with standard deviations $\sigma_{iq} = 0.5 + E_{qi}$, where $E_{qi} \overset{i.i.d.}{\sim}$ exponential(1) and $q = 1, 2, 3$. For the two-dimensional example, we specify the means to all be $(0, 0)$ and the standard deviations are 2x2 matrices with diagonal entries all i.i.d. shifted exponentials of the same form as σ_{iq}. Note that this implies that the K_i are not symmetric. Also, note that Gaussian-like smoothing represents one of the worst cases as it exponentially down-weights the β_j for large j.

We consider two one-dimensional signals for estimation, which we refer to as θ^{smooth} and θ^{peaked} (Figure 2) with $p = 256$. The first signal, θ^{smooth}, is the sum of two Gaussians that are filtered by a Gaussian-tapered filter. This filter is additionally enforced to be zero above the $p/2$ frequency. Hence, θ^{smooth} is very smooth and compactly supported in the frequency domain. This example is instructive as a smooth function should be well represented by the eigenvectors Ψ of the smoothing operators K_i. Also, a compact representation in frequency domain will reveal the effectiveness of the soft-thresholding in zeroing out the appropriate B_{nj}, ie: those that correspond to the β_j that are zero. See the first row of Figure 2 for a plot of θ^{smooth} (left column) along with a typical example of a noisy, smoothed version that comprises the recorded data (right column).

Additionally, we consider the opposite situation by defining a signal θ^{peaked} that is the sum of three sharp, non-smooth, peaks. This signal is difficult to represent with the eigenvectors of smoothing matrices but is common in signal processing as it corresponds to both spectra from biochemical analysis and nuclear magnetic resonance imaging (nMRI). Note that the smallest peak is completely obscured by the smoothing and noise. See the second row of Figure 2

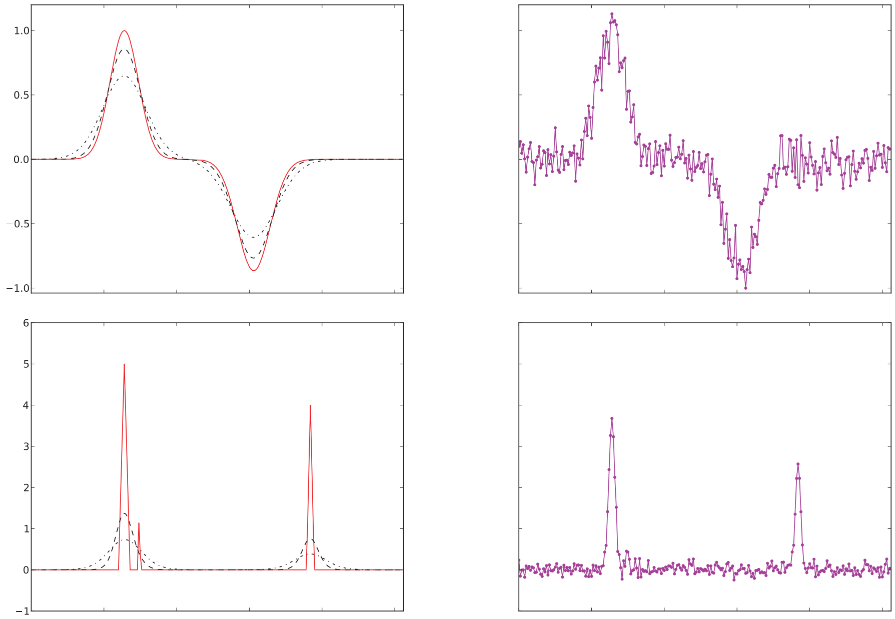


Fig 2. *The left column corresponds to the unobserved signal we wish to recover (solid, red line), along with the smallest and largest amount of smoothing considered (dashed and dashed-dotted black lines, respectively) The right column is an example observed signal. The top row corresponds to θ^{smooth} and the bottom to θ^{peaked}. Notice that in θ^{peaked}, the smaller peak is completely obscured in the observed data.*

for a plot of θ^{peaked} (left column) along with an example of a noisy, smoothed version (right column).

Lastly, we consider a two-dimensional signal, θ^{text}, that has dimension 1075×1075 (which means $\theta^{\text{text}} \in \mathbb{R}^p$, where $p = 1075^2 = 1,155,625$) which is a short section of text. We choose this due to the relative ease of visually evaluating the effect of an estimator at recovering the underlying signal. See Figure 3 for a plot of θ^{text} and an example of a noisy, smoothed version.

4.2. Results

In estimating either one-dimensional signal, θ^{smooth} or θ^{peaked}, the estimator $\hat{\theta}_n$ converges rapidly to the truth. See Table 1 for the RR of $\hat{\theta}_n$ and $\hat{\theta}_{\text{ridge}}$ used on both signals. In each case, for $n = 50$, the RR are approximately the same, with $\hat{\theta}_{\text{ridge}}$ having a slight edge. Every sample size thereafter shows substantial advantage of $\hat{\theta}_n$ over $\hat{\theta}_{\text{ridge}}$, culminating with a factor of two improvement in RR after $n = 300$ observations.

For estimating θ^{smooth}, both estimators have substantial oscillations for low sample sizes. However, due to $\hat{\theta}_n$ having a soft-thresholding effect, some of the entries in our estimator of β are zeroed out. In contrast, $\hat{\theta}_{\text{ridge}}$ only shrinks the

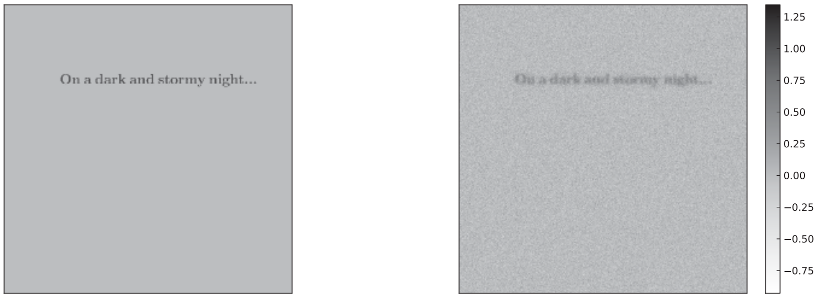


Fig 3. *The left column corresponds to θ^{text} and the right column corresponds to a noisy, smoothed version of θ^{text}.*

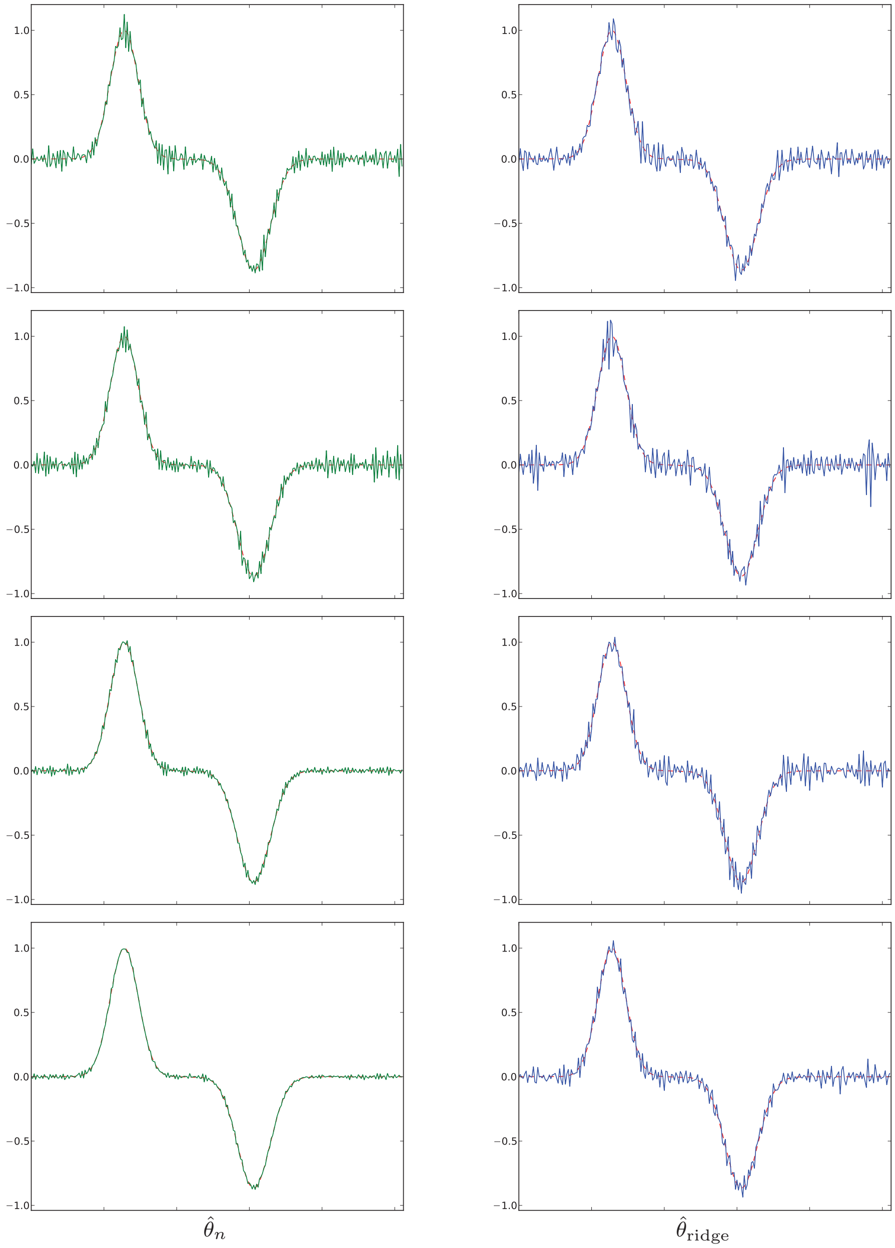
Table 1
The RR for the considered simulations. These are estimated by averaging 100 runs of our simulations

	Sample Size			
	$n = 50$	$n = 100$	$n = 200$	$n = 300$
$RR(\hat{\theta}_n, \theta^{\text{smooth}})$	0.291	0.210	0.149	0.120
$RR(\hat{\theta}_{\text{ridge}}, \theta^{\text{smooth}})$	0.288	0.223	0.199	0.173
$RR(\hat{\theta}_n, \theta^{\text{peaked}})$	0.148	0.116	0.092	0.079
$RR(\hat{\theta}_{\text{ridge}}, \theta^{\text{peaked}})$	0.151	0.150	0.149	0.141
$RR(\hat{\theta}_n, \theta^{\text{text}})$	0.743	0.689	0.646	0.634
$RR(\hat{\theta}_{\text{AIRY}}, \theta^{\text{text}})$	0.636	0.609	0.590	0.585

coefficients and hence still has substantial fluctuations after $n = 300$ observations. See Figure 4 for graphical results.

For the signal θ^{peaked}, $\hat{\theta}_n$ estimates the true height of the peaks accurately and quickly. In particular, the secondary small peak is definitively identified with the correct shape and height for $n = 50$ observations, while for $\hat{\theta}_{\text{ridge}}$, the secondary peak is much less clear. There are still some remaining oscillations at $n = 300$, resulting from unavoidable consequence of using the eigenvector basis. This is a well-known phenomenon in Fourier analysis known as the 'Gibbs effect.' Even with this obstacle, $\hat{\theta}_n$ converges quickly to θ^{peaked}. See Figure 5 for graphical results.

Our method extends to the two-dimensional case in a straight-forward manner. See Figure 6 for graphical results of $\hat{\theta}_n$ (left column) and $\hat{\theta}_{\text{AIRY}}$ with tuning parameter chosen as the argmin of RR (right column) on reconstructing θ^{text} for $n = 50$, 100, 200, 300. Additionally, see Table 1 for a comparison of the methods' RRs. The estimator $\hat{\theta}_{\text{AIRY}}$ has smaller RR than $\hat{\theta}_n$. This is not surprising as



$\hat{\theta}_n$

$\hat{\theta}_{\text{ridge}}$

FIG 4. *Estimation of θ^{smooth} by $\hat{\theta}_n$ (left column) and $\hat{\theta}_{ridge}$ (right column). The sample sizes range from top to bottom, $n = 50, 100, 200, 300$. Our estimator, $\hat{\theta}_n$, quickly converges to θ^{smooth}. However, $\hat{\theta}_{ridge}$, which doesn't zero out any coefficients, still has substantial fluctuations after $n = 300$ observations. See Table 1 for RR results for this simulation.*

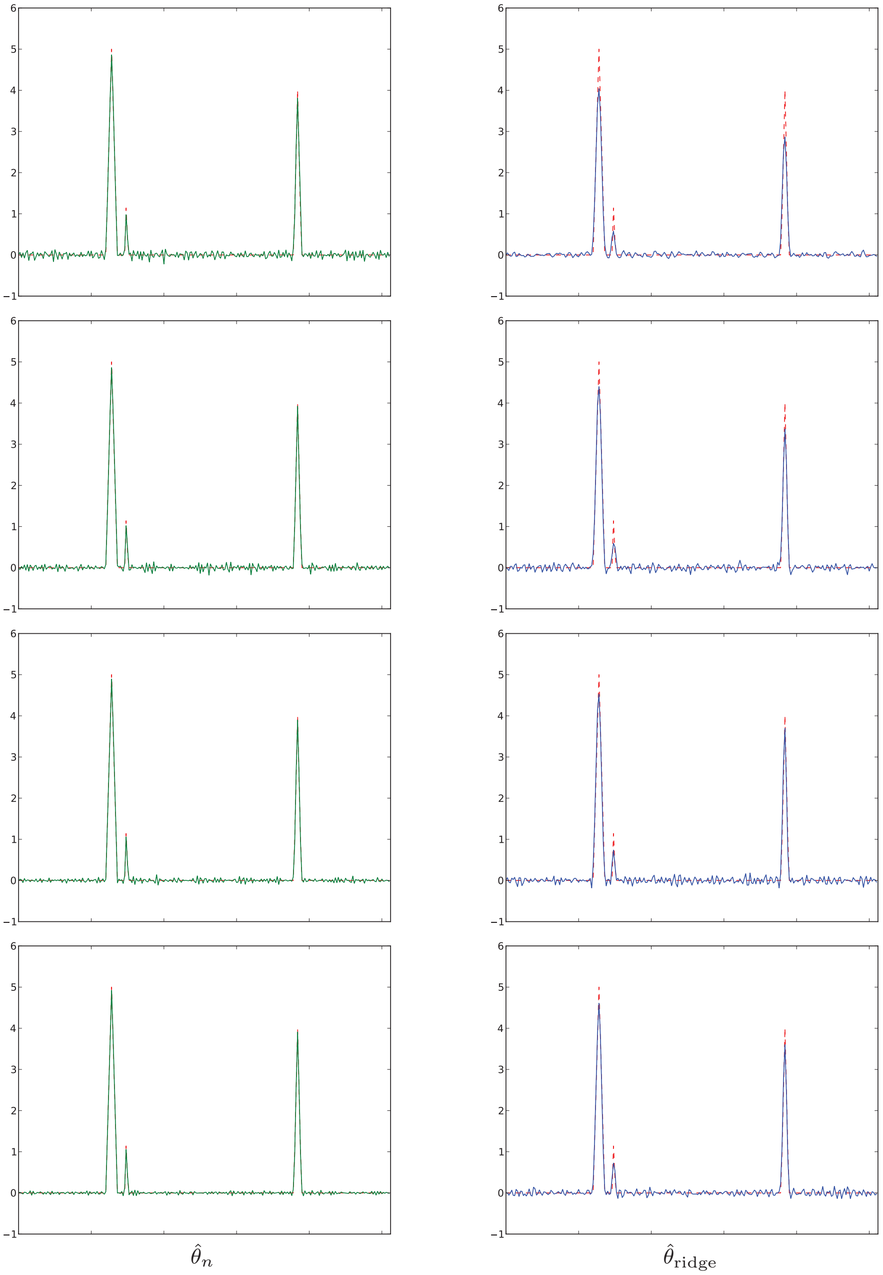


FIG 5. *Estimation of* θ^{peaked} *by* $\hat{\theta}_n$ *(left column) and* $\hat{\theta}_{ridge}$ *(right column). The sample sizes range from top to bottom,* $n =$ *50, 100, 200, 300. Our estimator,* $\hat{\theta}_n$*, estimates the true height of the peaks accurately and quickly. In particular, the secondary small peak is definitively identified with the correct shape and height. There are still some remaining oscillations at* $n = 300$*, resulting from an unavoidable Gibbs effect from using the eigenvectors as a basis. See Table 1 for RR results for this simulation.*

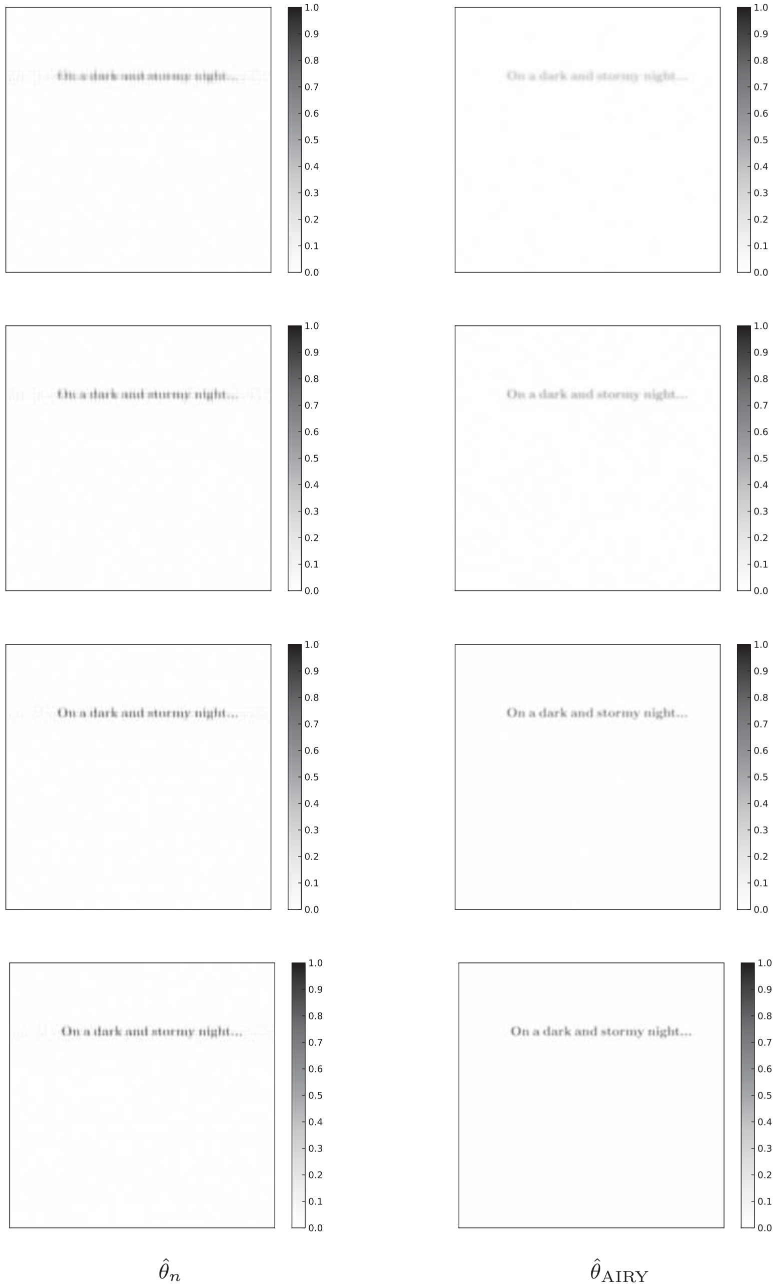


FIG 6. *Estimation of θ^{text} by $\hat{\theta}_n$ (left column) and $\hat{\theta}_{AIRY}$ (right column). The sample sizes range from top to bottom, $n = 50, 100, 200, 300$. Our estimator, $\hat{\theta}_n$, performs slightly worse than $\hat{\theta}_{AIRY}$ (see Table 1). However, considering $\hat{\theta}_{AIRY}$ uses oracle information, $\hat{\theta}_n$ compares very favorably, particularly for larger n.*

$\hat{\theta}_{\text{AIRY}}$, in this case, has access to oracle information unavailable to $\hat{\theta}_n$. However, in practice, a data-dependent tuning parameter would need to be selected for $\hat{\theta}_{\text{AIRY}}$, necessarily decreasing its performance. Additionally, even with this advantage, the difference in the RR of the two methods decreases by a factor of two as the sample size goes from $n = 50$ to $n = 300$ (0.107 to 0.049).

5. Discussion

In this paper, we provide a general method for recovering an unknown signal given a sequence of noisy observations that are only indirectly of that signal of interest. Our estimator, $\hat{\theta}_n$, has many favorable properties. It has computational efficiency in the sense that it can be updated with a new observation without need to reference the entire sequence of observations. Instead, it relies on only a few summary statistics that need to be maintained and updated. Though its computation is predicated on finding the eigenvectors and eigenvalues of potentially large matrices, the implementation is straightforward and generalizable to higher dimensional signals such as images. Additionally, there exist methods for the approximate computation of the eigenvectors of matrices that could in principle be used to speed up the computation of Ψ.

Also, $\hat{\theta}_n$ comes with theorectical support. The uniform consistency and oracle inequality results show that it is making about as good of use of the data as possible. Likewise, $\hat{\theta}_n$ has worked very well in our experiments so far, as evidenced by the results in Figures 1, 4, 5, and 6.

The development of this estimator is a novel contribution to the field of inverse problems by combining data-driven tuning parameter selection, computational efficiency, and statistical guarantees. Additionally, we develop some results about the estimation of a mean of a complex Gaussian random vector that are interesting in their own right.

Interesting future work would be to more fully explore the relationship between $\hat{\theta}_n$ and the other methods that are defined in equation (10). Different choices of the regularizing set $\mathcal{G}$ leads to different estimators, which encode different assumptions about the signal, θ, we hope to recover. Particular applications might benefit from leveraging these varying assumptions.

Appendix A

This section gives warrant for assumption (A3) in Section 2. Although a slightly weaker version of assumption (A3) is all that is actually required (that only the right eigenvectors need be the same instead of both left and right eigenvectors) we leave it in its current form for simplicity of exposition and conditions.

Two real matrices A, B share the same eigenvectors if they are simultaneously unitarily diagonalizable; that is, if there exists two diagonal matrices Σ_1, Σ_2 and a unitary matrix Ψ such that $A = \Psi\Sigma_1\Psi^*$ and $B = \Psi\Sigma_2\Psi^*$. Note that A and B must of course be unitarily diagonalizable, which implies by the spectral theorem that A and B are normal; that is $A^\top A = AA^\top$ and $B^\top B = BB^\top$. The following theorem characterizes simultaneous diagonalizability.

Lemma 7. *Let $\mathcal{K}$ be a commuting family of normal matrices. Then $\mathcal{K}$ is also simultaneously unitarily diagonalizable.*

Proof of Lemma 7. By the Schur unitary triangularization theorem (Horn and Johnson, 1985, Theorem 2.3.1) if $\mathcal{K}$ is a commuting family of matrices, then there is a unitary Ψ such that $\Psi K \Psi^*$ is upper triangular for every $K \in \mathcal{K}$. Hence, as normality is preserved under unitary congruence and a triangular normal matrix must be diagonal, the result follows. □

Though all Toeplitz matrices commute asymptotically as the number of rows and columns increases, not all Toeplitz matrices commute for a fixed size. Many subsets of the family of Toeplitz matrices satisfy Lemma 7, however. In particular, all circulant matrices commute (Gray, 2001, Chapter 3.1). This shows Theorem 1.

Appendix B

We use the following notation in several of the below proofs. We use $\lesssim$ to indicate 'less than or equal to up to a constant independent of n.' Also, it is convenient to think of a complex number $a = a_1 + a_2 i$ as an element $(a_1, a_2) \in \mathbb{R}^2$. In this case, we use $|||a|||^2 = a_1^2 + a_2^2$ as a norm on $\mathbb{R}^2$, as the complex modulus is not technically defined on elements of $\mathbb{R}^2$. Additionally, $Z \sim N(0, I_2)$ is the two dimensional standard normal. Lastly, we define $s_{nj}^2 := \Omega_n^2 \sigma^2 / \Delta_{nj}$, where Ω_n^2 is defined in equation (12).

We begin with a lemma that will be used in the proofs of Theorem 2 and Theorem 3:

Lemma 8. *Let $\mu \in \mathbb{R}^2$ be a vector, $\Sigma = diag(\sigma_1^2, \sigma_2^2)$ be a diagonal matrix with positive entries, and c^2 be a real, positive constant. Then*

$$\mathbb{P}(|||\mu + \Sigma^{1/2} Z|||^2 \leq c^2) \leq \mathbb{P}(|||\mu + \sigma_{max} Z|||^2 \leq c^2)$$

if $|||\mu||| > c$ and $\sigma_{max} = \max\{\sigma_1, \sigma_2\}$.

Here, we don't give a formal proof but provide intuition. The probability in Lemma (8) corresponds to the amount of the mass of an elliptical normal, aligned with the canonical axis, that resides in a ball of radius c at the origin. Hence, if $|||\mu||| > c$ (that is, the mean is outside the ball) a more spread out the normal results in more mass inside the ball.

Proof of Theorem 2. For simplicity, write $\hat{\beta}_n := \hat{\boldsymbol{\lambda}}(\mathbf{B}_n)$. Then

$$\sup_{\theta \in \Theta} R_n(\hat{\theta}_n, \theta) = \sup_{\beta \in \mathcal{B}} R_n(\hat{\beta}_n, \beta),$$

where $\mathcal{B} := \{\beta : ||\beta||^2 \leq T^2\} = \Psi^* \Theta$. Then we wish to show that

$$\lim_{n \to \infty} \sup \sup_{\beta \in \mathcal{B}} R_n(\hat{\beta}_n, \beta) = 0, \tag{22}$$

where the subscript n on R has been included to emphasize the dependence on the sample size.

We begin by defining the following set

$$A_j := \{\omega : |B_{nj}(\omega)|^2 > s_{nj}^2\}$$

where ω ranges over the measure space on which the random variable B_{nj} is defined. The utility of defining A_j is

$$\hat{\beta}_{nj}\mathbf{1}_{A_j} = \left(1 - \frac{\Omega_n^2\sigma^2}{\Delta_{nj}|B_{nj}|^2}\right) B_{nj}\mathbf{1}_{A_j} \tag{23}$$

Additionally, write $B_{nj} = \beta_j + Z_{nj}$ as a mean term plus stochastic term, where Z_{nj} is the j^{th} entry in the complex normal $\sigma\Delta_n^{-1}\sum_i(D_i^*\Psi^*W_i)$. Then the following bound on the j^{th} term in the loss holds:

$$\begin{aligned}
|\hat{\beta}_{nj} - \beta_j|^2 &= \mathbf{1}_{A_j}|\hat{\beta}_{nj} - \beta_j|^2 + \mathbf{1}_{A_j^c}|\hat{\beta}_{nj} - \beta_j|^2 \\
&= \mathbf{1}_{A_j}\left|\left(1 - \frac{s_{nj}^2}{|B_{nj}|^2}\right) B_{nj} - \beta_j\right|^2 + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&= \mathbf{1}_{A_j}\left|Z_{nj} - \left(\frac{s_{nj}^2(\beta_j + Z_{nj})}{|\beta_j + Z_{nj}|^2}\right)\right|^2 + \mathbf{1}_{A_j^c}|\beta_j|^2 \qquad (24) \\
&\le \mathbf{1}_{A_j}\left(|Z_{nj}| + \frac{s_{nj}^2}{|\beta_j + Z_{nj}|}\right)^2 + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&\le \mathbf{1}_{A_j}\left(|Z_{nj}| + s_{nj}\right)^2 + \mathbf{1}_{A_j^c}|\beta_j|^2.
\end{aligned}$$

To show that the expected value of the first term goes to zero in expectation, observe:

$$\begin{aligned}
\mathbb{E}\mathbf{1}_{A_j}\left(|Z_{nj}| + s_{nj}\right)^2 &\le \mathbb{E}|Z_{nj}|^2 + 2s_{nj}\mathbb{E}|Z_{nj}| + s_{nj}^2 \\
&\le \mathbb{E}|Z_{nj}|^2 + 2s_{nj}\sqrt{\mathbb{E}|Z_{nj}|^2} + s_{nj}^2 \\
&= \frac{\sigma^2}{\Delta_{nj}} + 2s_{nj}\sqrt{\frac{\sigma^2}{\Delta_{nj}}} + s_{nj}^2 \\
&\le \frac{\sigma^2}{\Delta_{nj}}\left(1 + 2\Omega_n + \Omega_n^2\right).
\end{aligned}$$

As $\Omega_n^2 < C < \infty$ for n large enough for some C by assumption (A5),

$$\mathbb{E}\mathbf{1}_{A_j}\left(|Z_{nj}| + s_{nj}\right)^2 = O(1/\Delta_{nj}) \tag{25}$$

uniformly in β.

For the second term, $\mathbf{1}_{A_j^c}|\beta_j|^2$, we need to show

$$\limsup_{n\to\infty} \sup_{\beta\in\mathcal{B}} \sum_{j=1}^{p} \mathbb{P}(A_j^c)|\beta_j|^2 = 0. \tag{26}$$

First, we compute the eigenvalue matrix Λ_{nj} of the covariance matrix of Z_{nj} as a vector in $\mathbb{R}^2$. By the properties of complex normals[2]

$$Z_{nj} \sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma^2/\Delta_{nj} & \Im C_{jj} \\ \Im C_{jj} & \sigma^2/\Delta_{nj} \end{array}\right)\right)$$

where C_{jj} is the j^{th} diagonal entry of the matrix $\sigma^2\Delta_n^{-1}\sum_i (D_i^*\Psi^*\overline{\Psi D_i})\Delta_n^{-1}$. Hence, the entries in Λ_{nj} are $\lambda_{nj,1}^2 = \sigma^2/\Delta_{nj} + \Im C_{jj}$ and $\lambda_{nj,2}^2 = \sigma^2/\Delta_{nj} - \Im C_{jj}$, which are both strictly positive. Also, define U to be the associated eigenvector matrix.

Though it is clear that $\mathbb{P}(A_j^c)|\beta_j|^2$ goes to zero pointwise, the worst β_j is arbitarily close to zero. Hence, to show uniform convergence, we define a parameter τ_{nj}^2. For each j, define $\mathcal{B}_j := \{\beta_j : |||\beta_j|||^2 \le T^2\}$ and split this set into $\mathcal{B}_j = \mathcal{B}_{jn} \cup \mathcal{B}_{jn}^c$, where

$$\mathcal{B}_{jn}^c := \{\beta_j : \tau_{nj}^2 \le |||\beta_j|||^2 \le T^2\}.$$

Also, as $|||\cdot|||$ is invariant under orthogonal operations, we can rotate everything by the eigenvectors U. Denote rotation by U by a tilde; that is, $\tilde{\beta}_j := U\beta_j$. Then,

$$\begin{aligned}
\sup_{\beta\in\mathcal{B}} \sum_{j=1}^{p} \mathbb{P}(A_j^c)|\beta_j|^2 &\le \sum_{j=1}^{p} \sup_{\beta_j\in\mathcal{B}_j} \mathbb{P}(A_j^c)|\beta_j|^2 \\
&\le \sum_{j=1}^{p} \max\left\{ \sup_{\beta_j\in\mathcal{B}_{nj}} \mathbb{P}_{\beta_j}(A_j^c)|\beta_j|^2, \sup_{\beta_j\in\mathcal{B}_{nj}^c} \mathbb{P}(A_j^c)|\beta_j|^2 \right\} \\
&\le \sum_{j=1}^{p} \max\left\{ \tau_{nj}^2, \sup_{\beta_j\in\mathcal{B}_{nj}^c} \mathbb{P}(A_j^c)|\beta_j|^2 \right\} \\
&= \sum_{j=1}^{p} \max\left\{ \tau_{nj}^2, \sup_{\beta_j\in\mathcal{B}_{nj}^c} \mathbb{P}(|||U(\beta_j + Z_n)|||^2 \le s_{nj}^2)|||\tilde{\beta}_j|||^2 \right\}.
\end{aligned}$$

Then continuing on with the second term in the max, and using Lemma 8 with $|||\tilde{\beta}_j||| > s_{nj}$, which happens if $\tau_{nj}^2 > s_{nj}^2$,

$$\begin{aligned}
&\sup_{\tilde{\beta}_j\in\mathcal{B}_{nj}^c} \mathbb{P}(|||\tilde{\beta}_j + \Lambda_{nj}^{1/2} Z|||^2 \le s_{nj}^2)|||\tilde{\beta}_j|||^2 \\
&\le \sup_{\tilde{\beta}_j\in\mathcal{B}_{nj}^c} \mathbb{P}(|||\tilde{\beta}_j + \lambda_{\max} Z|||^2 \le s_{nj}^2)|||\tilde{\beta}_j|||^2
\end{aligned}$$

[2] Technically, this covariance matrix is off by a constant, but this is not relevant for our current purposes.

$$
\begin{aligned}
&\le \sup_{\tilde{\beta}_j \in \mathcal{B}^c_{jn}} \left(1 - \Phi\left(|||\tilde{\beta}_j/\lambda_{\max}||| - s_{nj}/\lambda_{\max}\right)\right) |||\tilde{\beta}_j|||^2 \\
&= \sup_{\tau^2_{nj} \le u^2 \le T^2} \left(1 - \Phi\left(1/\lambda_{\max}(u - s_{nj})\right)\right) u^2 \\
&= \sup_{\frac{\tau_{nj}}{\lambda_{\max}} - s_{nj} \le t \le \frac{T}{\lambda_{\max}} - s_{nj}} (1 - \Phi(t))(\lambda_{\max}(t + s_{nj}))^2 \\
&= \lambda^2_{\max} \sup_{\frac{\tau_{nj}}{\lambda_{\max}} - s_{nj} \le t \le \frac{T}{\lambda_{\max}} - s_{nj}} (1 - \Phi(t))(t + s_{nj})^2 \\
&\le \lambda^2_{\max} \sup_{0 \le t \le \infty} (1 - \Phi(t))(t + 1)^2 \qquad \text{for } n \text{ large enough} \\
&\lesssim \lambda^2_{\max}
\end{aligned}
$$

The last inquality needs some explanation. Letting $g(t) = (1 - \Phi(t))(t + 1)^2$, then it is clear that g is continuous, $g(0) = 0.5$, and $\lim_{t\to\infty} g(t) = 0$ (this last claim follows by an application of L'Hôpital's rule). Therefore, g is uniformly bounded in t by a finite constant, from which the claim follows.

Thus,

$$\sup_{\beta \in \mathcal{B}} \mathbb{P}(A^c_j)|\beta_j|^2 \lesssim \max\{\tau^2_n, \lambda^2_{\max}\} \tag{27}$$

Hence, it is sufficient to choose $\tau^2_{nj} = 2s^2_{nj}$ and to note that

$$\lambda^2_{\max} \asymp s^2_{nj} \asymp \sigma^2/\Delta_{nj}.$$

This implies

$$\sup_{\beta \in \mathcal{B}} \mathbb{P}(A^c_j)|\beta_j|^2 = O(s^2_{nj}). \tag{28}$$

As we are summing over j in the risk, we conclude that

$$\limsup_{n\to\infty} \sup_{\beta \in \mathcal{B}} \gamma_n^{-1} R(\hat{\beta}_n, \beta) < \infty$$

where

$$\gamma_n = \min_j \frac{\sigma^2}{\Delta_{nj}}. \qquad \square$$

Proof of Theorem 3. We use the same notations and conventions as in the proof of Theorem 2. Note that if we define $\epsilon^2_{nj} = \sigma^2/\Delta_{nj}$, then the linear oracle risk is

$$R(\beta_*, \beta) = \min_{\tilde{\beta} = \boldsymbol{\lambda}(\mathbf{B}_n)} R(\tilde{\beta}, \beta) = \sum_{j=1}^p \frac{|\beta_j|^2 \epsilon^2_{nj}}{\epsilon^2_{nj} + |\beta_j|^2} = \sum_{j=1}^p \frac{|\beta_j|^2 s^2_{nj}}{s^2_{nj} + \Omega^2_n |\beta_j|^2}. \tag{29}$$

We can bound the j^{th} term in the loss as follows

$$
\begin{aligned}
&|\hat{\beta}_j - \beta|^2 \\
&= \mathbf{1}_{A_j} \left[|Z_{nj}|^2 - \frac{\overline{Z_{nj}} s^2_{nj} (\beta_j + Z_{nj})}{|\beta_j + Z_{nj}|^2} - \frac{Z_{nj} s^2_{nj} \overline{(\beta_j + Z_{nj})}}{|\beta_j + Z_{nj}|^2} + \frac{s^4_{nj}}{|\beta_j + Z_{nj}|^2} \right]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&= \mathbf{1}_{A_j}\left[|Z_{nj}|^2 - \frac{|Z_{nj}|^2 s_{nj}^2}{|\beta_j + Z_{nj}|^2} - \frac{s_{nj}^2(|Z_{nj}|^2 + \beta_j\overline{Z_{nj}} + \overline{\beta_j}Z_{nj})}{|\beta_j + Z_{nj}|^2} + \frac{s_{nj}^4}{|\beta_j + Z_{nj}|^2}\right] \\
&\quad + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&= \mathbf{1}_{A_j}\left[|Z_{nj}|^2 - \left(\frac{s_{nj}^2(|\beta_j + Z_{nj}|^2 - |\beta_j|^2)}{|\beta_j + Z_{nj}|^2}\right) + \frac{s_{nj}^4}{|\beta_j + Z_{nj}|^2}\right] + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&= \mathbf{1}_{A_j}\left[|Z_{nj}|^2 - s_{nj}^2 + \left(\frac{s_{nj}^2|\beta_j|^2}{|\beta_j + Z_{nj}|^2}\right) + \frac{s_{nj}^4}{|\beta_j + Z_{nj}|^2}\right] + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&\le |Z_{nj}|^2 + \mathbf{1}_{A_j}\left(\frac{s_{nj}^2|\beta_j|^2}{|\beta_j + Z_{nj}|^2}\right) + \mathbf{1}_{A_j^c}|\beta_j|^2 \\
&= |Z_{nj}|^2 + \mathbf{1}_{A_j}\left(\frac{s_{nj}^2|\beta_j|^2}{s_{nj}^2 + \Omega_n^2|\beta_j|^2}\right)\left(\frac{s_{nj}^2 + \Omega_n^2|\beta_j|^2}{|\beta_j + Z_{nj}|^2}\right) + \mathbf{1}_{A_j^c}|\beta_j|^2.
\end{aligned}
$$

By the previous proof, we see that the expected value of the first and third term go to zero uniformly over $\beta \in \mathcal{B}$ at rate $O(1/\Delta_{nj})$; the same rate as the oracle. For the second term, notice that

$$
\mathbf{1}_{A_j}\left(\frac{s_{nj}^2 + \Omega_n^2|\beta_j|^2}{|\beta_j + Z_{nj}|^2}\right) \le \mathbf{1}_{A_j}\left(1 + \frac{\Omega_n^2|\beta_j|^2}{|\beta_j + Z_{nj}|^2}\right) \lesssim \frac{\mathbf{1}_{A_j}|\beta_j|^2}{|\beta_j + Z_{nj}|^2} =: G_{nj}
$$

for n large enough, by assumption (A5). Then our goal is to show that

$$
\limsup_{n\to\infty} \sup_{\beta\in\mathcal{B}} \mathbb{E}G_{nj} < \infty.
$$

First, due to G_{nj} being rotationally symmetric (once we use $|||\cdot|||$ instead of $|\cdot|$), we renormalize to transform Z_{nj} into a vector $\tilde{Z}$ with independent standard normal components

$$
\begin{aligned}
G_{nj} &= \mathbf{1}_{A_j}\left(\frac{|||\beta_j|||^2}{|||\beta_j + Z_{nj}|||^2}\right) \\
&= \mathbf{1}_{\tilde{A}_j}\left(\frac{|||U^\top\beta_j|||^2}{|||U^\top\beta_j + U^\top Z_{nj}|||^2}\right) \\
&= \mathbf{1}_{\tilde{A}_j}\left(\frac{|||\tilde{\beta}_j|||^2}{|||\tilde{\beta}_j + \Lambda_{nj}^{1/2}\tilde{Z}|||^2}\right).
\end{aligned}
$$

We define Λ_{nj} and U in the previous proof as the eigenvalues and eigenvectors, respectively, of the covariance matrix of Z_{nj} and $\tilde{A}_j := \{||\tilde{\beta}_j + \Lambda_{nj}^{1/2}\tilde{Z}||^2 > s_{nj}^2\}$.

We break bounding $\mathbb{E}G_{nj}$ into cases.

Case 1. $|||\tilde{\beta}_j|||^2 \le s_{nj}^2$.

We see from the definition of $\mathbf{1}_{\tilde{A}_j}$ that

$$G_{nj} \leq \mathbf{1}_{\tilde{A}_j}\left(\frac{s_{nj}^2}{|||\tilde{\beta} + \Lambda_{nj}\tilde{Z}|||^2}\right) < \mathbf{1}_{\tilde{A}_j}\left(\frac{s_{nj}^2}{s_{nj}^2}\right) \leq 1.$$

Case 2. $|||\tilde{\beta}|||^2 > s_{nj}^2$.

Note that by the nonnegativity of G_{nj}

$$\mathbb{E}G_{nj} = \int_0^\infty \mathbb{P}(G_{nj} > \tau)\, d\tau.$$

For $\tau > 0$,

$$\begin{aligned}\mathbb{P}(G_{nj} > \tau) &= \mathbb{P}\left(s_{nj}^2 \leq |||\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{\tau}\right) \\ &= \begin{cases} 0 & \tau \geq \frac{|||\tilde{\beta}|||^2}{s_{nj}^2} \\ \mathbb{P}\left(s_{nj}^2 \leq |||\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{\tau}\right) & \text{o.w.}\end{cases}\end{aligned}$$

Therefore, for any $c^2 > 0$,

$$\begin{aligned}\mathbb{E}G_{nj} &= \int_0^\infty \mathbb{P}(G_{nj} > \tau)\, d\tau \\ &= \int_0^{c^2} \mathbb{P}(G_{nj} > \tau)\, d\tau + \int_{c^2}^{\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}} \mathbb{P}(G_{nj} > \tau)\, d\tau \\ &\leq c^2 + \left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\mathbb{P}(G_{nj} > c^2) \\ &\leq c^2 + \left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\mathbb{P}\left(|||\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{c^2}\right)\end{aligned}$$

If $c^2 > 1$, then the mean of of the random variable $\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}$ will be outside of the circle centered at zero with radius $||\tilde{\beta}||/c$. Hence, by Lemma 8, if we define $\lambda_{\max}^2 := \max\{\operatorname{diag}(\Lambda_{nj})\}$, then it follows that

$$\mathbb{P}\left(|||\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{c^2}\right) \leq \mathbb{P}\left(|||\tilde{\beta} + \lambda_{\max}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{c^2}\right). \tag{30}$$

Using this, observe

$$\begin{aligned}&\left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\mathbb{P}\left(|||\tilde{\beta} + \Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{c^2}\right) \\ &\leq \left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\mathbb{P}\left(|||\tilde{\beta}/\lambda_{\max} + \tilde{Z}|||^2 < \frac{|||\tilde{\beta}/\lambda_{\max}|||^2}{c^2}\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\left(1-\Phi\left(\left(1-\frac{1}{c}\right)|||\tilde{\beta}/\lambda_{\max}|||\right)\right)\\
&= \left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\left(1-\Phi\left(\left(1-\frac{1}{c}\right)|||\tilde{\beta}/\lambda_{\max}|||\right)\right)\\
&= \left(\frac{(\lambda_{\max}t)^2}{s_{nj}^2}\right)\left(1-\Phi\left(\left(1-\frac{1}{c}\right)t\right)\right)\\
&= \left(\frac{\lambda_{\max}^2}{s_{nj}^2}\right)\left[t^2\left(1-\Phi\left(\left(1-\frac{1}{c}\right)t\right)\right)\right]
\end{aligned}
$$

Where we have transformed $t = |||\tilde{\beta}|||/\lambda_{\max}$. Hence, as $s_{nj}^2 \asymp \lambda_{\max}^2$ and

$$
\sup_{\frac{s_{nj}}{\lambda_{\max}} \leq t \leq \frac{T}{\lambda_{\max}}} t^2\left(1-\Phi\left(\left(1-\frac{1}{c}\right)t\right)\right) \leq \sup_{0\leq t\leq\infty} t^2\left(1-\Phi\left(\left(1-\frac{1}{c}\right)t\right)\right) \leq 1
$$

we see that

$$
\left(\frac{|||\tilde{\beta}|||^2}{s_{nj}^2}\right)\mathbb{P}\left(|||\tilde{\beta}+\Lambda_{nj}^{1/2}\tilde{Z}|||^2 < \frac{|||\tilde{\beta}|||^2}{c^2}\right) = O(1),
$$

independent of β. And we conclude that

$$
\mathbb{E}G_{nj} = O(1),
$$

again, independent of β. This ends the proof. □

Proof of Theorem 4. Observe

$$
\begin{aligned}
\lim_{n\to\infty}\sup_{\beta\in\mathcal{B}}\mathbb{E}_{(D_i),X_n}||\hat{\beta}-\beta||^2 &= \lim_{n\to\infty}\sup_{\beta\in\mathcal{B}}\mathbb{E}_{(D_i)}\mathbb{E}_{X_n|(D_i)}||\hat{\beta}-\beta||^2\\
&\leq \lim_{n\to\infty}\mathbb{E}_{(D_i)}\sup_{\beta\in\mathcal{B}}R(\hat{\beta},\beta). \qquad (31)
\end{aligned}
$$

Therefore, to show the result, it suffices to exchange the limit and integral. We appeal to the following bound from equation (24). For this proof, we use the following result. Let $\{X_t : t \in \mathcal{T}\}$ be a family of random variables. If $X_t \to X$ with probability one and $\{X_t : t \in \mathcal{T}\}$ is uniformly integrable, then $\mathbb{E}X_t \to \mathbb{E}X$. As a reminder, $\{X_t : t \in \mathcal{T}\}$ is *uniformly integrable* if

$$
\lim_{x\to\infty}\sup_{t\in\mathcal{T}}\mathbb{E}|X_t|\mathbf{1}_{|X_t|>x} = 0.
$$

Define $f_n := \sup_{\beta\in\mathcal{B}} R(\hat{\beta},\beta)$. Then, if we can show that f_n is uniformly integrable, we can exchange the limit and integral. The result follows by Theorem 2. Observe the following bound for f_n

$$
|f_n| = \sup_{\beta\in\mathcal{B}}\mathbb{E}_{X_n|(D_i)}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j|^2
$$

$$
\begin{aligned}
&\leq \sup_{\beta \in \mathcal{B}} \mathbb{E}_{X_n|(D_i)} \sum_{j=1}^{p} \left[\mathbf{1}_{A_j} \left(|Z_{nj}| + s_{nj} \right)^2 + \mathbf{1}_{A_j^c} |\beta_j|^2 \right] \\
&\leq \sum_{j=1}^{p} \left(\frac{\sigma^2}{\Delta_{nj}} + 2 s_{nj} \sqrt{\frac{\sigma^2}{\Delta_{nj}}} + s_{nj}^2 + T^2 \right) \\
&\leq \sum_{j=1}^{p} \left(\frac{\sigma^2}{\Delta_{nj}} \left(1 + 2\Omega_n^2 + (\Omega_n^2)^2 \right) + T^2 \right) \\
&= \sum_{j=1}^{p} \left(\frac{\sigma^2}{\Delta_{nj}} (\Omega_n^2 + 1)^2 + T^2 \right) \\
&=: \sum_{j=1}^{p} g_j =: g_n.
\end{aligned}
$$

Therefore, it is sufficient to show that g_n is uniformly integrable in order to show that f_n is uniformly integrable. Note that

$$
\mathbb{E}|g_j| \mathbf{1}_{|g_j| > x} \leq x \mathbb{P}(g_j > x) + \int_x^\infty \mathbb{P}(g_j > y) dy.
$$

For large x, $x > T^2$ and for large n, $\Omega_n \asymp 1$. Therefore, we only need deal with the term σ^2 / Δ_{nj}.

Using assumption (B4), continuing the above with relevant terms, and noticing that $\sup_n f_n$ occurs at $n = 1$, it follows that for x large enough

$$
\begin{aligned}
x \mathbb{P}\left(\frac{1}{|D_{1j}|^2} > x \right) + \int_x^\infty \left(\frac{1}{|D_{1j}|^2} > y \right) dy &= x \left(\frac{1}{x^\rho} \right) + \int_x^\infty \left(\frac{1}{y^\rho} \right) dy \\
&= \left(\frac{1}{x^{\rho - 1}} \right) + \int_x^\infty \left(\frac{1}{y^\rho} \right) dy \\
&\to 0.
\end{aligned}
$$

This allows for the exchange of integration end hence shows the desired result. □

Proof of Proposition 5. We can expand (8) for any $\boldsymbol{\lambda}(\mathbf{B}_n) \in \mathcal{E}$ as

$$
R(\boldsymbol{\lambda}) := R_\beta(\boldsymbol{\lambda}(\mathbf{B}_n)) = \sum_{j=1}^{p} \left[(\lambda_j - 1)^2 |\beta_j|^2 + \frac{\sigma^2 \lambda_j^2}{\Delta_{nj}} \right]. \tag{32}
$$

To form an estimator of R, we notice that $\mathbb{E}_{\beta_j}(|B_{nj}|^2 - \sigma^2/\Delta_{nj}) = |\beta_j|^2$. Hence,

$$
\hat{R}(\boldsymbol{\lambda}) := \sum_{j=1}^{p} \left[(\lambda_j - 1)^2 \left(|B_{nj}|^2 - \frac{\sigma^2}{\Delta_{nj}} \right) + \frac{\sigma^2 \lambda_j^2}{\Delta_{nj}} \right] \tag{33}
$$

is an unbiased estimate of $R(\boldsymbol{\lambda})$. We can make a substitution

$$
\hat{\psi}_j := (|B_{nj}|^2 - \sigma^2/\Delta_{nj})/|B_{nj}|^2,
$$

which produces

$$\hat{R}(\boldsymbol{\lambda}) = \sum_{j=1}^{p} \left[(\lambda_j - \hat{\psi}_j)^2 |B_{nj}|^2 \right] + \sigma^2 \sum_{j=1}^{p} \left(\frac{\hat{\psi}_j}{\Delta_{nj}} \right). \tag{34}$$

Finally, note that the second term in $\hat{R}$ doesn't depend on $\boldsymbol{\lambda}$, so it can be ignored for minimization purposes. Define

$$\hat{R}_n(\boldsymbol{\lambda}) := \sum_{j=1}^{p} (\lambda_j - \hat{\psi}_j)^2 |B_{nj}|^2 \tag{35}$$

which is proportional to $\hat{R}(\boldsymbol{\lambda})$. This is our objective function for formulating estimators.

However, there are some natural restrictions. First, define $\mathcal{L} := [0,1]^p$. If we consider a transformed version of (32) by making the substitution $\psi_j := |\beta_j|^2 / (|\beta_j|^2 + \Delta_{nj})$, then

$$R(\boldsymbol{\lambda}) = \sum_{j=1}^{p} \left[(\lambda_j - \psi_j)^2 \left(|\beta_j|^2 + \frac{\sigma^2}{\Delta_{nj}} \right) + \sigma^2 \left(\frac{\psi_j}{\Delta_{nj}} \right) \right]. \tag{36}$$

By inspection, the minimizer of (36) falls in $\mathcal{L}$ as $\psi_j \in [0,1]$ for each j. Hence, we cannot get a lower risk by considering any more general sets and thus confine our attention to $\boldsymbol{\lambda} \in \mathcal{L}$. □

Proof of Theorem 6. Direct computation shows that

$$R_1 = \min_{\lambda} \left(\sum_{j=1}^{p} (1 - \lambda_j)^2 |B_j|^2 + \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j^2}{\Delta_{nj}} \right)$$

and

$$R_2 = \min_{\lambda} \left(\sum_{j=1}^{p} (1 - \lambda_j)^2 |B_j|^2 + \frac{\sigma^2}{n} \sum_{j=1}^{p} \frac{\lambda_j^2}{|D_n|_j^2} \right).$$

This implies that

$$R_1 = \sum_{j=1}^{p} \frac{\frac{\sigma^2}{\Delta_{nj}} |\beta_j|^2}{|\beta_j|^2 + \frac{\sigma^2}{\Delta_{nj}}} = \sum_{j=1}^{p} \frac{|\beta_j|^2}{\frac{\Delta_{nj}}{\sigma^2} |\beta_j|^2 + 1}$$

and

$$R_2 = \sum_{j=1}^{p} \frac{\frac{\sigma^2}{n|D_n|_j^2} |\beta_j|^2}{|\beta_j|^2 + \frac{\sigma^2}{n|D_n|_j^2}} = \sum_{j=1}^{p} \frac{|\beta_j|^2}{\frac{n|D_n|_j^2}{\sigma^2} |\beta_j|^2 + 1}.$$

Hence, the result reduces to comparing Δ_{nj} to $n|D_n|_j^2$. Note

$$|D_n|^2 = D_n^* D_n = \frac{1}{n^2} \sum_{i,q} D_i^* D_q$$

and therefore

$$n|D_n|_j^2 = \frac{1}{n}\sum_{i,q} D_{ij}^* D_{qj}.$$

Observe

$$\begin{aligned}
n|D_n|_j^2 - \Delta_{nj} =& \frac{1}{n}\sum_{i,q} D_{ij}^* D_{qj} - \sum_{i=1}^n |D_{ij}|^2 \\
=& \left(\frac{1}{n} - 1\right)\Delta_{nj} + \sum_{i\neq q} D_{ij}^* D_{qj} \\
\leq& \frac{1}{n}\left(-(n-1)\Delta_{nj} + \sum_{i\neq q} D_{ij}^* D_{qj}\right) \\
\lesssim& -(n-1)\sum_{i=1}^n |D_{ij}|^2 + \sum_{i\neq q} D_{ij}^* D_{qj} \\
\leq& -(n-1)\sum_{i=1}^n |D_{ij}|^2 + \sum_{i\neq q}(|D_{ij}|^2 + |D_{qj}|^2)/2 \\
=& -(n-1)\sum_{i=1}^n |D_{ij}|^2 + (n-1)\sum_{i=1}^n |D_{ij}|^2 \\
\leq& 0
\end{aligned}$$

where the last inequality follows as $|D_{ij}||D_{iq}| \leq (|D_{ij}|^2 + |D_{qj}|^2)/2$ by the arithmetic-geometric inequality. □

References

Backus, G. and Gilbert, F. (1968). The resolving power of gross Earth data. *Geophysical Journal of the Royal Astronomical Society* **16** 169–205.

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *The Annals of Mathematical Statistics* **25** 423–462. MR0063630

Beran, R. (2000). Scatterplot smoothers: superefficiency through basis economy. *Journal of the American Statistical Association* **95** 155–171. MR1803148

Berenstein, C. and Patrick, E. V. (1990). Exact deconvolution for multiple convolution operators – an overview, plus performance characterizations for imaging sensors. *Proceedings of the IEEE* **78** 723–734.

Bertero, M. and Boccacci, P. (1998). *Introduction to Inverse Problems in Imaging*. IOP Publishing, Bristol. MR1640759

Bertero, M. and Boccacci, P. (2000a). Image restoration methods for the Large Binocular Telescope. *Astronomy and Astrophysics Supplement Series* **147** 323–333.

Bertero, M. and Boccacci, P. (2000b). Application of the OS-EM method to the restoration of Large Binocular Telescope images. *Astronomy and Astrophysics Supplement Series* **144** 181–186.

Brown, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters. *Journal of the American Statistical Association* **70** 417–427. MR0373082

Brown, L. D., Nie, H. and Xie, X. (2012). Ensemble minimax estimation for multivariate normal means. *The Annals of Statistics*.

Candés, E. J. and Donoho, D. L. (2002). Recovering edges in ill-posed inverse problems: optimality of curvelet frames. *Annals of Statistics* **30** 784–842. MR1922542

Casey, S. and Walnut, D. (1994). Systems of convolution equations, deconvolutions, shannon sampling, and the wavelet and Gabor transforms. *SIAM Review* **36** 537–577. MR1306923

Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems* **24**. MR2421941

Cavalier, L. and Tsybakov, A. B. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields* **123** 323–354. MR1918537

Cavalier, L., Golubev, G. K., Picard, D. and Tsybakov, A. B. (2002). Oracle inequlities for inverse problems. *Annals of Statistics* **30** 843–874. MR1922543

Correia, S., Carbillet, M., Boccacci, P., Bertero, M. and Fini, L. (2002). Restoration of interferometric images. *Astronomy & Astropysics* **387** 733–743.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis* 101–126. MR1325535

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224. MR1379464

Gallager, R. G. (2008). *Principals of Digital Communication*. Cambridge.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution. *The Annals of Mathematical Statistics* **34** 152–177. MR0145618

Gray, R. M. (2001). *Toeplitz and circulant matrices: a review*.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *California Inst. Tech., Sep. 2009* **ACM Report 2009-05**.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press. MR0832183

Ólafsson, G. and Quinto, E. T. (2005). *The Radon Transform, Inverse Problems, and Tomography: Short Course*. American Mathematical Society, Atlanta Georgia. MR2207138

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1** 502–527. MR0874480

Piana, M. and Bertero, M. (1996). Regularized deconvolution of multiple images of the same object. *J. Opt. Soc. Am. A* **13** 1516–1523.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley and Sons, Great Britain. MR0961262

Schreier, P. J. and Scharf, L. L. (2003). Second-order analysis of improper complex random vectors and processes. *Signal Processing, IEEE Transactions on* **51** 714-725. MR1963873

Schreier, P. J., Scharf, L. L. and Mullis, C. T. (2005). Detection and estimation of improper complex random signals. *Information Theory, IEEE Transactions on* **51** 306-312. MR2234588

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. John Wiley and Sons, Great Britain. MR1958247

Starck, J. L., Pantin, E. and Murtagh, F. (2002). Deconvolution in Astronomy: a review. *Publications of the Astronomy Society of the Pacific* **114** 1051-1069.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **90** 1247–1256. MR0630098

Tenorio, L. (2001). Statistical regularization of inverse problems. *SIAM Review* **43** 347–366. MR1861086

van Dyk, D., Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., Kashyap, V., Siemiginowska, A. and Zezas, A. (2006). Deconvolution in high-energy Astrophysics: science, instrumentation, and methods. *Bayesian Analysis* **1** 189-236.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA. MR1045442

Wooding, R. (1956). The multivariate distribution of complex normal variables. *Biometrika* **43** 212-215. MR0077007