Song *et al. EURASIP Journal on Advances in Signal Processing* (2018) 2018:32
https://doi.org/10.1186/s13634-018-0551-y

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

On the effect of model mismatch for sequential Info-Greedy Sensing

Ruiyang Song[1], Yao Xie[2*] and Sebastian Pokutta[2]

Abstract

We characterize the performance of sequential information-guided sensing (Info-Greedy Sensing) when the model parameters (means and covariance matrices) are estimated and inaccurate. Our theoretical results focus on Gaussian signals and establish performance bounds for signal estimators obtained by Info-Greedy Sensing, in terms of conditional entropy (related to the estimation error) and additional power required due to inaccurate models. We also show covariance sketching can be used as an efficient initialization for Info-Greedy Sensing. Numerical examples demonstrate the good performance of Info-Greedy Sensing algorithms compared with random measurement schemes in the presence of model mismatch.

Keywords: Sequential compressed sensing, Adaptive sensing, Mutual information, Model mismatch

1 Introduction

Sequential compressed sensing is a promising new information acquisition and recovery technique to process big data that arises in various applications such as compressive imaging [1–3], power network monitoring [4], and large-scale sensor networks [5]. The sequential nature of the problems is either because the measurements are taken one after another or due to the fact that the data is obtained in a streaming fashion so that it has to be processed in one pass.

To harvest the benefits of adaptivity in sequential compressed sensing, various algorithms have been developed (see [6] for a review). We may classify these algorithms as (1) being agnostic about the signal distribution and, hence, random measurements are used [7–10], (2) exploiting additional structure of the signal (such as graphical structure [11], sparse [12–14], low rank [15], and tree-sparse structure [16, 17]) to design measurements, and (3) exploiting the distributional information of the signal in choosing measurements [18], possibly through maximizing mutual information. The additional knowledge about signal structure or distributions are various forms of *information* about the unknown signal. Such work includes the seminal Bayesian compressive sensing work [19], Gaussian mixture models (GMM) [20, 21], the classic information gain maximization [22] based on quadratic approximation to the information gain function, and our earlier work [6] which is referred to as *Info-Greedy Sensing*. Info-Greedy Sensing is a framework that aims at designing subsequent measurements to maximize the mutual information conditioned on previous measurements. Conditional mutual information is a natural metric here, as it captures exclusively useful new information between the signal and the resulted measurements disregarding noise and what has already been learned from previous measurements. Information may play a distinguishing role: as the compressive imaging example demonstrated in Fig. 1 (see Section 4 for more details), with a bit of (albeit inaccurate) information estimated via random samples of small patches of the image, our Info-Greedy Sensing is able to recover details of a high-resolution image, whereas random measurements completely miss the image. As shown in [6], Info-Greedy Sensing for a Gaussian signal becomes a simple iterative algorithm: choosing the measurement as the leading eigenvector of the conditional signal covariance matrix in that iteration and then updating the covariance matrix via a simple rank-one update or, equivalently, choosing measurement vectors $a_1, a_2, \ldots$ as the orthonormal eigenvectors of the signal covariance matrix Σ in a decreasing order of eigenvalues. Different from the earlier literature

*Correspondence: yao.xie@isye.gatech.edu
[2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Full list of author information is available at the end of the article

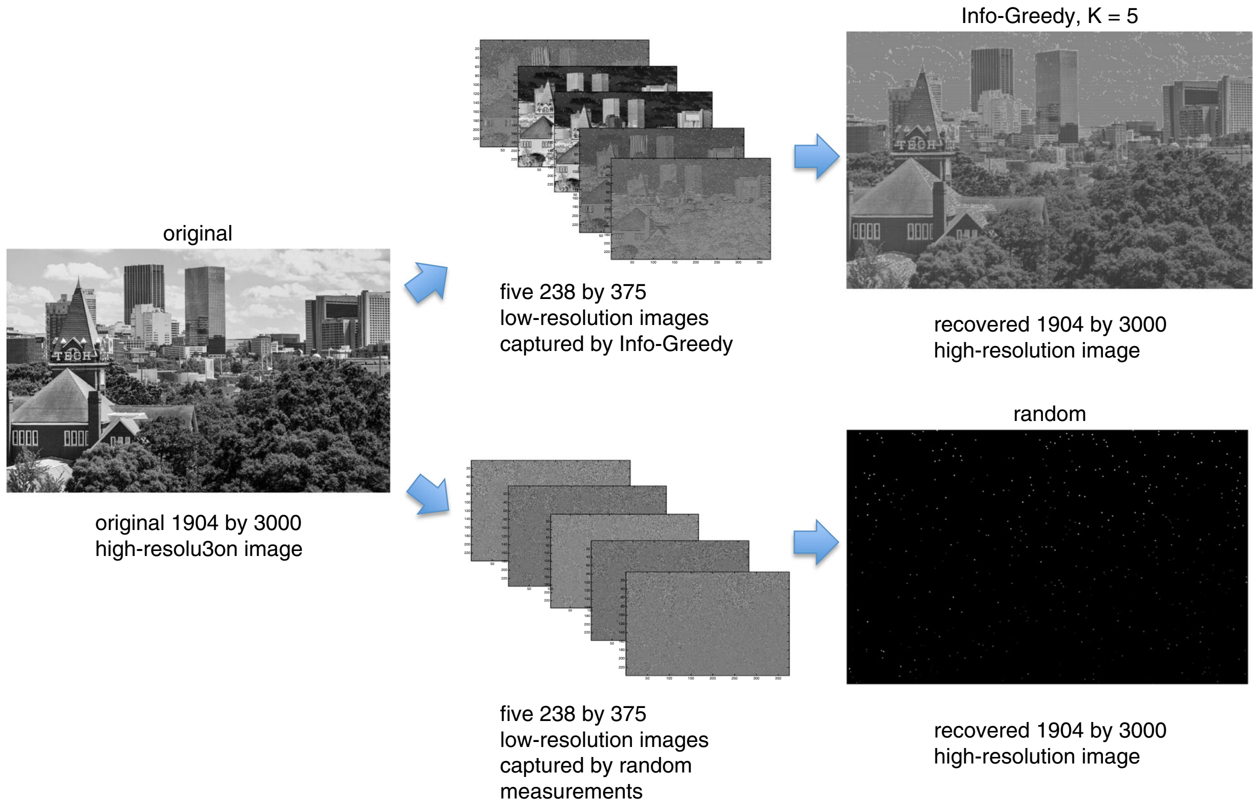

Fig. 1 Value of information in sensing a high-resolution image of size 1904 × 3000. Here, compressive linear measurements correspond to extracting the so-called *features* in compressive imaging [1–3]. In this example, the compressive imaging system captures five low-resolution images of size 238 × 275 using masks designed by Info-Greedy Sensing or random sensing (this corresponds to compressing the data into 8.32% of its original dimensionality). Info-Greedy Sensing performs much better than random features and preserves richer details in the recovered image. Details are explained in Section 4.3.2

[22], Info-Greedy Sensing determines not only the direction but also the precise magnitude of the measurements.

In practice, we usually need to estimate the signal covariance matrix, e.g., through a training session. For Gaussian signals, there are two possible approaches: either using training samples of the same dimension or through the new "covariance sketching" technique [23–25], which uses low-dimensional random sketches of the samples. Due to the inaccuracy of the estimated covariance matrices, measurement vectors usually deviate from the optimal directions as they are calculated as eigenvectors of the estimated covariance matrix. Hence, to understand the performance of information-guided algorithms in practice, it is crucial to quantify the performance of algorithms with model mismatch. This may also shed some light on how to properly initialize the algorithm.

In this paper, we aim at quantifying the performance of Info-Greedy Sensing when the parameters (in particular, the covariance matrices) are estimated. We focus on analyzing deterministic model mismatch, which is a reasonable assumption since we aim at providing instance-specific performance guarantees with sample estimated or sketched initial parameters. We establish a set of theoretical results including (1) studying the bias and variance of the signal estimator via posterior mean, by relating the error in the covariance matrix $\|\Sigma - \widehat{\Sigma}\|$ to the entropy of the signal posterior distribution after each sequential measurement, (2) establishing an upper bound on the additional power needed to achieve the signal precision $\|x - \hat{x}\| \leq \varepsilon$, where power is defined as the square of the norm of the measurement vector, and (3) translating these into requirements on the choice of the sample covariance matrix through direct estimation or through covariance sketching. Note that the power allocated for the measurements here is the minimum power required in order to achieve a prescribed precision for signal recovery within a fixed number of iterations. Furthermore, we also study Info-Greedy Sensing in a special setting when the measurement vector is desired to be one-sparse and establish analogously a set of theoretical results. Such a requirement arises from applications such as nondestructive testing (NDT) [26] or network tomography. We also present numerical examples to demonstrate the good performance of Info-Greedy Sensing compared to a batch method (where measurements are not adaptive) when there is mismatch. The main contribution of the paper is to study and understand the performance of Info-Greedy algorithm [6] in the presence of perturbed parameters, rather than proposing new algorithms.

Some other related works include [27], where adaptive methods for recovering structured sparse signals with Gaussian and Gaussian joint posterior are discussed, and [28], which analyzes the recovery of Gaussian mixture models with estimated mean and covariance using maximum a posteriori estimation. In [29], the orthogonal matching pursuit which aims at detecting the support of sparse signals while suffering from faulty measurements is studied. In this work, we focus on the case where the estimated mean, covariance, as well as the prior probability for each separate Gaussian component are available. Another work [20] discusses an adaptive sensing method for GMM, which is a two-step strategy that first adaptively detects the classification of the GMM, and then reconstructs the signal assuming it falls in the category determined in the previous step. While [20] assumes that there are sufficient samples for the first step in the first place, our early work [6] and this paper are different in that, sensing for GMM signal works on signal recovery directly without trying to identify the signal class as a first step. Hence, in general, our method is more tolerant to inaccuracy of the estimated parameters, and our algorithm can achieve good performance even without a large number of samples as demonstrated by numerical examples. The design of information-guided sequential sensing is related to the design of sequential experiments (see [15, 30, 31]) and large computer experiment approximation (see [32]). However, compared to the literature on design of experiments (e.g., [30]), our work does not use a statistical criterion based on the output of each iteration. In order words, we are designing our measurements based on the knowledge of the assumed model of the signal instead of the outputs of measurement.

Our notations are standard. Denote $[n] \triangleq \{1, 2, \ldots, n\}$; $\|X\|$, $\|X\|_F$, and $\|X\|_*$ represent the spectral norm, the Frobenius norm, and the nuclear norm of a matrix X, respectively; let $\nu_i(\Sigma)$ denote the ith largest eigenvalue of a positive semi-definite matrix Σ; $\|x\|_0$, $\|x\|_1$, and $\|x\|$ represent the ℓ_0, ℓ_1 and ℓ_2 norm of a vector x, respectively; let χ_n^2 be the quantile function of the chi-squared distribution with n degrees of freedom; let $\mathbb{E}[x]$ and $\mathrm{Var}[x]$ denote the mean and the variance of a random variable x; we write $X \succeq 0$ to indicate that the matrix is positive semi-definite; $\phi(x|\mu, \Sigma)$ denotes the probability density function of the multivariate Gaussian with mean μ and covariance matrix Σ; let e_j denote the jth column of identity matrix I (i.e., e_j is a vector with only one non-zero entry at location j); and $(x)^+ \triangleq \max\{x, 0\}$ for $x \in \mathbb{R}$.

2 Method: Info-Greedy Sensing

A typical sequential compressed sensing setup is as follows. Let $x \in \mathbb{R}^n$ be an unknown n-dimensional signal. We make K measurements of x sequentially

$$y_k = a_k^\intercal x + w_k, \quad k = 1, \ldots, K,$$

and the power of the measurement vector is $\|a_k\|^2 = \beta_k$. The goal is to recover x using measurements $\{y_k\}_{k=1}^K$. Consider a Gaussian signal $x \sim \mathcal{N}(0, \Sigma)$ with known zero mean and covariance matrix Σ (here without loss of generality we have assumed the signal has zero mean). Assume the rank of Σ is s and the signal is low rank, i.e. $s \ll n$ (however, the algorithm does not require the covariance to be low rank):

$$\mathrm{rank}(\Sigma) = s \ll n.$$

Our goal is to estimate the signal x using sequential and adaptive measurements. Info-Greedy Sensing introduced in [6] is one of such adaptive methods which chooses each measurement to maximize the conditional mutual information

$$a_k \leftarrow \underset{a}{\mathrm{argmax}} \left\{ \mathbb{I}\left[x; a^\intercal x + w | y_j, a_j, j < k\right] / a^\intercal a \right\}. \qquad (1)$$

The goal of this sensing scheme is to use a minimum number of measurements (or to use the minimum total power) so that the estimated signal is recovered with precision ε; i.e., $\|\widehat{x} - x\| < \varepsilon$ with a high probability p. Define

$$\chi_{n,p,\varepsilon} \triangleq \varepsilon^2 / \chi_n^2(p),$$

and we will show in the following that this is a fundamental quantity that determines the termination condition of our algorithm to achieve the precision ε with the confidence level p. Note that $\chi_{n,p,\varepsilon}$ is a precision ε adjusted by the confidence level.

2.1 Gaussian signal

In [6], we have devised a solution to (1) when the signal is Gaussian. The measurement will be made in the directions of the eigenvectors of Σ in a decreasing order of eigenvalues, and the powers (or the number of measurements) will be such that the eigenvalues after the measurements are sufficiently small (i.e., less than ε). The power allocation depends on the noise variance, signal recovery precision ε, and confidence level p, as given in Algorithm 1. Note that in Step 6, the update of covariance matrix can also be implemented, equivalently, via $\lambda\sigma^2 uu^\intercal / \left(\beta\lambda + \sigma^2\right) + \Sigma^{\perp u}$, as explained in (6). In the algorithm, the initializations μ and Σ are estimated and may not be very accurate.

Algorithm 1 Info-Greedy Sensing for Gaussian signals

```
Require: assumed signal mean $\mu$ (initialize with $\hat{x} = \mu$) and covariance matrix $\Sigma$, noise variance $\sigma^2$, recovery accuracy $\varepsilon$, confidence level $p$ close to 1
repeat
    $(\lambda, u) \leftarrow$ largest eigenvalue and associated normalized eigenvector of $\Sigma$
    $\beta \leftarrow \sigma^2(1/\chi_{n,p,\varepsilon} - 1/\lambda)^+$
    $a = \sqrt{\beta}u, y = a^\intercal x + w$ {measure}
    $\hat{x} \leftarrow \hat{x} + \Sigma a(y - a^\intercal \hat{x}) / \left(\beta\lambda + \sigma^2\right)$ {mean}
    $\Sigma \leftarrow \Sigma - \Sigma a a^\intercal \Sigma / \left(\beta\lambda + \sigma^2\right)$ {covariance}
until $\|\Sigma\| \leq \chi_{n,p,\varepsilon}$ {all eigenvalues small}
return signal estimate $\hat{x}$
```

2.2 One-sparse measurement

The problem of Info-Greedy Sensing with sparse measurement constraint, i.e., each measurement has only k_0 non-zero entries $\|a\|_0 = k_0$, has been examined in [6] and solved using outer approximation (cutting planes). Here, we will focus on one-sparse measurements, $\|a\|_0 = 1$, as it is an important instance arising in applications such as nondestructive testing (NDT).

Algorithm 2 Info-Greedy Sensing with sparse measurement $\|a\|_0 = 1$, for Gaussian signals

```
Require: assumed signal mean $\mu$ and covariance matrix $\Sigma$, noise variance $\sigma^2$, recovery accuracy $\varepsilon$, confidence level $p$
repeat
    $j^* \leftarrow \arg\max_j \Sigma_{jj}$
    $a \leftarrow \sqrt{\beta}e_{j^*}, y = a^\intercal x + w$ {measure}
    $\mu \leftarrow \mu + \Sigma a(y - a^\intercal \mu) / \left(\beta\Sigma_{j^*j^*} + \sigma^2\right)$ {mean}
    $\Sigma \leftarrow \Sigma - \Sigma a a^\intercal \Sigma / \left(\beta\Sigma_{j^*j^*} + \sigma^2\right)$ {covariance}
until $\|\Sigma\| \leq \chi_{n,p,\varepsilon}$ {all eigenvalues small}
return signal estimate $\hat{x} = \mu$
```

Info-Greedy Sensing with one-sparse measurements can be readily derived. Note that the mutual information between x and the outcome using one-sparse measurement $y_1 = e_j^\intercal x + w_1$ is given by

$$\mathbb{I}[x; y_1] = \frac{1}{2}\ln\left(\Sigma_{jj}/\sigma^2 + 1\right),$$

where Σ_{jj} denote the jth diagonal entry of matrix Σ. Hence, the measurement that maximizes the mutual information is given by e_{j^*} where $j^* \triangleq \arg\max_j \Sigma_{jj}$, i.e., measuring in the signal coordinate with the largest variance or largest uncertainty. Then Info-Greedy Sensing measurements can be found iteratively, as presented in Algorithm 2. Note that the correlation of signal coordinates are reflected in the update of the covariance matrix: if the ith and jth coordinates of the signal are highly correlated, then the uncertainty in j will also be greatly reduced if we measure in i. Similar to the previous two algorithms, the initial parameters are not required to be accurate.

2.3 Updating covariance with sequential data

If our goal is to estimate a sequence of data $x_1, x_2, \ldots$ (versus just estimating a single instance), we may be able to update the covariance matrix using the already estimated signals simply via

$$\widehat{\Sigma}_t = \alpha\widehat{\Sigma}_{t-1} + (1-\alpha)\hat{x}_t\hat{x}_t^\intercal, \quad t = 1, 2, \ldots, \tag{2}$$

and the initial covariance matrix is specified by our prior knowledge $\widehat{\Sigma}_0 = \widehat{\Sigma}$. Using the updated covariance matrix $\widehat{\Sigma}_t$, we design the next measurement for signal x_{t+1}. This way, we may be able to correct the inaccuracy of $\widehat{\Sigma}$ by including new samples. Here, α is a parameter for the update step-size. We refer to this method as "Info-Greedy-2" hereafter.

2.4 Gaussian mixture model signals

In this subsection we introduce the case of sensing Gaussian mixture model (GMM) signals. The probability density function of GMM is given by

$$p(x) = \sum_{c=1}^{C} \pi_c \phi(x|\mu_c, \Sigma_c),$$

where C is the number of classes, and π_c is the probability that the sample is drawn from class c. Unlike for Gaussian signals, the mutual information of GMM has no explicit form. However, for GMM signals, there are two approaches that tend to work well: Info-Greedy Sensing derived based on a gradient descent approach [6, 21] uses the fact that the gradient of the conditional mutual information with respect to a is a linear transform of the minimum mean square error (MMSE) matrix [33, 34], and the so-called *greedy heuristic* [6], which approximately maximizes the mutual information, shown in Algorithm 3. The greedy heuristic picks the Gaussian component with the highest posterior π_c at that moment and chooses the next measurement a as its eigenvector associated with the maximum eigenvalue. The greedy heuristic can be implemented more efficiently compared to the gradient descent approach and sometimes has competitive performance [6]. Also, the initialization for means, covariances, and weights can be off from the true values.

3 Performance bounds

In the following, we establish performance bounds, for cases when we (1) sense Gaussian signals using estimated covariance matrices and (2) sense Gaussian signals with one-sparse measurements.

Algorithm 3 Heuristic Info-Greedy Sensing for GMM signals

```
Require: number of components $C$, assumed means $\{\mu_c\}$,
    covariances $\{\Sigma_c\}$, initial weights $\{\pi_c\}$,
    noise variance $\sigma^2$,
    confidence level $p$
repeat
    $c^* \leftarrow \arg\max_c \pi_c$
    $(\lambda, u) \leftarrow$ largest eigenvalue and associated normalized eigenvector of $\Sigma_{c^*}$
    $\beta \leftarrow \sigma^2(1/\chi_{n,p,\varepsilon} - 1/\lambda)^+$
    $a = \sqrt{\beta}u, y = a^\intercal x + w$ {measurement}
    for $c = 1, \ldots, C$ do
        $\mu_c \leftarrow \mu_c + \left[(y - a^\intercal \mu_c)/\left(a^\intercal \Sigma_c a + \sigma^2\right)\right] \Sigma_c a$
        $\Sigma_c \leftarrow \Sigma_c - \Sigma_c a a^\intercal \Sigma_c/\left(a^\intercal \Sigma_c a + \sigma^2\right)$
        $\pi_c \leftarrow K\pi_c \exp\left\{-\frac{1}{2}(y - a^\intercal \mu_c)^2/\left(a^\intercal e \Sigma_c a + \sigma^2\right)\right\}$
        ($K$: normalizing constant)
    end for
until $\|\Sigma_{c^*}\| \le \chi_{n,p,\varepsilon}$
return signal class $c^* = \arg\max_c \pi_c$, estimate $\hat{x} = \mu_{c^*}$
```

3.1 Gaussian case with model mismatch

To analyze the performance of our algorithms when the assumed covariance $\widehat{\Sigma}$ used in Algorithm 1 is different from the true signal covariance matrix Σ, we introduce the following notations. Let the eigenpairs of Σ with the eigenvalues (which can be zero) ranked from the largest to the smallest to be $(\lambda_1, u_1), (\lambda_2, u_2), \ldots, (\lambda_n, u_n)$, and let the eigenpairs of $\widehat{\Sigma}$ with the eigenvalues (which can be zero) ranked from the largest to the smallest to be $(\hat{\lambda}_1, \hat{u}_1), (\hat{\lambda}_2, \hat{u}_2), \ldots, (\hat{\lambda}_n, \hat{u}_n)$. Let the updated covariance matrix in Algorithm 1 starting from $\widehat{\Sigma}$ after k measurements be $\widehat{\Sigma}_k$ and the true posterior covariance matrix of the signal conditioned on these measurements be Σ_k.

Note that since each time we measure in the direction of the dominating eigenvector of the posterior covariance matrix, $(\hat{\lambda}_k, \hat{u}_k)$ and (λ_k, u_k) correspond to the largest eigenpair of $\widehat{\Sigma}_{k-1}$ and Σ_{k-1}, respectively. Furthermore, define the difference between the true and the assumed conditional covariance matrices after k measurements as

$$E_k \triangleq \widehat{\Sigma}_k - \Sigma_k, \quad k = 1, \ldots, K,$$

and their sizes

$$\delta_k \triangleq \|E_k\|, \quad k = 1, \ldots, K.$$

Let the eigenvalues of E_k be $e_1 \ge e_2 \ge \cdots \ge e_n$, then the spectral norm of E_k is the maximum of the absolute values of the eigenvalues. Hence, $\delta_k = \max\{|e_1|, |e_n|\}$. Let

$$\delta_0 \triangleq \|\widehat{\Sigma} - \Sigma\|$$

denote the size of the initial mismatch.

3.1.1 Deterministic mismatch

First, we assume the mismatch is deterministic and find bounds for bias and variance of the estimated signal. It is common in practice to use estimated covariance matrices, which may have deterministic bias from the true covariances. Assume the initial mean is $\hat{\mu}$ and the true signal mean is μ, the updated mean using Algorithm 1 after k measurements is $\hat{\mu}_k$, and the true posterior mean is μ_k.

Theorem 1 (Unbiasedness) *After k measurements, the expected difference between the updated mean and the true posterior mean is given by*

$$\mathbb{E}[\hat{\mu}_k - \mu_k] = (\hat{\mu} - \mu) \cdot \prod_{j=1}^{k} \left(I_n - \frac{\beta_j \hat{\lambda}_j}{\beta_j \hat{\lambda}_j + \sigma^2} \hat{u}_j \hat{u}_j^\intercal\right).$$

Moreover, if $\hat{\mu} = \mu$, i.e., the assumed mean is accurate, the estimator is unbiased throughout all the iterations $\mathbb{E}[\hat{\mu}_k - \mu_k] = 0$, for $k = 1, \ldots, K$.

Next, we show that the variance of the estimator, when the initial mismatch $\|\widehat{\Sigma} - \Sigma\|$ is sufficiently small, reduces gracefully. This is captured through the reduction of entropy, which is also a measure of the uncertainty in the estimator. In particular, we consider the posterior entropy of the signal conditioned on the previous measurement outcomes. Since the entropy of a Gaussian signal $x \sim \mathcal{N}(\mu, \Sigma)$ is given by $\mathbb{H}[x] = \ln\left[(2\pi e)^{n/2} \det^{1/2}(\Sigma)\right]$, the conditional mutual information is the log of the determinant of the conditional covariance matrix, or equivalently the log of the volume of the ellipsoid defined by the covariance matrix. Here, to accommodate the scenario where the covariance matrix is low rank (our earlier assumption), we consider a modified definition for conditional entropy, which is the logarithm of the volume of the ellipsoid on the low-dimensional space that the signal lies on:

$$\mathbb{H}[x|y_j, a_j, j \le k] = \ln[(2\pi e)^{s/2} \mathsf{Vol}(\Sigma_k)],$$

where $\mathsf{Vol}(\Sigma_k)$ is the volume of the ellipse, which equals to the product of the non-zero eigenvalues of Σ_k:

$$\mathsf{Vol}(\Sigma_k) = \lambda_1 \cdots \lambda_{s_k},$$

where $\mathrm{rank}(\Sigma_k) = s_k$.

Theorem 2 (Entropy of estimator) *If for some constant $\delta \in (0, 1)$ the initial error satisfies*

$$\|\widehat{\Sigma} - \Sigma\| \le \frac{\delta}{4^{K+1}} \chi_{n,p,\varepsilon}, \tag{3}$$

then for $k = 1, \ldots, K$,

$$\mathbb{H}[x|y_j, a_j, j \le k] \le \frac{s}{2} \left\{\ln[2\pi e\, tr(\Sigma)] - \sum_{j=1}^{k} \ln(1/f_j)\right\}, \tag{4}$$

where

$$f_k \triangleq 1 - \frac{1-\delta}{s}\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2} \in (0,1), \quad k = 1,\ldots,K. \tag{5}$$

Note that in (3), the allowable initial error decreases with K. This is due to that larger K means the recovery precision criterion gets stricter, and hence, the maximum tolerable initial bias gets smaller. In the proof of Theorem 2, we track the trace of the underlying actual covariance matrix $\mathrm{tr}(\Sigma_k)$ as the cost function, which serves as a surrogate for the product of eigenvalues that determines the volume of the ellipsoid and hence the entropy, since it is much easier to calculate the trace of the observed covariance matrix $\mathrm{tr}(\widehat{\Sigma}_k)$. The following recursion is crucial for the derivation: for an assumed covariance matrix Σ, after measuring in the direction of a unit norm eigenvector u with eigenvalue λ using power β, the updated matrix takes the form of

$$\begin{aligned}\Sigma - \Sigma\sqrt{\beta}u\left(\sqrt{\beta}u^{\mathsf{T}}\Sigma\sqrt{\beta}u+\sigma^2\right)^{-1}\sqrt{\beta}u^{\mathsf{T}}\Sigma \\ = \frac{\lambda\sigma^2}{\beta\lambda+\sigma^2}uu^{\mathsf{T}} + \Sigma^{\perp u},\end{aligned} \tag{6}$$

where $\Sigma^{\perp u}$ is the component of Σ in the orthogonal complement of u. Thus, the only change in the eigendecomposition of Σ is the update of the eigenvalue of u from λ to $\lambda\sigma^2/(\beta\lambda+\sigma^2)$. Based on (6), after one measurement, the trace of the covariance matrix becomes

$$\mathrm{tr}\left(\widehat{\Sigma}_k\right) = \mathrm{tr}\left(\widehat{\Sigma}_{k-1}\right) - \frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k+\sigma^2}. \tag{7}$$

Remark 1 *The upper bound of the posterior signal entropy in (4) shows that the amount of uncertainty reduction by the kth measurement is roughly* $(s/2)\ln(1/f_k)$.

Remark 2 *Using the inequality* $\ln(1-x) \le -x$ *for* $x \in (0,1)$, *we have that in (4)*

$$\begin{aligned}\mathbb{H}[x|y_j,a_j,j\le k] &\le \frac{s}{2}\ln[2\pi e tr(\Sigma)] - \frac{1-\delta}{2}\sum_{j=1}^{k}\frac{\beta_j\hat{\lambda}_j}{\beta_j\hat{\lambda}_j+\sigma^2} \\ &= \frac{s}{2}\ln[2\pi e tr(\Sigma)] - \frac{k(1-\delta)}{2} \\ &\quad + \frac{(1-\delta)}{2}\sum_{j=1}^{k}\frac{\chi_{n,p,\varepsilon}}{\hat{\lambda}_j}.\end{aligned}$$

On the other hand, in the ideal case if the true covariance matrix is used, the posterior entropy of the signal is given by

$$\mathbb{H}_{ideal}\left[x,\left|y_j,a_j,j\le k\right.\right] = \frac{1}{2}\ln\left[(2\pi e)^s\prod_{j=1}^{s}\lambda_j\right] - \frac{1}{2}\sum_{j=1}^{k}\frac{\lambda_j}{\chi_{n,p,\varepsilon}}, \tag{8}$$

where $\tilde{\beta}_j = (1/\chi_{n,p,\varepsilon} - 1/\lambda_j)^+\sigma^2$. *Hence, we have*

$$\begin{aligned}\mathbb{H}[x|y_j,a_j,j\le k] &\le \mathbb{H}_{ideal}\left[x,\left|y_j,a_j,j\le k\right.\right] \\ &+ C - \frac{1}{2}\sum_{j=1}^{k}\left[\frac{\lambda_j}{\chi_{n,p,\varepsilon}} + (1-\delta)\left(1-\frac{\chi_{n,p,\varepsilon}}{\hat{\lambda}_j}\right)\right].\end{aligned} \tag{9}$$

where $C \triangleq (s/2)\ln[tr(\Sigma)/(\prod_{j=1}^{s}\lambda_j)^{1/s}]$ *is a constant independent of measurements. This upper bound has a nice interpretation: it characterizes the amount of uncertainty reduction with each measurement. For example, when the number of measurements required when using the assumed covariance matrix versus using the true covariance matrix are the same, we have* $\lambda_j \ge \chi_{n,p,\varepsilon}$ *and* $\hat{\lambda}_j \ge \chi_{n,p,\varepsilon}$. *Hence, the third term in (9) is upper bounded by* $-k/2$, *which means that the amount of reduction in entropy is roughly 1/2 nat per measurement.*

Remark 3 *Consider the special case where the errors only occur in the eigenvalues of the matrix but not in the eigenspace U, i.e.,* $\widehat{\Sigma} - \Sigma = U diag\{e_1,\cdots,e_s\}U^{\mathsf{T}}$ *and* $\max_{1\le j\le s}|e_j| = \delta_0$, *then the upper bound in (8) can be further simplified. Suppose only the first* K ($K \le s$) *largest eigenvalues of* $\widehat{\Sigma}$ *are larger than the stopping criterion* $\chi_{n,p,\varepsilon}$ *required by the precision, i.e., the algorithm takes K iterations in total. Then,*

$$\begin{aligned}\mathbb{H}[x|y_j,a_j,j\le k] &\le \mathbb{H}_{ideal}\left[x,\left|y_j,a_j,j\le k\right.\right] \\ &+ K\ln(1+\delta_K/\chi_{n,p,\varepsilon}) + \sum_{j=K+1}^{s}\ln(1+(\delta_0+\delta_K)/\lambda_j).\end{aligned}$$

The additional entropy relative to the ideal case $\mathbb{H}_{ideal}$ *is typically small, because* $\delta_K \le \delta_0 4^K$ *(according to Lemma 7 in the Appendix 2),* δ_0 *is on the order of* ε^2, *and hence the second term is on the order of* K^2; *the third term will be small because* δ_0 *and* δ_K *are small compare to* λ_j.

Note that, however, if the power allocations β_i are calculated using the eigenvalues of the assumed covariance matrix $\widehat{\Sigma}$, after $K = s$ iterations, we are not guaranteed to reach the desired precision ε with probability p. However, this becomes possible if we increase the total power slightly. The following theorem establishes an upper bound on the amount of extra total power needed to reach the same precision ε compared to the total power P_{ideal} if we use the correct covariance matrix.

Theorem 3 (Additional power required) *Assume* $K \le s$ *eigenvalues of* Σ *are larger than* $\chi_{n,p,\varepsilon}$. *If*

$$\|\widehat{\Sigma} - \Sigma\| \le \frac{1}{4^{s+1}}\chi_{n,p,\varepsilon},$$

then to reach a precision ε *at confidence level* p, *the total power* $P_{mismatch}$ *required by Algorithm* 1 *when using* $\widehat{\Sigma}$ *is upper bounded by*

$$P_{mismatch} < P_{ideal} + \left[\frac{20}{51}s + \frac{1}{272}K\right]\frac{\sigma^2}{\chi_{n,p,\varepsilon}}.$$

Note that in Theorem 3, when $K = s$ eigenvalues of Σ are larger than $\chi_{n,p,\varepsilon}$, under the conditions of Theorem 3, we have a simpler expression for the upper bound

$$P_{\text{mismatch}} < P_{\text{ideal}} + \frac{323}{816}\frac{\sigma^2}{\chi_{n,p,\varepsilon}}s.$$

Note that the additional power required is quite small and is only linear in s.

3.2 One-sparse measurement

In the following, we provide performance bounds for the case of one-sparse measurements in Algorithm 2. Assume the signal covariance matrix is known precisely. Now that $\|a_k\|_0 = 1$, we have $a_k = \sqrt{\beta_k}u_k$, where $u_k \in \{e_1, \cdots, e_n\}$. Suppose the largest diagonal entry of $\Sigma^{(k-1)}$ is determined by

$$j_{k-1} = \arg\max_t \Sigma_{tt}^{(k-1)}.$$

From the update equation for the covariance matrix in Algorithm 2, the largest diagonal entry of $\Sigma^{(k)}$ can be determined from

$$j_k = \arg\max_t \left\{\Sigma_{tt}^{(k-1)} - \frac{\left(\Sigma_{tj_{k-1}}^{(k-1)}\right)^2}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)} + \sigma^2/\beta_k}\right\}.$$

Let the correlation coefficient be denoted as

$$\rho_{ij}^{(k)} \triangleq \frac{\left(\Sigma_{ij}^{(k)}\right)^2}{\Sigma_{ii}^{(k)}\Sigma_{jj}^{(k)}},$$

where the covariance of the ith and jth coordinate of x after k measurements is denoted as $\Sigma_{ij}^{(k)}$.

Lemma 1 (One sparse measurement. Recursion for trace of covariance matrix) *Assume the minimum correlation for the kth iteration is* $\rho^{(k-1)} \in [0, 1)$ *such that* $\rho^{(k-1)} \le \left|\rho_{ij_{k-1}}^{(k-1)}\right|$ *for any* $i \in [n]$. *Then, for a constant* $\gamma > 0$, *if the power of the kth measurement* β_k *satisfies* $\beta_k \ge \sigma^2/\left(\gamma \max_t \Sigma_{tt}^{(k-1)}\right)$, *we have*

$$\text{tr}(\Sigma_k) \le \left[1 - \frac{(n-1)\rho^{(k-1)} + 1}{n(1+\gamma)}\right]\text{tr}(\Sigma_{k-1}). \tag{10}$$

Lemma 1 provides a good bound for a one-step ahead prediction for the trace of the covariance matrix, as demonstrated in Fig. 2. Using the above lemma, we can obtain an upper bound on the number of measurements needed for one-sparse measurements.

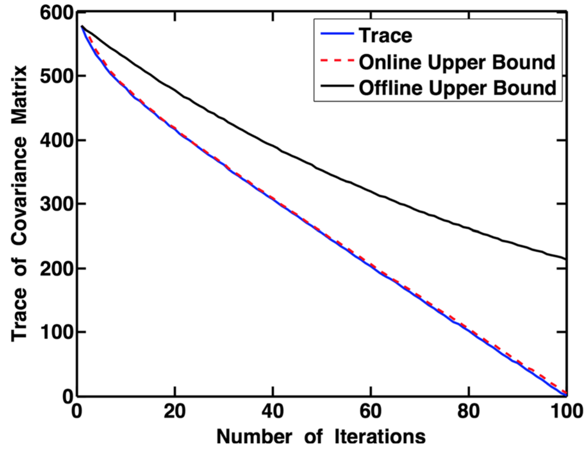

Fig. 2 One-step ahead prediction for the trace of the covariance matrix: the offline bound corresponds to applying (10) iteratively k times, and the online bound corresponds to predicting $\text{tr}(\Sigma_k)$ using $\text{tr}(\Sigma_{k-1})$. Here $n = 100$, $p = 0.95$, $\varepsilon = 0.1$, $\Sigma = dd^\intercal + 5I_n$ where $d = [1, \cdots, 1]^\intercal$

Theorem 4 (Gaussian, one-sparse measurement) *For constant* $\gamma > 0$, *when power is allocated satisfying* $\beta_k \ge \sigma^2/(\gamma \max_t \Sigma_{tt}^{(k-1)})$ *for* $k = 1, 2, \ldots, K$, *we have* $\|\hat{x} - x\| \le \varepsilon$ *with probability p as long as*

$$K \ge \frac{\ln[\,tr(\Sigma)/\chi_{n,p,\varepsilon}]}{\ln\frac{1}{1-1/[n(1+\gamma)]}}. \tag{11}$$

The above theorem requires the number of iterations to be on the order of $\ln(1/\varepsilon)$ to reach a precision of ε (recall that $\chi_{n,p,\varepsilon} = \varepsilon^2/\chi_n^2(p)$), as expected. It also suggests a method of power allocation, which sets β_k to be proportional to $\sigma^2/\max_t \Sigma_{tt}^{(k-1)}$. This captures the inter-dependence of the signal entries as the dependence will affect the diagonal entries of the updated covariance matrix.

4 Results: numerical examples

In the following, we have three sets of numerical examples to demonstrate the performance of Info-Greedy Sensing when there is mismatch in the signal covariance matrix, when the signal is sampled from Gaussian, and from GMM models, respectively. Below, in all figures, we present sorted estimation errors from the smallest to the largest over all trials.

4.1 Sensing Gaussian with mismatched covariance matrix

In the two examples below, we generate true covariance matrices using random positive semi-definite matrices. When the assumed covariance matrix for the signal x is equal to its true covariance matrix, Info-Greedy Sensing is identical to the batch method [21] (the batch method measures using the largest eigenvectors of the signal covariance matrix). However, when there is a mismatch between the two, Info-Greedy Sensing outperforms the batch method due to its adaptivity, as shown by the example demonstrated in Fig. 3 (with $K = 20$). Further performance improvement can be achieved by updating the covariance matrix using estimated signal sequentially such as described in (2). Info-Greedy Sensing also outperforms the sensing algorithm where a_i are chosen to be random Gaussian vectors with the same power allocation, as it uses prior knowledge (albeit being imprecise) about the signal distribution.

Figure 4 demonstrates an effect that when there is a mismatch in the assumed covariance matrix, better performance can be achieved if we make many lower power measurements than making one full power measurement because we update the assumed covariance matrix in between. Performance of these scenarios are compared with the case without mismatch. And it is also shown in the figure that many lower power measurements and one full power measurement perform the same when the assumed model is exact.

4.2 Measure Gaussian mixture model signals using one-sparse measurements

In this example, we sense a GMM signal with a one-sparse measurement. Assume there are $C = 3$ components and we know the signal covariance matrix exactly. We consider two cases of generating the covariance matrix for each signal: when the low-rank covariance matrices for each component are generated completely at random and when it has certain structure. In this example, we expect “Info-Greedy” to have much better performance than “Random” in the second case (b) because there is a structure in the covariance matrix. Since Info-Greedy has an advantage in exploiting structure in covariance, it should have larger performance gain. In the first case (a), the covariance matrix is generated randomly, and thus, the performance gain is not significant.

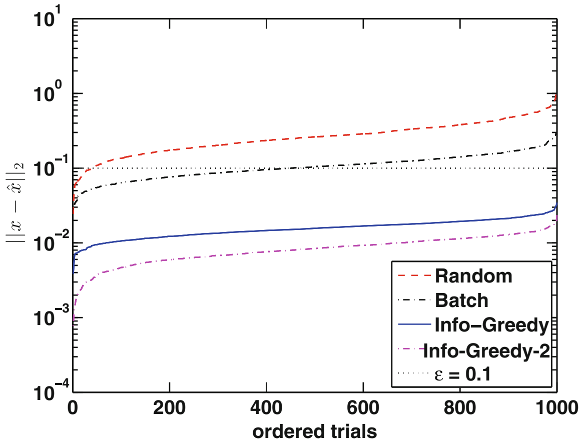

Fig. 3 Sensing a Gaussian signal of dimension $n = 100$, when there is mismatch between the assumed covariance matrix and the true covariance matrix: $\widehat{\Sigma} \propto \Sigma + RR^{\intercal}$, where $R \in \mathbb{R}^{n\times 3}$ and each entry of $R_{ij} \sim \mathcal{N}(0, 1)$. We repeat 1000 Monte Carlo trials, and for each trial, we use $K = 20$ measurements. The Info-Greedy-2 method corresponds to (2), where we update the assumed covariance matrix sequentially each time we recover a signal and $\alpha = 0.5$

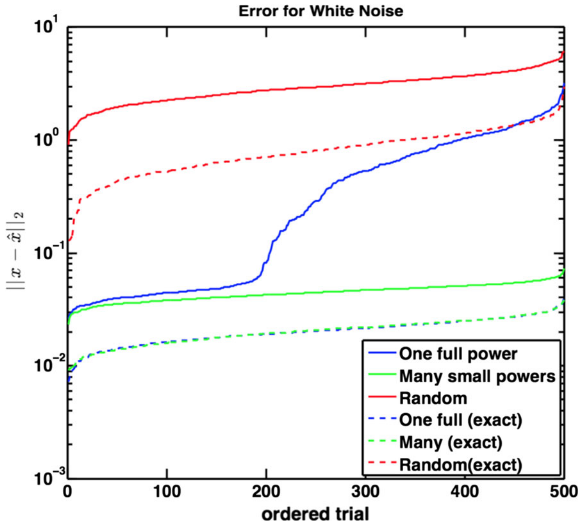

Fig. 4 Comparison of sensing a Gaussian signal with dimension $n = 100$ using unit power measurements along the eigenvector direction, versus splitting each unit power measurement into five smaller ones, each with amplitude $\sqrt{1/5}$, and we update the covariance matrix in between. The mismatched covariance matrix is $\widehat{\Sigma} \propto \Sigma + rr^{\intercal}$, where $r \in \mathbb{R}^{n\times 5}$ and each entry of r is i.i.d. $\mathcal{N}(0, 1)$, and $\widehat{\Sigma}$ is normalized to have unit spectral norm. Performance of the algorithm in the presence of mismatch is compared with that with exact parameters

Figure 5 shows the reconstruction error $\|\hat{x} - x\|$, using $K = 40$ one-sparse measurements for GMM signals. Note that Info-Greedy Sensing (Algorithm 2) with unit power $\beta_j = 1$ can significantly outperform the random approach with unit power (which corresponds to randomly selecting coordinates of the signal to measure). The experiment results validate our expectation.

4.3 Real data

4.3.1 Sensing of a video stream using Gaussian model

In this example, we use a video from the Solar Data Observatory. In this scenario, one aims to compress the high-resolution video (before storage and transmission). Each measurement corresponds to a linear compression of a frame. The frame is of size 232×292 pixels. We use the first 50 frames to form a sample covariance matrix $\widehat{\Sigma}$ and use it to perform Info-Greedy Sensing on the rest of the frames. We take $K = 90$ measurements. As demonstrated in Fig. 6, Info-Greedy Sensing performs much better in

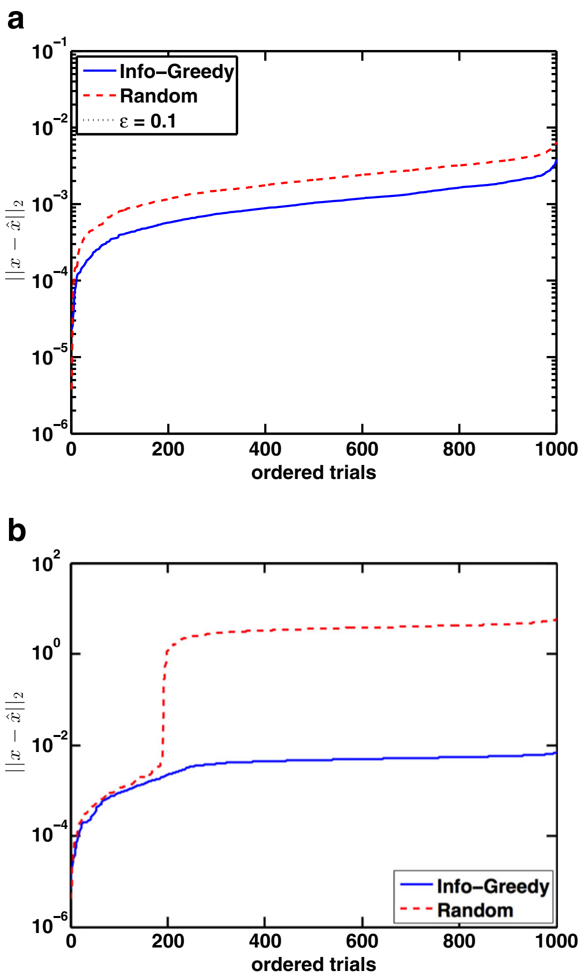

Fig. 5 Sensing a low-rank GMM signal of dimension $n = 100$ using $K = 40$ measurements with $\sigma = 0.001$, when the covariance matrices are generated **a** completely randomly, $\Sigma_c \propto RR^{\intercal}, R \in \mathbb{R}^{n\times 3}, R_{ij} \sim \mathcal{N}(0,1)$ or **b** having certain structure, $\Sigma_c \propto \left(11^{\intercal} + 20\alpha^2 \cdot \text{ diag}\{n, n-1, \cdots, 1\}\right), \alpha \sim \mathcal{N}(0,1)$. The covariance matrices Σ_c are normalized so that their spectral norms are 1

that it acquires more information such that the recovered image has much richer details.

4.3.2 *Sensing of a high-resolution image using GMM*

The second example is motivated by computational photography [35], where one takes a sequence of measurements and each measurement corresponds to the integrated light intensity through a designed mask. We consider a scheme for sensing a high-resolution image that exploits the fact that the patches of the image can be approximated using a Gaussian mixture model, as demonstrated in Fig. 1. We break the image into 8 × 8 patches, which resulted in 89250 patches. We randomly select 500 patches (0.56% of the total pixels) to estimate a GMM model with $C = 10$ components, and then based on the estimated GMM, initialize Info-Greedy Sensing with $K = 5$ measurements and sense the rest of the patches. This means we can use a compressive imaging system to capture five low-resolution images of size 238 × 275 (this corresponds to compressing the data into 8.32% of its original dimensionality). With such a small number of measurements, the recovered image from Info-Greedy Sensing measurements has superior quality compared with those with random sensing measurements.

5 Covariance sketching

We may be able to initialize $\widehat{\Sigma}$ with desired precision via covariance sketching, i.e., using fewer samples to reach a "rough" estimate of the covariance matrix. In this section, we present the covariance sketching scheme, by adapting the covariance sketching in earlier works [24, 25]. The goal here is not to present completely new covariance sketching algorithms, but rather to illustrate how to efficiently obtain initialization for Info-Greedy.

Consider the following setup for covariance sketching. Suppose we are able to form a measurement in the form of $y = a^{\intercal}x + w$ like we have in the Info-Greedy Sensing algorithm.

Suppose there are N copies of Gaussian signal, we would like to sketch $\tilde{x}_1, \ldots, \tilde{x}_N$ that are i.i.d. sampled from $\mathcal{N}(0, \Sigma)$, and we sketch using M random vectors: $b_1, \ldots, b_M$. Then, for each fixed sketching vector b_i and fixed copy of the signal $\tilde{x}_j$, we acquire L noisy realizations of the projection result y_{ijl} via

$$y_{ijl} = b_i^{\intercal}\tilde{x}_j + w_{ijl}, \quad l = 1, \ldots, L.$$

We choose the random sampling vectors b_i as i.i.d. Gaussian with zero mean and covariance matrix equal to an identity matrix. Then, we average y_{ijl} over all realizations l=$1, \ldots, L$ to form the ith sketch y_{ij} for a single copy $\tilde{x}_j$:

$$y_{ij} = b_i^{\intercal}\tilde{x}_j + \underbrace{\frac{1}{L}\sum_{l=1}^{L} w_{ijl}}_{w_{ij}}.$$

The average is introduced to suppress measurement noise, which can be viewed as a generalization of sketching using just one sample. Denote $w_{ij} \triangleq \frac{1}{L}\sum_{l=1}^{L} w_{ijl}$, which is distributed as $\mathcal{N}(0, \sigma^2/L)$. Then, we will use the average energy of the sketches as our data γ_i, $i = 1, \ldots, M$, for covariance recovery $\gamma_i \triangleq \frac{1}{N}\sum_{j=1}^{N} y_{ij}^2$. Note that γ_i can be further expanded as

$$\gamma_i = \operatorname{tr}\left(\widehat{\Sigma}_N b_i b_i^{\intercal}\right) + \frac{2}{N}\sum_{j=1}^{N} w_{ij} b_i^{\intercal}\tilde{x}_j + \frac{1}{N}\sum_{j=1}^{N} w_{ij}^2, \tag{12}$$

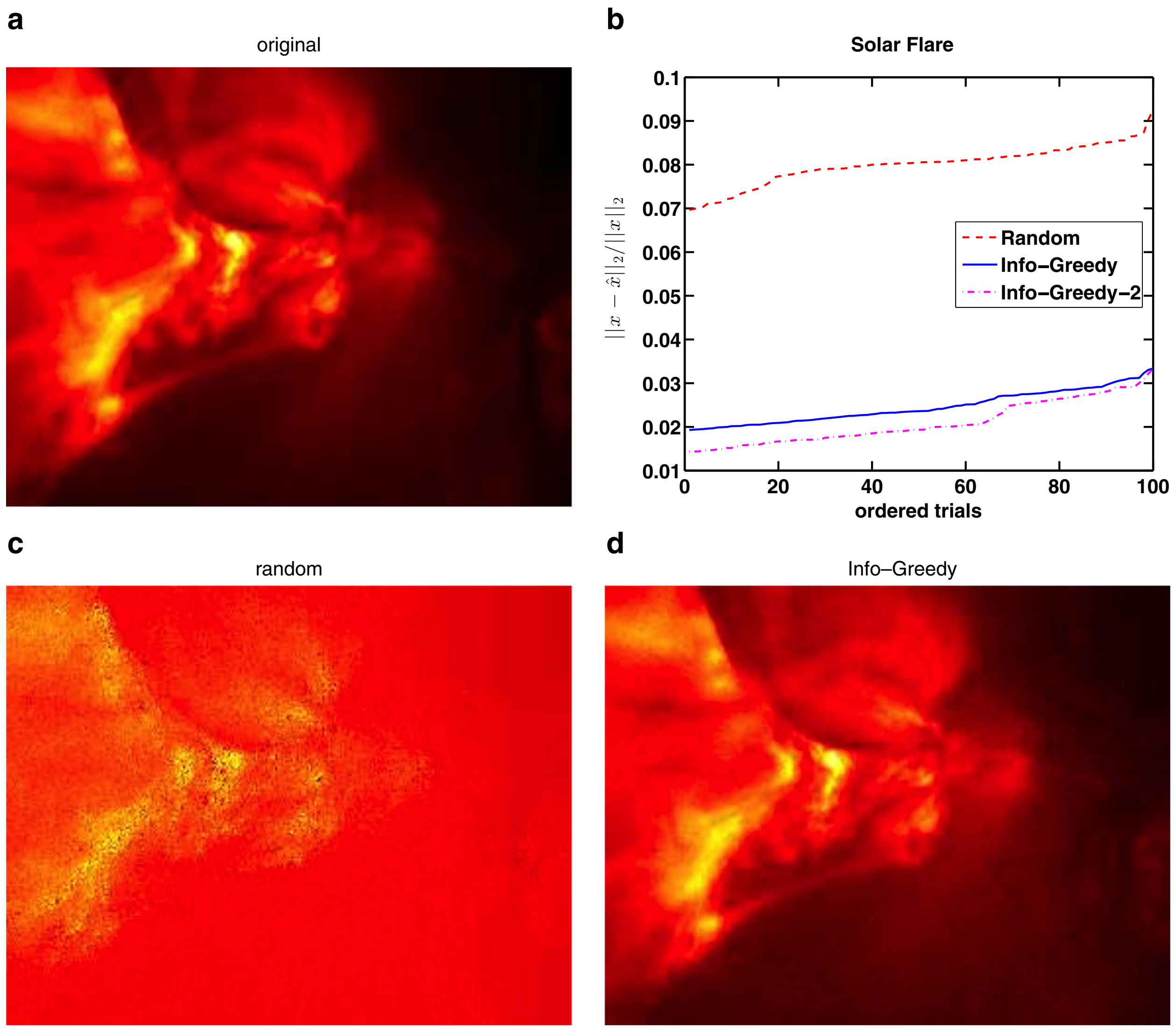

Fig. 6 Recovery of solar flare images of size 224 by 288 with $K = 90$ measurements and no sensing noise. We used the first 50 frames to estimate the mean and covariance matrix of a single Gaussian. **a** original image for 300th frame. **b** Ordered relative recovery error of the 200th to the 300th frames. **c** Recovered the 300th frame using random measurement. **d** Recovered the 300th frame using Info-Greedy Sensing

where $\widehat{\Sigma}_N \triangleq \frac{1}{N}\sum_{j=1}^N \tilde{x}_j\tilde{x}_j^\top$ is the maximum likelihood estimate of Σ (and is also unbiased). We can write (12) in vector matrix notation as follows. Let $\gamma = [\gamma_1, \cdots \gamma_M]^\top$. Define a linear operator $\mathcal{B} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^M$ such that $[B(X)]_i = \operatorname{tr}\left(X b_i b_i^\top\right)$. Thus, we can write (12) as a linear measurement of the true covariance matrix Σ $\gamma = \mathcal{B}(\Sigma) + \eta$, where $\eta \in \mathbb{R}^M$ contains all the error terms and corresponds to the noise in our covariance sketching measurements, with the ith entry given by

$$\eta_i = b_i^\top(\widehat{\Sigma}_N - \Sigma)b_i + \frac{2}{N}\sum_{j=1}^N w_{ij} b_i^\top \tilde{x}_j + \frac{1}{N}\sum_{j=1}^N w_{ij}^2.$$

Note that we can further bound the ℓ_1 norm of the error term as

$$\|\eta\|_1 = \sum_{i=1}^M |\eta_i| \le \|\widehat{\Sigma}_N - \Sigma\| b + 2\sum_{i=1}^M |z_i| + w,$$

where $b \triangleq \sum_{i=1}^M \|b_i\|^2$, $\mathbb{E}[b] = Mn$, $\mathrm{Var}[b] = 2Mn$, $w \triangleq \frac{1}{N}\sum_{i=1}^M \sum_{j=1}^N w_{ij}^2$, $\mathbb{E}[w] = M\sigma^2/L$, and $\mathrm{Var}[w] = \frac{2M\sigma^4}{NL^2}$, and

$$z_i \triangleq \frac{1}{N}\sum_{j=1}^N w_{ij} b_i^\top \tilde{x}_j,\ \mathbb{E}[z_i] = 0 \text{ and } \mathrm{Var}[z_i] = \frac{\sigma^2 \mathrm{tr}(\Sigma)}{NL}.$$

We may recover the true covariance matrix from the sketches γ using the convex optimization problem (13).

We need L to be sufficiently large to reach the desired precision. The following Lemma 2 arises from a simple tail probability bound of the Wishart distribution (since the sample covariance matrix follows a Wishart distribution).

Lemma 2 (Initialize with sample covariance matrix) *For any constant $\delta > 0$, we have $\|\widehat{\Sigma} - \Sigma\| \le \delta$ with probability exceeding $1 - 2n\exp(-\sqrt{n})$, as long as*

$$L \ge 4n^{1/2}\mathrm{tr}(\Sigma)\left(\|\Sigma\|/\delta^2 + 4/\delta\right).$$

Lemma 2 shows that the number of measurements needed to reach a precision δ for a sample covariance matrix is $\mathcal{O}(1/\delta^2)$ as expected.

We may also use a covariance sketching scheme similar to that described in [23–25] to estimate $\widehat{\Sigma}$. Covariance sketching is based on random projections of each training sample, and hence, it is memory efficient when we are not able to store or operate on the full vectors directly. The covariance sketching scheme is described below. Assume training samples $\tilde{x}_i$, $i = 1, \ldots, N$ are drawn from the signal distribution. Each sample, $\tilde{x}_i$ is sketched M times using random sketching vectors b_{ij}, $j = 1, \ldots, M$, through a noisy linear measurement $\left(b_{ij}^{\intercal}x_i + w_{ijl}\right)^2$, and we repeat this for L times ($l = 1, \ldots, L$) and compute the average energy to suppress noise[1]. This sketching process can be shown to be a linear operator $\mathcal{B}$ applied on the original covariance matrix Σ. We may recover the original covariance matrix from the vector of sketching outcomes $\gamma \in \mathbb{R}^M$ by solving the following convex optimization problem

$$\begin{aligned} \widehat{\Sigma} = \ & \mathrm{argmin}_X \ \mathrm{tr}(X) \\ & \text{subject to } X \succeq 0, \ \|\gamma - \mathcal{B}(X)\|_1 \le \tau, \end{aligned} \tag{13}$$

where τ is a user parameter that depends on the noise level. In the following theorem, we further establish conditions on the covariance sketching parameters N, M, L, and τ so that the recovered covariance matrix $\widehat{\Sigma}$ may reach the required precision in Theorem 2, by adapting the results in [25].

Lemma 3 (Initialize with covariance sketching) *For any $\delta > 0$ the solution to (13) satisfies $\|\widehat{\Sigma} - \Sigma\| \le \delta$, with probability exceeding $1 - 2/n - 2/\sqrt{n} - 2n\exp(-\sqrt{n}) - \exp(-c_1 M)$, as long as the parameters M, N, L and τ satisfy the following conditions*

$$M > c_0 ns, \quad N \ge 4n^{1/2}\mathrm{tr}(\Sigma)\left(\frac{36M^2n^2\|\Sigma\|}{\tau^2} + \frac{24Mn}{\tau}\right),$$

$$L \ge \max\left\{\frac{M}{4n^2\|\Sigma\|}\sigma^2, \frac{1}{\sqrt{2[\,\mathrm{tr}(\Sigma)/\|\Sigma\|\,]\,Mn^2}}\sigma^2, \frac{6M}{\tau}\sigma^2\right\}, \tag{14}$$

$$\tau = M\delta/c_2, \tag{15}$$

where c_0, c_1, and c_2 are absolute constants.

Finally, we present one numerical example to validate covariance sketching as initialization for Info-Greedy, as shown in Fig. 7. We compare it with the case ("direct" in the figure) when sample covariance matrix is directly estimated using original samples. The parameters are signal dimension $n = 10$; there are 30 samples and $m = 40$ sketches for each sample (thus the dimensionality reduction ratio is $40/10^2 = 0.4$); precision level $\epsilon = 0.1$; the confidence level $p = 0.95$; and noise standard deviation $\sigma_0 = 0.01$. The covariance matrix $\widehat{\Sigma}$ is obtained by solving the optimization problem (13) using standard optimization solver CVX, a package for specifying and solving convex programs [36, 37]. Note that the covariance sketching has a higher error level (to achieve dimensionality reduction); however, the errors are still below the precision level ($\epsilon = 0.1$) thus the performance of covariance sketching is acceptable.

6 Conclusions and discussions

In this paper, we have studied the robustness of sequential compressed sensing algorithm based on conditional mutual information maximization, the so-called Info-Greedy Sensing [6], when the parameters are learned from data. We quantified the algorithm performances in the presence of estimation errors. We further presented covariance sketching based scheme for initializing covariance matrices. Numerical examples demonstrated the robust performance of Info-Greedy.

Our results for Gaussian and GMM signals are quite general in the following sense. In high-dimensional prob-

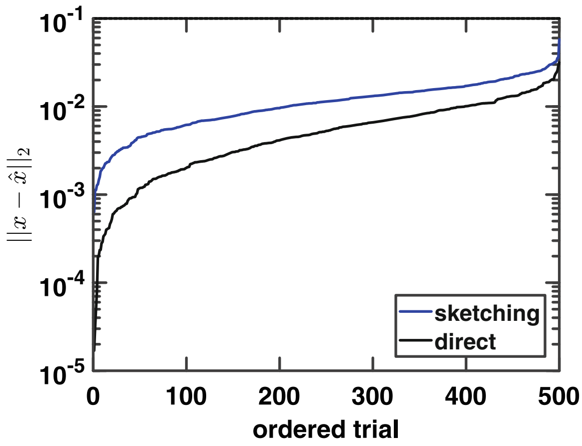

Fig. 7 Covariance sketching as initialization for Info-Greedy Sensing. Sorted estimation error in 500 trials. In this example, signal dimension $n = 10$, there are $m = 40$ sketches; thus, the dimensional reduction ratio is $40/10^2 = 0.4$. The errors of covariance sketching are higher than using the direct covariance estimation as initialization (to achieve the goal of dimensionality reduction); however, note that the errors of covariance sketching are still much below the pre-specified error tolerance $\epsilon = 0.1$ and thus are acceptable

lems, a commonly used low-dimensional signal model for x is to assume the signal lies in a subspace plus Gaussian noise, which corresponds to the case where the signal is Gaussian with a low-rank covariance matrix; GMM is also commonly used (e.g., in image analysis and video processing) as it models signals lying in a union of multiple subspaces plus Gaussian noise. In fact, parameterizing via low-rank GMMs is a popular way to approximate complex densities for high-dimensional data.

Endnote

[1] Our sketching scheme is slightly different from that used in [25] because we would like to use the square of the noisy linear measurements y_i^2 (where as the measurement scheme in [25] has a slightly different noise model). In practice, this means that we may use the same measurement scheme in the first stage as training to initialize the sample covariance matrix.

Appendix 1

Backgrounds

Lemma 4 (Eigenvalue of perturbed matrix [38]) *Let Σ, $\widehat{\Sigma} \in \mathbb{R}^{n\times n}$ be symmetric,with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$, respectively. Let $E \triangleq \widehat{\Sigma} - \Sigma$ have eigenvalues $e_1 \geq \cdots \geq e_n$. Then for each $i \in \{1, \cdots, n\}$, the perturbed eigenvalues satisfy $\hat{\lambda}_i \in [\,\lambda_i + e_n, \lambda_i + e_1]$.*

Lemma 5 (Stability conditions for covariance sketching [25]) *Denote $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ a linear operator and for $X \in \mathbb{R}^{n\times n}$, $\mathcal{A}(X) = \{a_i^T X a_i\}_{i=1}^m$. Suppose the measurement is contaminated by noise $\eta \in R^m$, i.e., $Y = \mathcal{A}(\Sigma) + \eta$ and assume $\|\eta\|_1 \leq \epsilon_1$. Then with probability exceeding $1 - \exp(-c_1 m)$ the solution $\widehat{\Sigma}$ to the trace minimization (13) satisfies*

$$\|\widehat{\Sigma} - \Sigma\|_F \leq c_0 \frac{\|\Sigma - \Sigma_r\|_*}{\sqrt{r}} + c_2 \frac{\epsilon_1}{m},$$

for all $\Sigma \in R^{n\times n}$, provided that $m > c_0 nr$. Here c_0, c_1, and c_2 are absolute constants and Σ_r represents the best rank-r approximation of Σ. When Σ_r is exactly rank-r

$$\|\widehat{\Sigma} - \Sigma\|_F \leq c_0 \frac{\epsilon_1}{m}.$$

Lemma 6 (Concentration of measure for Wishart distribution [39]) *If $X \in \mathbb{R}^{n\times n} \sim \mathcal{W}_n(N, \Sigma)$, then for $t > 0$,*

$$P\left\{\left\|\frac{1}{N}X - \Sigma\right\| \geq \left(\sqrt{\frac{2t(\theta+1)}{N}} + \frac{2t\theta}{N}\right)\|\Sigma\|\right\} \leq 2n\exp(-t),$$

where $\theta = \mathrm{tr}(\Sigma)/\|\Sigma\|$.

Appendix 2

Proofs

Gaussian signal with mismatch

Proof of Theorem 1 Let $\xi_k \triangleq \hat{\mu}_k - \mu_k$. From the update equation for the mean $\hat{\mu}_k = \hat{\mu}_{k-1} + \widehat{\Sigma}_{k-1} a_k \left(y_k - a_k^\intercal \hat{\mu}_{k-1}\right) / \left(\hat{a}_k^\intercal \widehat{\Sigma}_{k-1} a_k + \sigma^2\right)$, since a_k is eigenvector of $\hat{\Sigma}_{k-1}$, we have the following recursion:

$$\begin{aligned}\xi_k = &\left(I_n - \frac{\hat{\lambda}_k a_k a_k^\intercal}{\beta_k \hat{\lambda}_k + \sigma^2}\right)\xi_{k-1} \\ &+ \left[-\hat{\lambda}_k \frac{a_k^\intercal E_{k-1} a_k}{\left(\beta_k \hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1} a_k\right)(\beta_k \hat{\lambda}_k + \sigma^2)} a_k \right. \\ &\left. + \frac{E_{k-1} a_k}{\beta_k \hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1} a_k}\right]\left(a_k^\intercal (x - \mu_{k-1}) + w_k\right).\end{aligned} \tag{16}$$

From the recursion of ξ_k in (16), for some vector C_k defined properly, we have that

$$\begin{aligned}\mathbb{E}[\,\xi_k] = &\left(I - \frac{\hat{\lambda}_k \beta_k}{\beta_k \hat{\lambda}_k + \sigma^2} u_k u_k^\intercal\right)\mathbb{E}[\,\xi_{k-1}] \\ &+ C_k \underbrace{\mathbb{E}\left[a_k^\intercal (x - \mu_{k-1}) + w_k\right]}_{0},\end{aligned} \tag{17}$$

where the expectation is taken over random variables x and w's. Note that the second term is equal to zero using an argument based on iterated expectation

$$\mathbb{E}\left[a_k^\intercal (x - \mu_{k-1}) + w_k\right] = a_k^\intercal \mathbb{E}[\,\mathbb{E}[\,x - \mu_{k-1} | y_1, \ldots, y_k]] = 0.$$

Hence, Theorem 1 is proved by iteratively apply the recursion (17). When $\hat{\mu}_0 - \mu_0 = 0$, we have $\mathbb{E}[\,\xi_k] = 0, k = 0, 1, \ldots, K$. □

In the following, Lemma 7 to Lemma 9 are used to prove Theorem 2.

Lemma 7 (Recursion in covariance matrix mismatch.) *If $\delta_{k-1} \leq 3\sigma^2/4\beta_k$, then $\delta_k \leq 4\delta_{k-1}$.*

Proof Let $\widehat{A}_k \triangleq a_k a_k^\intercal$. Hence, $\|\widehat{A}_k\| = \beta_k$. Recall that a_k is the eigenvector of $\widehat{\Sigma}_{k-1}$, using the definition of $E_k \triangleq \widehat{\Sigma}_k - \Sigma_k$, together with the recursions of the covariance matrices

$$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} - \widehat{\Sigma}_{k-1} a_k a_k^\intercal \Sigma_{k-1} / (\hat{\lambda}_k + \sigma^2), \tag{18}$$

$$\Sigma_k = \Sigma_{k-1} - \Sigma_{k-1} a_k a_k^\intercal \Sigma_{k-1} / \left(a_k^\intercal \Sigma_{k-1} a_k + \sigma^2\right), \tag{19}$$

we have

$$E_k = E_{k-1} + \frac{\Sigma_{k-1} a_k a_k^\intercal \Sigma_{k-1}}{a_k^\intercal \Sigma_{k-1} a_k + \sigma^2} - \frac{\hat{\lambda}_k a_k a_k^\intercal \widehat{\Sigma}_{k-1}}{\beta_k \hat{\lambda}_k + \sigma^2}.$$

Based on this recursion, using $\delta_k = \|E_k\|$, the triangle inequality, and inequality $\|AB\| \leq \|A\|\|B\|$, we have

$$\begin{aligned}\delta_k &\leq \delta_{k-1} + \frac{\beta_k\hat{\lambda}_k a_k E_{k-1} a_k}{\left(\beta_k\hat{\lambda}_k + \sigma^2\right)\left(\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1} a\right)}\\ &\quad \cdot \|\widehat{A}_k\widehat{\Sigma}_{k-1}\| + \frac{1}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}\\ &\quad \cdot[\,\hat{\lambda}_k(\|\widehat{A}_k E_{k-1}\| + \|E_{k-1}\widehat{A}_k\|) + \|E_{k-1}\widehat{A}_k E_{k-1}\|]\\ &\leq \delta_{k-1} + \frac{\beta_k^2\hat{\lambda}_k^2\delta_{k-1}}{\left(\beta_k\hat{\lambda}_k + \sigma^2\right)(\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1})}\\ &\quad + \frac{\beta_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}[\,2\hat{\lambda}_k\delta_{k-1} + \delta_{k-1}^2]\\ &\leq \left(1 + \frac{3\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}\right)\delta_{k-1}\\ &\quad + \frac{\beta_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}\delta_{k-1}^2.\end{aligned}$$

Hence, if we set $\delta_{k-1} \leq 3\sigma^2/(4\beta_k)$, i.e., $\delta_{k-1}\beta_k \leq \frac{3}{4}\sigma^2$, the last inequality can be upper bounded by

$$\left(1 + 3\cdot\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2/4}\right)\delta_{k-1} + 3\cdot\frac{\sigma^2/4}{\beta_k\hat{\lambda}_k + \sigma^2/4}\delta_{k-1} = 4\delta_{k-1}.$$

Hence, if $\delta_{k-1} \leq 3\sigma^2/(4\beta_k)$, we have $\delta_k \leq 4\delta_{k-1}$. □

Lemma 8 (Recursion for trace of the true covariance matrix) *If* $\delta_{k-1} \leq \hat{\lambda}_k$,

$$\mathrm{tr}(\Sigma_k) \leq \mathrm{tr}(\Sigma_{k-1}) - \frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k + \sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}. \tag{20}$$

Proof Let $\widehat{A}_k \triangleq a_k a_k^\intercal$. Using the definition of E_k and the recursions (18) and (19), the perturbation matrix E_k after k iterations is given by

$$\begin{aligned}E_k = E_{k-1} + \hat{\lambda}_k^2\widehat{A}_k \cdot \frac{a_k^\intercal E_{k-1}a_k}{\left(\beta_k\hat{\lambda}_k + \sigma^2\right)\left(\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k\right)}\\ - \frac{\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}\cdot(\widehat{A}_k E_{k-1} + E_{k-1}\widehat{A}_k)\\ + \frac{1}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}E_{k-1}\widehat{A}_k E_{k-1}.\end{aligned} \tag{21}$$

Note that $\mathrm{rank}(\widehat{A}_k) = 1$, thus $\mathrm{rank}(\widehat{A}_k E_{k-1}) \leq 1$; therefore, it has at most one non-zero eigenvalue,

$$\begin{aligned}\left|\mathrm{tr}\left(\widehat{A}_k E_{k-1}\right)\right| = \left|\mathrm{tr}\left(E_{k-1}\widehat{A}_k\right)\right| = \left\|\widehat{A}_k E_{k-1}\right\| \leq \left\|\widehat{A}_k\right\|\|E_{k-1}\|\\ = \beta_k\delta_{k-1}.\end{aligned}$$

Note that E_{k-1} is symmetric and $\hat{A}_k$ is positive semi-definite, we have $\mathrm{tr}(E_{k-1}\widehat{A}_k E_{k-1}) \geq 0$. Hence, from (21) we have

$$\begin{aligned}\mathrm{tr}(E_k) = \mathrm{tr}(\widehat{\Sigma}_k) - \mathrm{tr}(\Sigma_k) &\geq \mathrm{tr}(E_{k-1})\\ &\quad - \frac{3\beta_k\hat{\lambda}_k\left(\beta_k\hat{\lambda}_k + \frac{2\sigma^2}{3}\right)\delta_{k-1}}{(\beta_k\hat{\lambda}_k + \sigma^2)\left(\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}\right)}\\ &\geq \mathrm{tr}(E_{k-1}) - \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}.\end{aligned}$$

After rearranging terms we obtain

$$\begin{aligned}\mathrm{tr}(\Sigma_k) &\leq \mathrm{tr}(\Sigma_{k-1}) + \left[\mathrm{tr}\left(\widehat{\Sigma}_k\right) - \mathrm{tr}\left(\widehat{\Sigma}_{k-1}\right)\right]\\ &\quad + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}.\end{aligned}$$

Together with the recursion for trace of $\mathrm{tr}(\widehat{\Sigma}_k)$ in (7), we have

$$\mathrm{tr}(\Sigma_k) \leq \mathrm{tr}(\Sigma_{k-1}) - \frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k + \sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}.$$

□

Lemma 9 *For a given positive semi-definite matrix* $X \in \mathbb{R}^{n\times n}$*, and a vector* $h \in \mathbb{R}^n$*, if*

$$Y = X - \frac{1}{h^\intercal X h + \sigma^2}Xhh^\intercal X,$$

then rank $(X) = $ *rank*(Y).

Proof Apparently, for all $x \in \ker(X)$, $Yx = 0$, i.e., $\ker(X) \subset \ker(Y)$. Decompose $X = Q^\intercal Q$. For all $x \in \ker(Y)$, let $b \triangleq Qh$, $z \triangleq Qx$. If $b = 0$, $Y = X$; otherwise, when $b \neq 0$, we have

$$0 = x^\intercal Yx = z^\intercal z - \frac{z^\intercal bb^\intercal z}{b^\intercal b + \sigma^2}.$$

Thus,

$$z^\intercal z = \frac{z^\intercal bb^\intercal z}{b^\intercal b + \sigma^2} \leq \frac{b^\intercal b}{b^\intercal b + \sigma^2}z^\intercal z.$$

Therefore $z = 0$, i.e., $x \in \ker(X)$, $\ker(Y) \subset \ker(X)$. This shows that $\ker(X) = \ker(Y)$ or equivalently $\mathrm{rank}(X) = \mathrm{rank}(Y)$. □

Proof of Theorem 2 Recall that for $k = 1, \ldots, K$, $\hat{\lambda}_k \geq \chi_{n,p,\varepsilon}$. Using Lemma 7, we can show that for some $0 < \delta < 1$, if

$$\delta_0 \leq \delta\chi_{n,p,\varepsilon}/4^{K+1} \leq 3\sigma^2/\left(4^{K+1}\beta_1\right), \tag{22}$$

then for the first K measurements, we have

$$\delta_k \leq \frac{1}{4^{K-k+1}}\frac{\delta\chi_{n,p,\varepsilon}}{4} \leq \frac{1}{4^{K-k}}\frac{3\sigma^2}{4\beta_1}, \quad k = 1, \ldots, K.$$

Note that the second inequality in (22) comes from the fact that $(1/\chi_{n,p,\varepsilon} - 1/\hat{\lambda}_1)\chi_{n,p,\varepsilon}\sigma^2 \le 3\sigma^2$. Clearly, $\delta_{k-1} \le \delta\chi_{n,p,\varepsilon}/16$. Hence, $(4+\delta)\delta_{k-1} \le \delta\lambda_k$. Note that $\beta_k\delta_{k-1} \le \sigma^2$ and $|\lambda_k - \hat{\lambda}_k| \le \delta_{k-1}$, we have $\beta_k\lambda_k \le \beta_k(\hat{\lambda}_k + \delta_{k-1}) \le \beta_k\hat{\lambda}_k + \sigma^2$. Thus, $4\delta_{k-1}\left(\beta_k\hat{\lambda}_k + \sigma^2\right) + \delta\beta_k\lambda_k\delta_{k-1} \le \delta\lambda_k\left(\beta_k\hat{\lambda}_k + \sigma^2\right)$. Then, we have $3\beta_k\hat{\lambda}_k\delta_{k-1}\left(\beta_k\hat{\lambda}_k + \sigma^2\right) \le \beta_k\hat{\lambda}_k(\delta\lambda_k - \delta_{k-1})(\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1})$, which can be rewritten as $\frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \le \frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}(\delta\lambda_k - \delta_{k-1})$. Hence, $\frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \le \frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}[(\delta-1)\lambda_k + \hat{\lambda}_k]$, which can be written as $-\frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k+\sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \le -(1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\lambda_k$. By applying Lemma 8, we have

$$\mathrm{tr}(\Sigma_k) \le \mathrm{tr}(\Sigma_{k-1}) - (1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\lambda_k \le \mathrm{tr}(\Sigma_{k-1}) - (1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\frac{\mathrm{tr}(\Sigma_{k-1})}{s} \triangleq f_k\mathrm{tr}(\Sigma_{k-1}),$$

where we have used the definition for f_k in (5). Subsequently,

$$\mathrm{tr}\left(\Sigma_k) \le \left(\prod_{j=1}^{k} f_j\right)\mathrm{tr}(\Sigma_0\right).$$

Lemma 9 shows that the rank of the covariance will not be changed by updating the covariance matrix sequentially: $\mathrm{rank}(\Sigma_1) = \cdots = \mathrm{rank}(\Sigma_k) = s$. Hence, we may decompose the covariance matrix $\Sigma_k = QQ^\intercal$, with $Q \in \mathbb{R}^{n\times s}$ being a full-rank matrix, then $\mathsf{Vol}(\Sigma_k) = \det(Q^\intercal Q)$. Since $\mathrm{tr}(Q^\intercal Q) = \mathrm{tr}(QQ^\intercal)$, we have

$$\mathsf{Vol}^2(\Sigma_k) = \det(Q^\intercal Q) \overset{(1)}{\le} prod_{j=1}^{s}(Q^\intercal Q)_{jj} \overset{(2)}{\le} \left(\frac{\mathrm{tr}(Q^\intercal Q)}{s}\right)^s = \left(\frac{\mathrm{tr}(\Sigma_k)}{s}\right)^s,$$

where (1) follows from the Hadamard's inequality and (2) follows from the inequality of arithmetic and geometric means. Finally, we can bound the conditional entropy of the signal as

$$\mathbb{H}[x|y_j, a_j, j \le k] = \ln(2\pi e)^{s/2}\mathsf{Vol}(\Sigma_k) \le \frac{s}{2}\ln\left\{2\pi e\left(\prod_{j=1}^{k} f_j\right)\mathrm{tr}(\Sigma_0)\right\}, \tag{23}$$

which leads to the desired result. □

Proof of Theorem 3 Recall that $\mathrm{rank}(\Sigma) = s$, and hence $\lambda_k = 0, k = s+1, \ldots, n$. Note that for each iteration, the eigenvalue of $\widehat{\Sigma}_k$ in the direction of a_k, which corresponds to the largest eigenvalue of $\widehat{\Sigma}_k$, is eliminated below the threshold $\chi_{n,p,\varepsilon}$. Therefore, as long as the algorithm continues, the largest eigenvalue of $\widehat{\Sigma}_k$ is exactly the $(k+1)$th largest eigenvalue of $\widehat{\Sigma}$. Now, if

$$\delta_0 \le \chi_{n,p,\varepsilon}/4^{s+1}, \tag{24}$$

using Lemma 4 and Lemma 7, we have that

$$|\hat{\lambda}_k - \lambda_k| \le \delta_0, \text{ for } k = 1, \ldots, s, \quad |\hat{\lambda}_j| \le \delta_0 \le \chi_{n,p,\varepsilon} - \delta_s, \text{ for } k = s+1, \ldots, n.$$

In the ideal case without perturbation, each measurement decreases the eigenvalue along a given eigenvector to be below $\chi_{n,p,\varepsilon}$. Suppose in the ideal case, the algorithm terminates at $K \le s$ iterations, which means

$$\lambda_1 \ge \cdots \ge \lambda_L \ge \chi_{n,p,\varepsilon} > \lambda_{K+1}(\Sigma) \ge \cdots \ge \lambda_s(\Sigma),$$

and the total power needed is

$$P_{\text{ideal}} = \sum_{k=1}^{K}\sigma^2\left(\frac{1}{\chi_{n,p,\varepsilon}} - \frac{1}{\lambda_k}\right). \tag{25}$$

On the other hand, in the presence of perturbation, the algorithm will terminate using more than K iterations since with perturbation, eigenvalues of Σ that are originally below $\chi_{n,p,\varepsilon}$ may get above $\chi_{n,p,\varepsilon}$. In this case, we will also allocate power while taking into account the perturbation:

$$\beta_k = \sigma^2\left(\frac{1}{\chi_{n,p,\varepsilon} - \delta_s} - \frac{1}{\hat{\lambda}_k}\right).$$

This suffices to eliminate even the smallest eigenvalue to be below threshold $\chi_{n,p,\varepsilon}$ since

$$\frac{\sigma^2\hat{\lambda}_{k-1}}{\beta_{k-1}\hat{\lambda}_{k-1} + \sigma^2} = \chi_{n,p,\varepsilon} - \delta_s < \chi_{n,p,\varepsilon}.$$

We first estimate the total amount of power used at most to eliminate eigenvalues $\hat{\lambda}_k$, for $K+1 \le k \le s$:

$$\beta_k = \sigma^2(1/(\chi_{n,p,\varepsilon} - \delta_s) - 1/\hat{\lambda}_k) \le \sigma^2(1/(\chi_{n,p,\varepsilon} - \delta_s) - 1/(\chi_{n,p,\varepsilon} + \delta_0)) \le \sigma^2\frac{(4^s+1)\delta_0}{(\chi_{n,p,\varepsilon} - 4^s\delta_0)(\chi_{n,p,\varepsilon} + \delta_0)} \le \frac{20}{51}\frac{\sigma^2}{\chi_{n,p,\varepsilon}}.$$

where we have used the fact that $\delta_s \le 4^s\delta_0$ (a consequence of Lemma 7), the assumption (24), and monotonicity of the upper bound in s. The total power to reach precision ε in the presence of mismatch can be upper bounded by

$$P_{\text{mismatch}} \le \sum_{k=1}^{s}\beta_k \le \sigma^2\left\{\sum_{k=1}^{K}\left(\frac{1}{\chi_{n,p,\varepsilon} - \delta_s} - \frac{1}{\hat{\lambda}_k}\right) + \frac{20(s-K)}{51}\frac{\sigma^2}{\chi_{n,p,\varepsilon}}\right\}.$$

In order to achieve precision ε and confidence level p, the extra power needed is upper bounded as

$$\begin{aligned}
P_{\text{mismatch}} - P_{\text{ideal}} &\le \sigma^2 \left\{ \sum_{k=1}^{K} \left(\frac{1}{3} \frac{1}{\chi_{n,p,\varepsilon}} + \frac{\delta_0}{\lambda_k^2} \right) \right. \\
&\qquad \left. + \frac{20(s-K)}{51} \frac{1}{\chi_{n,p,\varepsilon}} \right\} \\
&\le \sigma^2 \left\{ \frac{1}{4^{s+1}} \sum_{k=1}^{K} \frac{\chi_{n,p,\varepsilon}}{\lambda_k^2} + \frac{20s-3K}{51} \frac{1}{\chi_{n,p,\varepsilon}} \right\} \\
&< \left(\frac{20}{51} s - \left(\frac{3}{51} - \frac{1}{4^{s+1}} \right) K \right) \frac{\sigma^2}{\chi_{n,p,\varepsilon}} \\
&\le \left(\frac{20}{51} s + \frac{1}{272} K \right) \frac{\sigma^2}{\chi_{n,p,\varepsilon}},
\end{aligned}$$

where we have again used $\delta_s \le 4^s \delta_0 \le 4^s \chi_{n,p,\varepsilon}/4^{s+1} = \chi_{n,p,\varepsilon}/4$, $1/\hat{\lambda}_k - 1/\lambda_k \le \delta_0/\lambda_k^2$, the fact that $\lambda_k \ge \chi_{n,p,\varepsilon}$ for $k = 1, \ldots, K$. □

Proof of Lemma 2 It is a direct consequence of Lemma 6. Let $\theta = \text{tr}(\Sigma)/\|\Sigma\| \ge 1$. For some constant $\delta > 0$, set

$$L \ge 4n^{1/2}\text{tr}(\Sigma)(\|\Sigma\|/\delta^2 + 4/\delta).$$

Then, from Lemma 6, we have

$$\begin{aligned}
P\left\{\|\widehat{\Sigma} - \Sigma\| \le \delta\right\} \ge P &\left\{ \|\widehat{\Sigma} - \Sigma\| \right. \\
&\left. \le \left(\sqrt{2n^{1/2}(\theta+1)/L} + 2\theta n^{1/2}/L \right) \|\Sigma\| \right\} \\
&> 1 - 2n\exp(-\sqrt{n}).
\end{aligned}$$

□

The following Lemma is used in the proof of Lemma 3.

Lemma 10 *If for some constants M, N, and L that satisfy the conditions in Lemma* 3*, then* $\|\eta\|_1 \le \tau$ *with probability exceeding* $1 - 2/n - 2/\sqrt{n} - 2n\exp(-c_1 M)$ *for some universal constant* $c_1 > 0$.

Proof Let $\theta \triangleq \text{tr}(\Sigma)/\|\Sigma\|$. With Chebyshev's inequality, we have that

$$\mathbb{P}\left\{|z_i| < \frac{\tau}{6M}\right\} \ge 1 - \frac{36M^2\sigma^2\text{tr}(\Sigma)}{NL\tau^2}, \quad i = 1, \ldots, K,$$

$$\mathbb{P}\left\{|w| < M\frac{\sigma^2}{L} + \frac{\tau}{6}\right\} \ge 1 - \frac{72\sigma^4 M}{NL^2\tau^2},$$

and

$$\mathbb{P}\left\{|b| < (M + \sqrt{M})n\right\} \ge 1 - \frac{2}{n}.$$

When

$$N \ge 4n^{1/2}\text{tr}(\Sigma)\left(\frac{36n^2M^2\|\Sigma\|}{\tau^2} + \frac{24nM}{\tau}\right), \tag{26}$$

with the concentration inequality for Wishart distribution in Lemma 6 and plugging in the lower bound for N in (26) and the definition for τ in (15), we have

$$\begin{aligned}
\mathbb{P}\{\|\widehat{\Sigma}_N - \Sigma\| \le \tau/[3n(M+\sqrt{M})]\} \ge \mathbb{P}\{&\|\widehat{\Sigma}_N - \Sigma\| \\
&\le \left(\sqrt{\frac{2n^{1/2}\theta}{N}} + \frac{2\theta n^{1/2}}{N} \right) \|\Sigma\|\} \\
&> 1 - 2n\exp(-\sqrt{n}).
\end{aligned}$$

Furthermore, when L satisfies (14), we have

$$\mathbb{P}\left\{|z_i| < \frac{\tau}{6M}\} \ge 1 - \frac{1}{M\sqrt{n}}, \quad \mathbb{P}\{|w| < \frac{\tau}{3}\right\} \ge 1 - \frac{1}{\sqrt{n}},$$

$$\mathbb{P}\left\{|b| < (M+\sqrt{M})n\right\} \ge 1 - \frac{2}{n}.$$

Therefore, $\|\eta\|_1 \le \tau$ holds with probability at least $1 - 2/n - 2/\sqrt{n} - 2n\exp(-\sqrt{n})$. □

Proof of Lemma 3 With Lemma 10, let $\tau = M\delta/c_2$, the choices of M, N, and L ensure that $\|\eta\|_1 \le M\delta/c_2$ with probability at least $1 - 2/n - 2/\sqrt{n} - 2n\exp(-\sqrt{n})$. By Lemma 5 in Appendix 1 and noting that the rank of Σ is s, we have $\|\widehat{\Sigma} - \Sigma\|_F \le \delta$. Therefore, with probability exceeding $1-2/n-2/\sqrt{n}-2n\exp(-\sqrt{n})-\exp(-c_0c_1ns)$, $\|\widehat{\Sigma} - \Sigma\| \le \|\widehat{\Sigma} - \Sigma\|_F \le \delta$. □

The proof will use the following two lemmas.

Lemma 11 (Moment generating function of multivariate Gaussian [40]) *Assume* $X \sim \mathcal{N}(0, \Sigma)$*. The moment generating function of* $\|X\|_2$ *is* $\mathbb{E}[e^{s\|X\|_2}] = 1/\sqrt{I - 2s\Sigma}$.

Note that $|\varrho_k|$ can be computed recursively. We may derive a recursion. Let $z_k \triangleq a_k^\intercal(x - \mu_{k-1}) + w_k = y_k - a_k^\intercal\mu_{k-1}$. Also Let $\varrho_k \triangleq a^\intercal(\hat{\mu}_k - \mu_k)$. Note that $\varrho_k = a^\intercal\xi_k$ for $\xi_k = \hat{\mu}_k - \mu_k$ in (16). Based on the recursion for ξ_k in (16) that we derived earlier, we have

$$\varrho_k = \frac{\sigma^2}{\beta_k\hat{\lambda}_k + \sigma^2}\left[\varrho_{k-1} + \frac{a_k^\intercal E_{k-1}a_k\left(y_k - a_k^\intercal\mu_{k-1}\right)}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}\right]$$

and

$$|\varrho_k| \le \frac{1}{\hat{\lambda}_k(\beta_k/\sigma^2) + 1}\left[|\varrho_{k-1}| + \frac{\delta_k}{(\hat{\lambda}_k - \delta_k) + \sigma^2/\beta_k}|z_k|\right].$$

Proof of Lemma 1 The recursion of the diagonal entries can be written as

$$\begin{aligned}
\Sigma_{ii}^{(k)} &= \Sigma_{ii}^{(k-1)} - \frac{\left(\Sigma_{ij_{k-1}}^{(k-1)}\right)^2}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)} + \sigma^2/\beta_k} \\
&= \frac{\Sigma_{ii}^{(k-1)}\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}\left(1 - \rho_{ij_{k-1}}^{(k-1)}\right) + \Sigma_{ii}^{(k-1)}\sigma^2/\beta_k}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)} + \sigma^2/\beta_k}.
\end{aligned}$$

Note that for $i = j_{k-1}$,

$$\Sigma^{(k)}_{j_{k-1}j_{k-1}} = \frac{\Sigma^{(k-1)}_{j_{k-1}j_{k-1}}\sigma^2/\beta_k}{\Sigma^{(k-1)}_{j_{k-1}j_{k-1}} + \sigma^2/\beta_k} \le \frac{\gamma}{1+\gamma}\Sigma^{(k-1)}_{j_{k-1}j_{k-1}},$$

and for $i \neq j_{k-1}$,

$$\begin{aligned}\Sigma^{(k)}_{ii} &\le \frac{\Sigma^{(k-1)}_{ii}\Sigma^{(k-1)}_{j_{k-1}j_{k-1}}\left(1-\rho^{(k-1)}\right) + \Sigma^{(k-1)}_{ii}\sigma^2/\beta_k}{\Sigma^{(k-1)}_{j_{k-1}j_{k-1}} + \sigma^2/\beta_k}\\ &\le \Sigma^{(k-1)}_{ii}\frac{\Sigma^{(k-1)}_{j_{k-1}j_{k-1}}\left(1-\rho^{(k-1)}\right) + \sigma^2/\beta_k}{\Sigma^{(k-1)}_{j_{k-1}j_{k-1}} + \sigma^2/\beta_k}\\ &\le \Sigma^{(k-1)}_{ii}\frac{1-\rho^{(k-1)}+\gamma}{1+\gamma}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{tr}(\Sigma_k) &\le \left(1 - \frac{\rho^{(k-1)}}{1+\gamma}\right)\mathrm{tr}(\Sigma_{k-1}) - \frac{1-\rho^{(k-1)}}{1+\gamma}\Sigma^{(k-1)}_{j_{k-1}j_{k-1}}\\ &\le \left[1 - \frac{(n-1)\rho^{(k-1)}+1}{n(1+\gamma)}\right]\mathrm{tr}(\Sigma_{k-1}).\end{aligned}$$

□

Proof of Theorem 4 Let $\varepsilon \ge \sqrt{\|\Sigma_K\| \cdot \chi^2_n(p)}$, i.e. $\|\Sigma_K\| \le \chi_{n,p,\varepsilon}$. Then, Theorem 4 follows from

$$\begin{aligned}&\mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[\,\|x-\mu_K\|_2 \le \varepsilon]\\ &\ge \mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[\,\|x-\mu_K\|_2 \le \sqrt{\|\Sigma_K\|\cdot\varepsilon^2}]\\ &\ge \mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[\,(x-\mu_K)^{\intercal}{\Sigma_K}^{-1}(x-\mu_K) \le \chi^2_n(p)] = p.\end{aligned} \tag{27}$$

This says that, if $\|\Sigma_K\| \le \chi_{n,p,\varepsilon}$, then (27) holds, we have $\|\hat{x}-x\| \le \varepsilon$ with probability at least p. From Lemma 1, we have that when the powers β_i are sufficiently large

$$\|\Sigma_K\| \le \mathrm{tr}(\Sigma_K) \le \left(1 - \frac{1}{n(1+\gamma)}\right)^K \mathrm{tr}(\Sigma).$$

Hence, for (27) to hold, we can simple require $\left(1-\frac{1}{n(1+\gamma)}\right)^K \mathrm{tr}(\Sigma) \le \chi_{n,p,\varepsilon}$, or equivalently (11) in Theorem 4. □

Abbreviations

GMM: Gaussian mixture models; NDT: Non-destructive testing

Acknowledgements

We would like to acknowledge Tsinghua University for supporting Ruiyang Song while he visited at Georgia Institute of Technology.

Funding

This work is partially supported by an NSF CAREER Award CMMI-1452463, an NSF grant CCF-1442635 and an NSF grant CMMI-1538746. Ruiyang Song was visiting the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology while working on this paper.

Availability of data and materials

The Georgia Tech campus image is available at www2.isye.gatech.edu/~yxie77/campus.mat and the data for solar flare image is at www2.isye.gatech.edu/~yxie77/data_193.mat.

Authors' contributions

We consider a class of mutual information maximization-based algorithms which are called Info-Greedy algorithms, and we present a rigorous performance analysis for these algorithms in the presence of model parameter mismatch. All authors read and approved the final manuscript.

Competing interests

The author declares that he/she has no competing interests.

Publisher's Note

Author details

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Received: 11 October 2017 Accepted: 1 May 2018
Published online: 05 June 2018

References

1. A Ashok, P Baheti, MA Neifeld, Compressive imaging system design using task-specific information. Appl. Opt. **47**(25), 4457–4471 (2008)
2. J Ke, A Ashok, M Neifeld, Object reconstruction from adaptive compressive measurements in feature-specific imaging. Appl. Opt. **49**(34), 27–39 (2010)
3. A Ashok, MA Neifeld, Compressive imaging: hybrid measurement basis design. J. Opt. Soc. Am. A. **28**(6), 1041–1050 (2011)
4. W Boonsong, W Ismail, Wireless monitoring of household electrical power meter using embedded RFID with wireless sensor network platform. Int. J. Distrib. Sens. Networks. **2014**(876914), 10 (2014)
5. B Zhang, X Cheng, N Zhang, Y Cui, Y Li, Q Liang, in *Sparse Target Counting and Localization in Sensor Networks Based on Compressive Sensing*. IEEE Int. Conf. Computer Communications (INFOCOM), (2014), pp. 2255–2258
6. G Braun, S Pokutta, Y Xie, Info-greedy sequential adaptive compressed sensing. IEEE J. Sel. Top. Signal Proc. **9**(4), 601–611 (2015)
7. J Haupt, R Nowak, R Castro, in *Adaptive Sensing for Sparse Signal Recovery*. IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop (DSP/SPE), (2009), pp. 702–707
8. A Tajer, HV Poor, Quick search for rare events. IEEE Transactions on Information Theory. **59**(7), 4462–4481 (2013)
9. D Malioutov, S Sanghavi, A Willsky, Sequential compressed sensing. IEEE J. Sel. Topics Sig. Proc. **4**(2), 435–444 (2010)
10. J Haupt, R Baraniuk, R Castro, R Nowak, in *Sequentially Designed Compressed Sensing*. Proc. IEEE/SP Workshop on Statistical Signal Processing, (2012)
11. A Krishnamurthy, J Sharpnack, A Singh, in *Recovering Graph-structured Activations Using Adaptive Compressive Measurements*. Annual Asilomar Conference on Signals, Systems, and Computers, (2013)
12. J Haupt, R Castro, R Nowak, in *International Conference on Artificial Intelligence and Statistics*. Distilled sensing: Selective sampling for sparse signal recovery, (2009), pp. 216–223
13. MA Davenport, E Arias-Castro, in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*. Compressive binary search, (2012), pp. 1827–1831
14. ML Malloy, RD Nowak, Near-optimal adaptive compressed sensing. IEEE Trans. Inf. Theory. **60**(7), 4001–4012 (2014)
15. S Jain, A Soni, J Haupt, in *Signals, Systems and Computers, 2013 Asilomar Conference On*. Compressive measurement designs for estimating structured signals in structured clutter: a Bayesian experimental design approach, (2013), pp. 163–167
16. E Tanczos, R Castro, Adaptive sensing for estimation of structure sparse signals. arXiv:1311.7118 (2013)
17. A Soni, J Haupt, On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. IEEE Trans. Info. Theory. **60**(1), 133–149 (2014)
18. HS Chang, Y Weiss, WT Freeman, Informative sensing. arXiv preprint arXiv:0901.4275 (2009)
19. S Ji, Y Xue, L Carin, Bayesian compressive sensing. IEEE Trans. Sig. Proc. **56**(6), 2346–2356 (2008)

20. JM Duarte-Carvajalino, G Yu, L Carin, G Sapiro, Task-driven adaptive statistical compressive sensing of gaussian mixture models. IEEE Trans. Signal Process. **61**(3), 585–600 (2013)
21. W Carson, M Chen, R Calderbank, L Carin, Communication inspired projection design with application to compressive sensing. SIAM J. Imaging Sci (2012)
22. DJC MacKay, Information based objective functions for active data selection. Comput. Neural Syst. **4**(4), 589–603 (1992)
23. G Dasarathy, P Shah, BN Bhaskar, R Nowak, in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference On*. Covariance Sketching, (2012)
24. G Dasarathy, P Shah, BN Bhaskar, R Nowak, Sketching sparse matrices. ArXiv ID:1303.6544 (2013)
25. Y Chen, Y Chi, AJ Goldsmith, Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Trans. Inf. Theory. **61**(7), 4034–4059 (2015)
26. C Hellier, *Handbook of Nondestructive Evaluation*. (McGraw-Hill, 2003)
27. P Schniter, in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop On*. Exploiting structured sparsity in bayesian experimental design, (2011), pp. 357–360
28. G Yu, G Sapiro, Statistical compressed sensing of Gaussian mixture models. IEEE Trans. Signal Process. **59**(12), 5842–5858 (2011)
29. Y Li, Y Chi, C Huang, L Dolecek, Orthogonal matching pursuit on faulty circuits. IEEE Transactions on Communications. **63**(7), 2541–2554 (2015)
30. H Robbins, in *Herbert Robbins Selected Papers*. Some aspects of the sequential design of experiments (Springer, 1985), pp. 169–177
31. CF Wu, M Hamada, *Experiments: Planning, Analysis, and Optimization*, vol. 552. (Wiley, 2011)
32. R Gramacy, D Apley, Local Gaussian process approximation for large computer experiments. J. Comput. Graph. Stat. **(just-accepted)**, 1–28 (2014)
33. D Palomar, Verdú, Gradient of mutual information in linear vector Gaussian channels. IEEE Trans. Info. Theory. **52**, 141–154 (2006)
34. Payaró, DP Palomar, Hessian and concavity of mutual information, entropy, and entropy power in linear vector Gaussian channels. IEEE Trans. Info. Theory, 3613–3628 (2009)
35. DJ Brady, *Optical Imaging and Spectroscopy*. (Wiley-OSA, 2009)
36. M Grant, S Boyd, CVX: Matlab Software for Disciplined Convex Programming, version 2.1 (2014). http://cvxr.com/cvx
37. M Grant, S Boyd, in *Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences*, ed. by V Blondel, S Boyd, and H Kimura. Graph implementations for nonsmooth convex programs (Springer, 2008), pp. 95–110
38. GW Stewart, J-G Sun, *Matrix Perturbation Theory*. (Academic Press, Inc., 1990)
39. S Zhu, A short note on the tail bound of Wishart distribution. arXiv:1212.5860 (2012)
40. T Vincent, L Tenorio, M Wakin, Concentration of measure: fundamentals and tools. Lect. Notes Rice Univ (2015)