

The Performance of a Matched Subspace Detector that Uses Subspaces Estimated from Finite, Noisy, Training Data

Nicholas Asendorf and Raj Rao Nadakuditi

Abstract—We analyze the performance of a matched subspace detector (MSD) where the test signal vector is assumed to reside in an unknown, low-rank k subspace that must be estimated from finite, noisy, signal-bearing training data. Under both a stochastic and deterministic model for the test vector, subspace estimation errors due to limited training data degrade the performance of the standard plug-in detector, relative to that of an oracle detector. To avoid some of this performance loss, we utilize and extend recent results from random matrix theory (RMT) that precisely quantify the quality of the subspace estimate as a function of the eigen-SNR, dimensionality of the system, and the number of training samples. We exploit this knowledge of the subspace estimation accuracy to derive from first-principles a new RMT detector and to characterize the associated ROC performance curves of the RMT and plug-in detectors. Using more than the a critical number of *informative* components, which depends on the training sample size and eigen-SNR parameters of training data, will result in a performance loss that our analysis quantifies in the large system limit. We validate our asymptotic predictions with simulations on moderately sized systems.

Index Terms—Matched subspace detector, deterministic, stochastic, random matrix theory, ROC analysis

I. INTRODUCTION

MANY signal processing [1] and machine learning [2] applications involve the task of detecting a signal of interest buried in high dimensional noise. A matched subspace detector (MSD) is commonly used to solve this problem when the target signal is assumed to lie in a low-rank subspace. The low-rank signal buried in noise model is ubiquitous in signal processing. In array processing, [3] and [4] use multiple array snapshots to detect a low-rank signal in the presence of both interference and noise when the noise power is known and unknown, respectively. Similarly in adaptive radar detection, [5] and [6] adaptively detect distributed low-rank targets given multiple snapshots of primary (signal plus noise) and secondary (noise only) data under partially homogeneous and homogeneous noise assumptions, respectively. Low rank signal models are also used in electroencephalography (EEG) and magnetoencephalography (MEG) source localization as in [7]

Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

N. Asendorf is with the Department of Electrical Engineering and Computer Science, 1301 Beal Avenue, Room 4313, University of Michigan, Ann Arbor, MI, 48109 USA (e-mail: asendorf@umich.edu).

R.R. Nadakuditi is with the Department of Electrical Engineering and Computer Science, 1301 Beal Avenue, Room 4118, University of Michigan, Ann Arbor, MI, 48109 USA (e-mail: rajnrao@umich.edu).

Manuscript received ??

and [8], respectively. In [3], [4], [5], [6], the signal subspace is known. The performance of a MSD when the signal subspace is known was studied in [9] and [10] under deterministic signal assumptions and in [11] and [12] under stochastic signal assumptions. This paper considers the performance of a MSD in the less studied setting where the signal subspace is unknown and must be estimated from finite, noisy, signal-bearing training data.

The setting we have in mind arises from machine learning related applications where the low-rank signal model is reasonable but the signal subspace is not parameterizable. This is in contrast to the array processing applications that motivated the original MSD work [9] where the signal subspace is explicitly parameterizable whenever the array geometry is known. The inferential problem is made tractable by the availability of a training dataset consisting of signal-bearing observations that have been collected in a variety of representative experimental (and thus noisy) conditions. In such a scenario, the truncated eigen-decomposition of the sample covariance matrix of this training data yields an estimate of the unknown low-rank signal subspace, which may then be used for signal versus noise discrimination.

An illustrating example of this is the classical problem of handwriting recognition [13, Chapter 10] where a MSD can be used to determine if an area of an image contains a digit 0–9 or is pure noise. Here, a database [14], containing a large number of handwritten samples of each of the digits written by many different writers, is used to form a low-rank subspace estimate of each digit. The samples are noisy because of digitization effects and the inherent variation between writers. A nearest-subspace classifier based on retaining only the first few (10–12, in this example) principal components (or leading eigenvectors of the digit’s training data sample covariance matrix) associated with each digit yields greater than 93% classification performance [13, Table 10.1, pp. 121], indicating that the low-rank signal buried in noise model is appropriate. The motivating setting described also arises in the context of image or waveform recognition applications (e.g. license plate character recognition) where the target and the camera are separated by a dynamic random medium and in hyperspectral imaging based anomaly detection [15], [16], [17] relative to a statistically stationary scene (e.g. toxic gas detection). Here too, a practitioner might have access to training samples collected over a variety of experimental conditions and might employ the MSD in a similar manner.

In these applications, the standard plug-in detector, which

substitutes an estimate of the signal subspace into the expression for the oracle MSD that was derived assuming the subspace is perfectly known, realizes a performance loss because additive noise and finite training data decrease the accuracy of the estimated subspace. This motivates questions such as: What is the expected plug-in detector performance? Is it possible to avoid some of this performance loss? How does the estimation of the signal subspace dimension influence detector performance? Is the “play-it-safe” overestimation of subspace dimension, to compensate for the potential underestimation of schemes discussed in [18], a good idea?

Our performance analysis, which relies on insights from random matrix theory (RMT), highlights the importance of using no more than k_{eff} *informative* signal subspace components, where k_{eff} is a number that depends on the system dimensionality, number of training samples, and eigen-SNR (signal-to-noise-ratio). We derive a new RMT detector that only utilizes the k_{eff} *informative* signal subspace components, thereby avoiding some of the possible performance loss suffered by the plug-in detector. Given the number and quality (i.e. SNR) of the training samples, our analysis also allows a practitioner to predict the expected receiver operating characteristic (ROC) performance of a general class of detectors. An outcome of this analysis is that we can accurately predict how many training samples are needed to get to within ϵ of the oracle MSD’s performance (see Figures 2, 6(a), and 6(b)). This performance characterization can provide the practitioner with experimental guidance and might be a starting point for the formulation of achievable system performance specifications.

This paper differs from previous works in several aspects. The focus and main contribution is analytically quantifying the performance of a general class of MSD’s as a function of the system dimensionality, number of training samples, and eigen-SNR. Theorem 5.1 and Corollary 5.1 extend recent results from RMT [19], [20], [21] to precisely quantify the accuracy of the subspace estimate. This quantification yields approximations that appear to hold for moderate system dimensions even though the theory is asymptotic, in the limit of large dimensionality and relatively large training sample size. We provide a first-principles derivation of a new RMT detector that incorporates this knowledge of the accuracy of the estimated subspace, thereby illuminating the asymptotic form of a detector that mitigates some of the potential performance loss suffered by the plug-in detector. These RMT insights also allow us to characterize the ROC performance of a MSD under both a deterministic and stochastic model for the test vector. This work builds on [22] by providing the proofs of Theorem 5.1 and Corollary 5.1, analyzing the performance of the general class of detectors given in (14), considering the deterministic test vector setting, and unifying the performance analysis of the stochastic and deterministic MSD’s.

The paper is organized as follows. We describe the generative models for the training data and test vector and also estimate unknown parameters in Section II. In Section III, we derive standard oracle and plug-in detectors for each testing setting and highlight how finite training data causes subspace estimation errors and subsequent performance loss. We formally pose the questions addressed herein in Section IV.

Section V contains pertinent results from RMT and our definition in (16) of k_{eff} . In Section VI we derive RMT detectors for the stochastic and deterministic test vector models. Aided by RMT and a saddlepoint approximation of the CDF of a weighted sum of chi-square random variables, we predict ROC performance curves for a general detector in Section VII. We validate our asymptotic ROC predictions and demonstrate the importance of using the k_{eff} informative subspace components in Section VIII. We provide concluding remarks in Section IX.

II. DATA MODELS AND PARAMETER ESTIMATION

Given an observation, we wish to discriminate between the H_0 hypothesis that the observation is purely noise and the H_1 hypothesis that the observation contains a target signal. We assume that the signal of interest lies in a low dimensional subspace as in [3], [4], [5], [6], [7], [8], [15], [16], [17]. However, this low-rank subspace and the SNR governing the subspace components are unknown. To design a detector to distinguish between the H_0 and H_1 hypotheses, we have access to a training dataset, recorded under similar noisy conditions, whose observations are known to contain the signal of interest (see, for example, [16], [17]). We use this training data to form estimates of the unknown low-rank subspace and each component’s SNR. This section will mathematically describe the training data models, how we estimate any unknown parameters, and a stochastic and deterministic model for the testing data. Both testing models share the same training data model.

A. Training Data Model

We model our unknown subspace with the complex matrix $U = [u_1, \dots, u_k]$ such that $\dim u_i = n$ and $\langle u_i, u_j \rangle = u_i^H u_j = \delta_{ij}$ for $i, j = 1, \dots, k$. Here δ_{ij} is the delta function such that $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ii} = 1$ for $i = j$. We are given m signal-bearing training vectors $y_i \in \mathbb{C}^{n \times 1}$, $i = 1, \dots, m$, modeled¹ as $y_i = Ux_i + z_i$ where $z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_n)$ and $x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \Sigma)$ where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_k^2)$ with $\sigma_1 > \sigma_2 > \dots > \sigma_k > 0$ unknown. Similar gaussian priors appear in [5], [6], [15], [16]. Σ models the SNR of each subspace component and z_i models the additive noise. For each observation, x_i and z_i are independent. The dimension, k , of our subspace is unknown and we assume throughout that $k \ll n$ so that we have a low-rank signal embedded in a high-dimensional observation vector.

B. Parameter Estimation

The parameters k , U , and Σ are all unknown in our training model. For the rest of the paper, we assume that we are given a dimension estimate, \hat{k} ; this may have been estimated from the training data or provided by a domain expert. Typically, \hat{k} is an overestimation of a dimension estimate provided by percent variance, scree plots [23], or robust techniques [24], [25],

¹For expositional simplicity, we have assumed that all our matrices and vectors are complex-valued; our results also hold for real-valued matrices and vectors.

[26]. This overestimation, or “play-it-safe” strategy, strives to include all signal subspace components at the expense of possibly including non-signal subspace components.

Given \hat{k} and the signal bearing training data $Y = [y_1 \dots y_m]$, we form the sample covariance matrix $S = \frac{1}{m}YY^H$. The covariance matrix of y_i is $U\Sigma U^H + I_n$ and it follows that the (classical) ML estimates (in the many-sample, small matrix setting) for U and Σ are given by [27]

$$\begin{aligned}\hat{U} &= [\hat{u}_1 \dots \hat{u}_{\hat{k}}] \\ \hat{\sigma}_i^2 &= \max(0, \hat{\lambda}_i - 1) \text{ for } i = 1, \dots, \hat{k}\end{aligned}\quad (1)$$

where $\hat{\lambda}_1, \dots, \hat{\lambda}_{\hat{k}}$ are the \hat{k} largest eigenvalues of the sample covariance matrix, S , and $\hat{u}_1, \dots, \hat{u}_{\hat{k}}$ are the corresponding eigenvectors. Define the signal covariance matrix estimate as $\hat{\Sigma} = \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_{\hat{k}}^2)$. We are now able to use the parameter estimates \hat{U} and $\hat{\Sigma}$ in detectors where necessary.

C. Testing Data Model

We will consider both a stochastic and deterministic model for a test vector. In both settings, parameter estimates are formed as described in (1) from training data modeled in Section II-A.

In the stochastic setting, the test vector $y \in \mathbb{C}^{n \times 1}$ is modeled as

$$\text{Stochastic Model: } y = \begin{cases} z & y \in H_0 : \text{Noise only} \\ Ux + z & y \in H_1 : \text{Signal-plus noise} \end{cases}, \quad (2)$$

where U , z , and x are modeled as described in Section II-A. This assumes that the signal, Ux , may lie anywhere in the subspace and whose position in the subspace is governed by the signal covariance matrix Σ .

In the deterministic setting, the test vector $y \in \mathbb{C}^{n \times 1}$ is modeled as

$$\text{Deterministic Model: } y = \begin{cases} z & y \in H_0 : \text{Noise only} \\ U\Sigma^{1/2}x + z & y \in H_1 : \text{Signal-plus noise} \end{cases}, \quad (3)$$

where U , Σ , and z are modeled as described in Section II-A. Here, in contrast to the stochastic setting, x is a non-random deterministic vector. Thus the signal, $U\Sigma^{1/2}x$, lies at a fixed point in the unknown subspace. Note that Σ still controls the SNR of each subspace component and that placing a mean zero, identity covariance Gaussian prior on x in (3) yields the stochastic model described in (2).

III. STANDARD DETECTOR DERIVATIONS

In this paper, we focus on the Neyman-Pearson setting (see [28]) where, given a test observation from (2) or (3), a MSD is a likelihood ratio test (LRT) taking the form

$$\Lambda(y) := \frac{f(y|H_1)}{f(y|H_0)} \stackrel{H_1}{\gtrless} \eta \quad (4)$$

where $\Lambda(y)$ is the test statistic, η is the threshold set to achieve a given false alarm rate, and f is the appropriate conditional density of the test observation. In the following section, for both testing data models we derive the standard oracle detector (assuming all parameters are known) and plug-in detector

(formed by substituting the parameter estimates of (1) in the oracle detector). The oracle detectors, while unrealizable, give an upper bound for the performance of a MSD. We will see that when only finite training data is available (as is the case in real applications), the plug-in detector will realize a performance loss relative to this bound.

A. Stochastic Testing Model

The LRT in (4) depends on the conditional distribution of the test vector, y . By properties of Gaussian random variables, when using the stochastic test model in (2), these distributions are $y|H_0 \sim \mathcal{N}(0, I_n)$ and $y|H_1 \sim \mathcal{N}(0, U\Sigma U^H + I_n)$. The resulting LRT statistic is

$$\Lambda(y) = \frac{\mathcal{N}(0, U\Sigma U^H + I_n)}{\mathcal{N}(0, I_n)}. \quad (5)$$

We derive an oracle detector by assuming that k , Σ , and U are all known in (5). After simplification of this expression (see Section 4.14 of [1]), the oracle statistic becomes

$$\Lambda_{\text{oracle}}(y) = y^H U (\Sigma^{-1} + I_k)^{-1} U^H y. \quad (6)$$

Note that the oracle statistic depends on the sufficient statistic $w := U^H y$. Using this notation, the oracle statistic is

$$\boxed{\Lambda_{\text{oracle}}(w) = w^H (\Sigma^{-1} + I_k)^{-1} w = \sum_{i=1}^k \left(\frac{\sigma_i^2}{\sigma_i^2 + 1} \right) w_i^2} \quad (7)$$

and the oracle detector is $\Lambda_{\text{oracle}}(w) \stackrel{H_1}{\gtrless} \gamma_{\text{oracle}}$ where the threshold γ_{oracle} is chosen in the usual manner, *i.e.*, so that it satisfies $P(\Lambda_{\text{oracle}}(w) > \gamma_{\text{oracle}} | H_0) = \alpha$ with α a desired false alarm rate.

However, as the parameters U and Σ are unknown, the oracle statistic in (7) cannot be computed. Given a dimension estimate \hat{k} , we substitute the ML estimates of U and Σ given in (1) for the unknown parameters in (6) as similarly done in [11] and [12]. This results in the plug-in detector’s LRT statistic: $\Lambda_{\text{plugin}}(y) = y^H \hat{U} (\hat{\Sigma}^{-1} + I_{\hat{k}})^{-1} \hat{U}^H y$. Simplifying this expression using the statistic $\hat{w} = \hat{U}^H y$, yields the plug-in statistic

$$\boxed{\Lambda_{\text{plugin}}(\hat{w}) = \hat{w}^H \text{diag} \left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + 1} \right) \hat{w} = \sum_{i=1}^{\hat{k}} \left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + 1} \right) \hat{w}_i^2} \quad (8)$$

and the plug-in detector takes the form $\Lambda_{\text{plugin}}(w) \stackrel{H_1}{\gtrless} \gamma_{\text{plugin}}$ where the threshold γ_{plugin} is chosen in the usual manner.

The plug-in detector assumes that the estimated signal subspace, \hat{U} , is equal to the true signal subspace, U , and that the estimated signal covariance, $\hat{\Sigma}$, is equal to the true signal covariance, Σ . In other words, the plug-in detector derivation assumes that $\hat{U}^H U = I_{\hat{k}}$, $\hat{\sigma}_i^2 = \sigma_i^2$ for $i = 1, \dots, \hat{k}$, and the provided subspace dimension estimate, \hat{k} , is equal to the true underlying dimension of our signal subspace, k . Perhaps unsurprisingly, (as discussed in Section V) incorrectly choosing \hat{k} degrades the performance of the plug-in detector.

B. Deterministic Testing Model

We now consider the alternative deterministic test vector model (3), which results in the following conditional distributions of the test vector $y|H_0 \sim \mathcal{N}(0, I_n)$ and $y|H_1 \sim \mathcal{N}(U\Sigma^{1/2}x, I_n)$. We begin by deriving an oracle detector, which assumes that U , Σ , x , and k are all known. The LRT statistic for such a scenario is $\Lambda(y) = \frac{\mathcal{N}(U\Sigma^{1/2}x, I_n)}{\mathcal{N}(0, I_n)}$. Simplifying this expression leads to the oracle statistic

$$\Lambda_{\text{oracle}}(y) = x^H \Sigma^{1/2} U^H y. \quad (9)$$

As in the stochastic setting, $w = U^H y$ is a sufficient statistic and the oracle statistic simplifies to

$$\boxed{\Lambda_{\text{oracle}}(w) = x^H \Sigma^{1/2} w = \sum_{i=1}^k x_i \sigma_i w_i.} \quad (10)$$

However, as the parameters U , Σ , and x are unknown, the oracle statistic in (10) cannot be computed. Since we must estimate x from the test vector, we employ the generalized likelihood ratio test (GLRT) where $\Lambda(y) = \frac{\max_x f(y|H_1)}{f(y|H_0)}$, resulting in the GLRT statistic

$$\Lambda(y) = \frac{\max_x \mathcal{N}(U\Sigma^{1/2}x, I_n)}{\mathcal{N}(0, I_n)}. \quad (11)$$

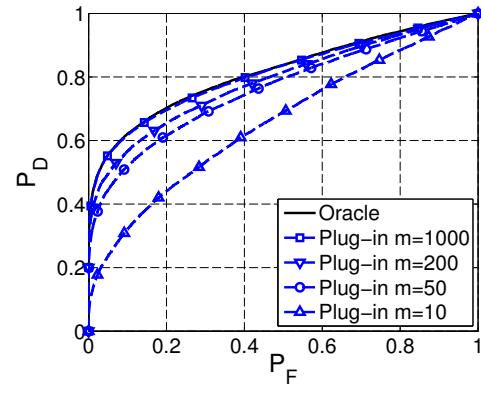
Employing maximum likelihood estimation on x in (11) yields the estimate $\hat{x} = \Sigma^{-1/2} U^H y$. Proceeding as in the stochastic setting, we substitute \hat{x} for the unknown x in (9) and then substitute the ML estimates of U and Σ given in (1) for the unknown U and Σ (see Section 4.11 of [1] for a similar treatment). This results in the plug-in statistic $\Lambda_{\text{plugin}}(y) = y^H \hat{U} \hat{U}^H y$. Again, $\hat{w} = \hat{U}^H y$ is a statistic that can be used to write the plug-in statistic as

$$\boxed{\Lambda_{\text{plugin}}(\hat{w}) = \hat{w}^H \hat{w} = \sum_{i=1}^{\hat{k}} \hat{w}_i^2,} \quad (12)$$

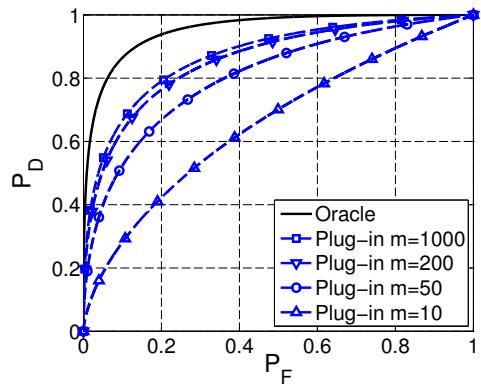
resulting in the detector $\Lambda_{\text{plugin}}(\hat{w}) \stackrel{H_1}{\gtrless} \stackrel{H_0}{\lesssim} \gamma_{\text{plugin}}$, where the threshold γ_{plugin} is chosen in the usual manner. The deterministic plug-in detector is an ‘energy detector’, which sums the energy of the test observation lying in the subspace \hat{U} .

C. Effect of the Number of Training Samples

In both the stochastic and deterministic testing settings, $\hat{w} = \hat{U}^H y$ is a statistic used in the plug-in statistics (8) and (12). This statistic relies on the estimated subspace \hat{U} formed from the top \hat{k} eigenvectors of the sample covariance matrix, S , of the training data. The stochastic detector also relies on the subspace-SNR estimate $\hat{\Sigma}$ formed from the top \hat{k} eigenvalues of S . For a fixed Σ , the accuracy of these estimates depends on the number of training data samples, m ; we will mathematically show this in Section V. If we had access to an infinite amount of training data, the parameter estimates would be exact ($\hat{U} \rightarrow U$ and $\hat{\Sigma} \rightarrow \Sigma$). However, when we have access to only a finite amount of training data, \hat{U} and $\hat{\Sigma}$ are inaccurate and will degrade the performance of the plug-in



(a) Stochastic Setting



(b) Deterministic Setting

Fig. 1. Empirical ROC curves for the plug-in and oracle detectors. Empirical ROC curves were simulated with $n = 200$, $\hat{k} = k = 2$, and $\Sigma = \text{diag}(10, 0.1)$. The empirical ROC curves were computed using 10000 test samples and averaged over 100 trials using algorithms 2 and 4 of [29]. (a) Shows results for the stochastic MSD. (b) Shows results for the deterministic MSD when $x = [0.75, 0.75]^T$. For both settings, as m decreases, the performance of the plug-in detector degrades.

detectors with respect to the oracle detector, which provides an upper bound on detector performance.

To illustrate this performance loss, we consider a moderately sized system where $n = 200$ and $\Sigma = \text{diag}(10, 0.1)$. We consider five detectors: the oracle detector and four plug-in detectors each using parameter estimates formed from varying amounts of training data. Figures 1(a) and 1(b) plot the empirical ROC curves for the stochastic and deterministic testing settings, respectively. The amount of training data drastically affects the performance of the plug-in detector. As m decreases, the plug-in detectors realize a significant performance loss. However, as $m \rightarrow \infty$, the plug-in detectors realize improved performance, closer to that of the oracle detectors.

For the stochastic detector, as $m \rightarrow \infty$, the plug-in detector achieves the same performance as the oracle detector. Examination of the statistics (7) and (8) shows that these statistics will be identical when $\hat{U} \rightarrow U$ and $\hat{\Sigma} \rightarrow \Sigma$, which is the case when infinite training data is available. However, this is not the case for the deterministic plug-in detector. Even with an infinite amount of training data, the plug-in detector will not achieve the oracle detector’s performance. The deterministic plug-in detector must estimate x given a noisy test observation y , which is independent from the training data. Even with infinite training data causing $\hat{U} \rightarrow U$ and $\hat{\Sigma} \rightarrow \Sigma$, \hat{x} does not

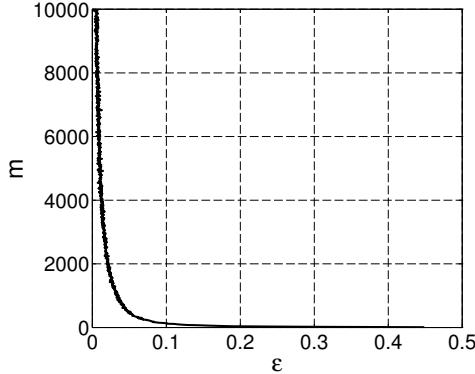


Fig. 2. Empirically determined number of training samples, m , needed for the stochastic plug-in detector to achieve a desired performance loss, ϵ , as defined in (13). The required false alarm rate is $P_F = 0.1$. Empirical ROC curves were generated for $n = 200$, $\Sigma = \text{diag}(10, 0.1)$, $\hat{k} = k = 2$ using 10000 testing samples and averaged over 100 trials using algorithms 2 and 4 of [29].

converge to x . Therefore, the deterministic plug-in detector cannot achieve the performance bound of the oracle detector, which assumes that x is known.

For a fixed probability of false alarm (P_F), we can explore this performance loss by comparing the achieved probability of detection (P_D) of the plug-in detector to that of the oracle detector. Let

$$\epsilon = 1 - \frac{P_D^{\text{plugin}}}{P_D^{\text{oracle}}} \quad (13)$$

be the performance loss of the plug-in detector. Figure 2 empirically plots the number of training samples needed to achieve a desired performance loss ϵ for the stochastic plug-in detector. There is an exponential relationship between ϵ and m indicating that we need infinite training samples to achieve zero performance loss ($\epsilon = 0$). However, in any practical application we will never have an infinite amount of training data and so the plug-in detector will realize some non-zero performance loss. The rest of the paper will mathematically predict how finite training data affects detector performance and will derive new detectors to avoid some of this performance loss.

IV. PROBLEM STATEMENTS

We saw in Section III that the plug-in detectors rely on the statistic $\hat{w} = \hat{U}^H y$. When only finite training data is available, the subspace estimate \hat{U} is inaccurate and subsequently degrades the performance of the plug-in detector. Motivated by this observation, we formulate the problems addressed in this paper.

A. Problem 1: Derive a New Detector that Exploits Predictions of Subspace Accuracy

We know that subspace estimation errors degrade the performance of the plug-in detector. Recent results from RMT specifically quantify the accuracy of \hat{U} relative to U . By deriving a new detector that accounts for this accuracy of the estimated subspace, we hope to avoid some of the performance loss associated with the plug-in detector. For both the stochastic and deterministic testing settings our goal is to

Design a new detector that exploits RMT predictions of subspace estimation accuracy.

The detector derivations in Section VI will provide insights on when, if, and how the performance of plug-in detectors that do not exploit the knowledge of subspace estimation accuracy can be improved.

B. Problem 2: Characterize ROC Performance Curves

We saw in Section III that both plug-in detectors took the form

$$\hat{w}^H D \hat{w} \stackrel{H_1}{\gtrless} \stackrel{H_0}{\eta} \quad (14)$$

where D is the appropriate diagonal matrix and the test statistic $\Lambda(\hat{w}) = \hat{w}^H D \hat{w}$ is compared against a threshold, η , set to achieve a prescribed false alarm rate. After solving Problem 1, we will see that the RMT detectors derived in Section VI also take the form of (14). In order to compare detectors of this form without training data or empirically generated test samples, we wish to analytically predict their ROC performance. Formally, for detectors with the form of (14) and for test vectors modeled as (2) or (3), our goal is to

$$\begin{aligned} \text{Predict } P_D &:= \mathbb{P}(\text{Detection}), \text{ for every} \\ P_F &:= \mathbb{P}(\text{False Alarm}) = \alpha \in (0, 1) \text{ given } n, m, \hat{k}, D \text{ and } \Sigma. \end{aligned}$$

For this problem, we assume that we are given Σ . We derive this theoretical prediction of ROC performance curves in Section VII and show that this performance prediction also relies on RMT results quantifying the accuracy of the subspace estimate \hat{U} , specifically the entries of the matrix $\hat{U}^H U$. In Section V we provide an asymptotic diagonal approximation to this matrix that makes the ROC prediction possible.

V. PERTINENT RESULTS FROM RANDOM MATRIX THEORY

In Section II-B we formed estimates \hat{U} and $\hat{\Sigma}$ of the unknown U and Σ by taking the eigen-decomposition of the sample covariance matrix S of the training data matrix Y . These estimates are inaccurate because the training data is noisy and contains only a finite number of observations. The following analysis specifically quantifies the accuracy of these estimates and is necessary to derive a new detector and predict ROC performance curves of detectors with the form of (14).

A. Eigenvector Aspects

The subspace estimate \hat{U} is formed from the eigenvectors corresponding to the \hat{k} largest eigenvalues of S . For an arbitrary non-random diagonal matrix D , we will be particularly interested in the matrix $\hat{U}^H U D U^H \hat{U}$ that appears in detector derivations and the ROC performance analysis in Sections VI and VII. The following proposition characterizes the limiting behavior (up to an arbitrary phase) of the diagonal entries of the matrix $\hat{U}^H U$.

Proposition 5.1: Assume that the columns of the training data matrix Y were generated as described in Section II-A. Let \hat{u}_i denote the eigenvector associated with the i -th largest

eigenvalue of S . Then for $i = 1, \dots, k$ and $n, m \rightarrow \infty$ with $n/m \rightarrow c$, we have that

$$|\langle u_i, \hat{u}_i \rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \frac{\sigma_i^4 - c}{\sigma_i^4 + \sigma_i^2 c} & \text{if } \sigma_i^2 > \sqrt{c} \\ 0 & \text{otherwise} \end{cases}. \quad (15)$$

Proof: This follows from Theorem 4 of [19] when $\gamma = c$, $\ell_\nu - 1 = \sigma_\nu^2$, $\tilde{e}_\nu = u_\nu$, and $p_\nu = \hat{u}_\nu$. This result also appears in Theorem 2.2 of [20]. ■

We note that $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence. The key insight from Proposition 5.1 is that only the eigenvectors corresponding to the signal variances, σ_i^2 , lying above the phase transition \sqrt{c} are *informative*. When a signal variance drops below this critical threshold, the corresponding eigenvector estimate is essentially noise-like (i.e. $|\langle u_i, \hat{u}_i \rangle|^2 = o_p(1)$ meaning $|\langle u_i, \hat{u}_i \rangle|^2 \xrightarrow{P} 0$ as $n \rightarrow \infty$, denoting convergence in probability) and thus *uninformative*. Decreasing the amount of training data, m , increases c , thereby decreasing the value of $|\langle u_i, \hat{u}_i \rangle|^2$; if this quantity became 0, the associated subspace component would become uninformative.

The term $|\langle u_i, \hat{u}_i \rangle|^2$ quantifies mismatch between the estimated and underlying eigenvectors and will play an important role in deriving a new RMT detector and in characterizing detector performance; a similar term also appears in the analysis of the resolving power of arrays due to model mismatch such as in [30].

Following [18], we define the effective number of (asymptotically) identifiable subspace components k_{eff} as:

$$k_{\text{eff}} = \text{Number of } \sigma_i^2 > \sqrt{c}. \quad (16)$$

We can form an estimate of k_{eff} , \hat{k}_{eff} , using ‘Algorithm 2’ of [24]. This algorithm assumes the same model of a low-rank signal buried in high dimensional noise as our training data. Given a desired significance level, the algorithm estimates the number of signals present in a finite number of samples. When the noise covariance matrix is not known a priori, we would instead use ‘Algorithm 1’ of [24]. Both algorithms rely on the Tracy-Widom distribution. Note that $\hat{k}_{\text{eff}} \leq k$ but that we allow $\hat{k} \geq \hat{k}_{\text{eff}}$ so we may understand the impact of a play-it-safe overestimation of the signal subspace dimension estimate \hat{k}_{eff} returned using RMT based detectors [24], [25], [26].

Proposition 5.1 only characterizes the limiting behavior (up to an arbitrary phase) of the diagonal entries of the matrix $\hat{U}^H U$. We now state a new theorem characterizing the limiting behavior of the off-diagonal entries in $\hat{U}^H U$.

Theorem 5.1: Assume the same hypothesis as in Proposition 5.1. Let $\hat{k} = k_{\text{eff}} = k$. For $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, and $i \neq j$, as $n, m \rightarrow \infty$ with $n/m \rightarrow c$, $\langle u_j, \hat{u}_i \rangle \xrightarrow{\text{a.s.}} 0$.

Proof: This is a new result. See Appendix for proof. ■

Claim 5.1: Assume the same hypothesis as in Proposition 5.1. For $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, and $i \neq j$, as $n, m \rightarrow \infty$ with $n/m \rightarrow c$, $\langle u_j, \hat{u}_i \rangle \xrightarrow{\text{a.s.}} 0$.

Remark 5.1: See Appendix for a brief discussion of this claim.

Together, Proposition 5.1 and Claim 5.1 characterize the limiting behavior of the entries of $\hat{U}^H U$. This permits ap-

proximation, in the large matrix limit, of $\hat{U}^H U D U^H \hat{U}$ by a suitable diagonal matrix.

Corollary 5.1: Suppose $\hat{k} \leq k$ and let D be a $k \times k$ (non-random) diagonal matrix such that $D = \text{diag}(d_1, \dots, d_k)$, independent of \hat{U} . Then as $n, m \rightarrow \infty$ with $n/m \rightarrow c$, we have that

$$\hat{U}^H U D U^H \hat{U} \xrightarrow{\text{a.s.}} \text{diag}(d_1 |\langle u_1, \hat{u}_1 \rangle|^2, \dots, d_{\hat{k}} |\langle u_{\hat{k}}, \hat{u}_{\hat{k}} \rangle|^2)$$

where for $i = 1, \dots, \hat{k}$ the quantity $|\langle u_i, \hat{u}_i \rangle|^2$ is given in Proposition 5.1.

Proof: This follows directly by applying Proposition 5.1 and Claim 5.1 to the entries of the matrix $U^H \hat{U}$. ■

This diagonal approximation of $\hat{U}^H U D U^H \hat{U}$ will be used in detector derivations and ROC performance analyses in Sections VI and VII.

B. Eigenvalue Aspects

The signal covariance estimate $\hat{\Sigma}$ is formed from the largest \hat{k} eigenvalues of S . To characterize the ROC performance curves of plug-in detectors that use $\hat{\Sigma}$ as the signal covariance estimate, we will also need to characterize the limiting behavior of $\hat{\Sigma}$. The following proposition gives the limiting behavior of these signal variance estimates.

Proposition 5.2: As $n, m \rightarrow \infty$ with $n/m \rightarrow c$ we have that:

$$\hat{\sigma}_i^2 \xrightarrow{\text{a.s.}} \begin{cases} \sigma_i^2 + c + \frac{c}{\sigma_i^2} & \text{if } \sigma_i^2 > \sqrt{c} \\ c + 2\sqrt{c} & \text{if } \sigma_i^2 \leq \sqrt{c} \end{cases}.$$

Proof: This follows from Theorems 1 and 2 in [19] for the real setting for $c < 1$ when $\gamma = c$, $\ell_\nu - 1 = \sigma_\nu^2$, and $\hat{\ell}_\nu - 1 = \hat{\sigma}_\nu^2$. See Theorem 2.6 in [21] for the complete result. ■

These limiting values will be used in Section VII when deriving the ROC performance of the plug-in detectors.

When only finite training data is available, c is non-zero and Proposition 5.2 shows that $\hat{\sigma}_i^2$ is biased. We wish to derive an improved signal variance estimate to use in a new RMT detector and to estimate $|\langle u_i, \hat{u}_i \rangle|^2$ in (15). As seen in Proposition 5.1, when $\sigma_i^2 \leq \sqrt{c}$ the eigenvector estimate is uninformative and we would not want to include that subspace component in a detector; the associated signal variance estimate is therefore unnecessary. For the \hat{k}_{eff} subspace components that are informative (i.e. when $\sigma_i^2 > \sqrt{c}$) we form an improved signal variance estimate using the following proposition that characterizes the fluctuations of these signal variance estimates.

Proposition 5.3: As $n, m \rightarrow \infty$ with $n/m \rightarrow c$, we have that for $i = 1, \dots, k_{\text{eff}}$

$$\sqrt{n} \left(\hat{\sigma}_i^2 - \left(\sigma_i^2 + c + \frac{c}{\sigma_i^2} \right) \right) \Rightarrow \mathcal{N} \left(0, \frac{2(\sigma_i^2 + 1)^2}{\beta} \left(1 - \frac{c}{\sigma_i^4} \right) \right),$$

where $\beta = 1$ when the data is real-valued and $\beta = 2$ when the data is complex-valued.

Proof: This follows from Theorem 3 in [19] for the real setting for $c < 1$ when $\gamma = c$, $\ell_\nu - 1 = \sigma_\nu^2$, $\hat{\ell}_\nu - 1 = \hat{\sigma}_\nu^2$, and p_ν is the limit of Theorem 2 of [19]. See Theorem 2.15 in [21] for the complete result. ■

For the \hat{k}_{eff} informative subspace components we form an improved estimate, $\hat{\sigma}_{i_{\text{rmt}}}^2$, of the unknown signal variance, σ_i^2 , by employing maximum-likelihood (ML) estimation on the distribution in Proposition 5.3. Specifically, for only the \hat{k}_{eff} signal eigenvalues, we form the RMT estimate:

$$\hat{\sigma}_{i_{\text{rmt}}}^2 = \underset{\sigma_i^2}{\operatorname{argmax}} \log \left(f_{\hat{\sigma}_i^2}(\sigma_i^2) \right) \quad (17)$$

where

$$f_{\hat{\sigma}_i^2}(\sigma_i^2) := \mathcal{N} \left(\left(\sigma_i^2 + c + \frac{c}{\sigma_i^2} \right), \frac{2(\sigma_i^2 + 1)^2}{n\beta} \left(1 - \frac{c}{\sigma_i^4} \right) \right).$$

We may then estimate $|\langle u_i, \hat{u}_i \rangle|^2$ in (15) by substituting the improved signal variance estimates, $\hat{\sigma}_{i_{\text{rmt}}}^2$, for the unknown σ_i^2 in Proposition 5.1. We refer to this estimate as $|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2$. For the $\hat{k} - \hat{k}_{\text{eff}}$ uninformative subspace components, we set $|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2 = 0$.

VI. DERIVATION OF NEW RMT MATCHED SUBSPACE DETECTORS

We saw in Section III that the plug-in detectors rely on the statistic $\hat{w} = \hat{U}^H y$. Instead of deriving the LRT statistic using the conditional distributions of y , we will instead use the conditional distributions of \hat{w} ; this will reveal the importance of the matrix $\hat{U}^H U$. The plug-in detectors assume that $\hat{U}^H U = I_{\hat{k}}$, however, the analysis in Section V-A shows that this assumption is incorrect. Knowing the importance of only using k_{eff} subspace components and armed with the asymptotic diagonal approximation of Corollary 5.1 and the improved signal variance estimates in (17), we are now in position to answer Problem 1 and derive a new RMT detector for both testing settings.

A. Stochastic RMT Detector

We begin with the stochastic test setting and form the test vector $\hat{w} = \hat{U}^H y$ where \hat{U} is the subspace estimated from (1) and y is generated from (2). The LRT statistic using \hat{w} depends on the conditional distributions under each hypothesis, which by properties of Gaussian random variables are simply

$$\hat{w}|H_0 \sim \mathcal{N}(0, I_{\hat{k}}) \quad \text{and} \quad \hat{w}|H_1 \sim \mathcal{N}\left(0, \hat{U}^H U \Sigma U^H \hat{U} + I_{\hat{k}}\right). \quad (18)$$

We immediately see the matrix of interest, $\hat{U}^H U \Sigma U^H \hat{U}$. The plug-in detector substitutes \hat{U} for U and $\hat{\Sigma}$ for Σ ; this results in $\hat{w}|H_1 \sim \mathcal{N}(0, \hat{\Sigma} + I_{\hat{k}})$. However, Corollary 5.1 shows that this is incorrect by providing the asymptotic limit of the covariance matrix in (18):

$$\hat{U}^H U \Sigma U^H \hat{U} + I_{\hat{k}} \xrightarrow{\text{a.s.}} \text{diag}(|\langle u_i, \hat{u}_i \rangle|^2 \sigma_i^2 + 1). \quad (19)$$

If σ_i^2 were assumed known, this limit would suffice because we could plug in the results in Proposition 5.1 to get the desired statistic. However, the signal variances are unknown so σ_i^2 and subsequently $|\langle u_i, \hat{u}_i \rangle|^2$ must be estimated from data. For the \hat{k}_{eff} subspace components estimated from ‘Algorithm 2’ of [24], we form an improved signal variance estimate, $\hat{\sigma}_{i_{\text{rmt}}}^2$, obtained via (17) and use it to estimate $|\langle u_i, \hat{u}_i \rangle|^2$, denoted by $|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2$. Of course, there are correction terms due to

finite system size effects, which we ignore, that affect the convergence properties but not the asymptotic form of the detector.

We obtain the RMT detector by computing the LRT statistic using the conditional distributions of (18). The covariance matrix of $\hat{w}|H_1$ is computed by substituting $|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2$ and $\hat{\sigma}_{i_{\text{rmt}}}^2$ into the diagonal covariance matrix (19). After some straightforward algebra we obtain the desired RMT statistic

$$\Lambda_{\text{rmt}}(\hat{w}) = \sum_{i=1}^{\hat{k}} \left(\frac{|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2}{|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2 + 1} \right) \hat{w}_i^2.$$

As seen in Proposition 5.1, when $i > k_{\text{eff}}$, $|\langle u_i, \hat{u}_i \rangle|^2 \xrightarrow{\text{a.s.}} 0$. The sum on the right hand side (asymptotically) discards the uninformative subspace components. Thus the RMT detector only uses the \hat{k}_{eff} informative components given by (16). Consequently, we obtain the test statistic

$$\Lambda_{\text{rmt}}(\hat{w}) = \sum_{i=1}^{\hat{k}_{\text{eff}}} \left(\frac{|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2}{|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2 + 1} \right) \hat{w}_i^2 \quad (20)$$

and the RMT detector becomes $\Lambda_{\text{rmt}}(\hat{w}) \xrightarrow[H_0]{H_1} \gamma_{\text{rmt}}$ where the threshold γ_{rmt} is chosen in the usual manner. Note that the stochastic RMT detector also takes the form of (14). The principal difference between the RMT test statistic in (20) and the plug-in test statistic in (8) is the role of \hat{k}_{eff} in the former. The scaling factors associated with each \hat{w}_i^2 for both detectors are about the same; this is why the plug-in detector that uses \hat{k}_{eff} components exhibits the same (asymptotic) performance as the RMT detector, which incorporates knowledge of the subspace estimate accuracy. However, our analysis shows that overcompensating and “playing-it-safe” by setting $\hat{k} > \hat{k}_{\text{eff}}$ can lead to performance loss.

B. Deterministic RMT Detector

When forming \hat{w} with y generated from (3), the conditional distributions of \hat{w} under each hypothesis are $\hat{w}|H_0 \sim \mathcal{N}(0, I_{\hat{k}})$ and $\hat{w}|H_1 \sim \mathcal{N}(\hat{U}^H U \Sigma^{1/2} x, I_{\hat{k}})$. Again, as x is unknown, we use a GLRT. Employing maximum likelihood estimation on x yields the estimate $\hat{x} = (\Sigma^{1/2} U^H \hat{U} \hat{U}^H U \Sigma^{1/2})^\dagger \Sigma^{1/2} U^H \hat{U} \hat{w}$ where \dagger denotes the Moore-Penrose pseudoinverse. After simplifying using \hat{x} and using the natural logarithm operator as a monotonic operation, the GLRT statistic becomes

$$\Lambda(\hat{w}) = \hat{w}^H \left(\hat{U}^H U \Sigma^{1/2} \left(\Sigma^{1/2} U^H \hat{U} \hat{U}^H U \Sigma^{1/2} \right)^\dagger \Sigma^{1/2} U^H \hat{U} \right) \hat{w}.$$

Consider the term $\hat{U}^H U$. By Proposition 5.1 and Claim 5.1 and by noting that the eigenvectors are unique up to a phase, we have that $\hat{U}^H U \xrightarrow{\text{a.s.}} BA$ where B is a $\hat{k} \times k$ matrix and A is a $k \times k$ matrix defined as

$$B_{i\ell} := \begin{cases} b_i = \exp(j\psi_i) & i = \ell \\ 0 & \text{otherwise} \end{cases}, \quad A_{i\ell} := \begin{cases} a_i = |\langle u_i, \hat{u}_i \rangle| & i = \ell \\ 0 & \text{otherwise} \end{cases}.$$

For some ψ_i , b_i denotes the random phase ambiguity in the eigenvector computation (since eigenvectors are unique up to a phase).

Detector	Detector Statistic $\Lambda(\hat{w})$	Distribution of ΛH_0	Distribution of ΛH_1
Plug-in	$\sum_{i=1}^{\hat{k}} \left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + 1} \right) \hat{w}_i^2$	$\sum_{i=1}^{\hat{k}} \left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + 1} \right) \chi_{1i}^2$	$\sum_{i=1}^{\hat{k}} \left(\frac{\hat{\sigma}_i^2 (\sigma_i^2 \langle u_i, \hat{u}_i \rangle ^2 + 1)}{\hat{\sigma}_i^2 + 1} \right) \chi_{1i}^2$
RMT	$\sum_{i=1}^{\hat{k}_{\text{eff}}} \left(\frac{ \langle u_i, \hat{u}_i \rangle _{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2}{ \langle u_i, \hat{u}_i \rangle _{\text{rmt}}^2 \hat{\sigma}_{i_{\text{rmt}}}^2 + 1} \right) \hat{w}_i^2$	$\sum_{i=1}^{\hat{k}_{\text{eff}}} \left(\frac{\hat{\sigma}_{i_{\text{rmt}}}^2 \langle u_i, \hat{u}_i \rangle _{\text{rmt}}^2}{\hat{\sigma}_{i_{\text{rmt}}}^2 \langle u_i, \hat{u}_i \rangle _{\text{rmt}}^2 + 1} \right) \chi_{1i}^2$	$\sum_{i=1}^{\hat{k}_{\text{eff}}} \left(\hat{\sigma}_{i_{\text{rmt}}}^2 \langle u_i, \hat{u}_i \rangle _{\text{rmt}}^2 \right) \chi_{1i}^2$

TABLE I
SUMMARY OF THE PLUG-IN AND RMT STOCHASTIC MSDS. SEE SECTIONS III-A AND VI-A FOR DERIVATIONS.

The plug-in detector assumes that $A = B = I_{\hat{k}}$, that is $b_i = 1$ and $|\langle u_i, \hat{u}_i \rangle| = 1$. However, as seen in Section V, we have knowledge of $|\langle u_i, \hat{u}_i \rangle|$ which we may exploit in deriving a new detector. Using the notation just developed, the GLRT statistic may be written as

$$\Lambda(\hat{w}) = \hat{w}^H B A \Sigma^{1/2} (\Sigma^{1/2} A^H B^H B A \Sigma^{1/2})^\dagger \Sigma^{1/2} A^H B^H \hat{w}.$$

We use (17) and Proposition 5.1 to estimate $a_i = \sqrt{|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2}$. Recall that \hat{k}_{eff} is an estimate for the number of σ_i^2 above the phase transition and note that $a_i = 0$ when $\sigma_i^2 \leq \sqrt{c}$. Incorporating this into the detector, and noting that A , B , and Σ contain only diagonal elements, the GLRT simplifies to

$$\boxed{\Lambda_{\text{rmt}}(\hat{w}) = \sum_{i=1}^{\hat{k}_{\text{eff}}} \hat{w}_i^2} \quad (21)$$

and the deterministic RMT detector is $\Lambda_{\text{rmt}}(\hat{w}) \stackrel[H_1]{>}{<} \gamma_{\text{rmt}}$ where the threshold γ_{rmt} is chosen in the usual manner. This addresses the problem posed in Section IV-A for the deterministic test vector setting. We note that this deterministic RMT detector also takes on the form of (14). In fact, in the deterministic setting, the plug-in and RMT detectors are both ‘energy detectors’ and have the same statistic except for the upper bound in the summation. As in the stochastic setting, the principal difference between the RMT test statistic in (21) and the plug-in test statistic in (12) is the role of \hat{k}_{eff} in the former. This is also why the plug-in detector that uses \hat{k}_{eff} components exhibits the same performance as the RMT detector, which incorporates knowledge of the subspace estimates.

VII. THEORETICAL ROC CURVE PREDICTIONS

We saw in Sections III and VI that the plug-in and RMT detectors under both testing settings are (exactly or asymptotically) of the form given by (14). Thus by answering the ROC curve prediction problem posed in Section IV-B, we have characterized the asymptotic (or large system) performance of the detectors considered herein. For the following analysis, we are given n , m , \hat{k} , D , Σ , and x (in the deterministic setting).

We first note that each previously derived detector corresponds to a specific choice of the diagonal matrix D in (14), which can be discerned by inspection of Tables I and II. In what follows, we solve the ROC prediction problem for general D ; direct substitution of the relevant parameters for D will yield the performance curves for individual detectors.

Recall that the ROC curve [29] for a test statistic $\Lambda(\hat{w})$ is obtained by computing

$$P_D = P(\Lambda(\hat{w}) \geq \gamma | \hat{w} \in H_1), \quad P_F = P(\Lambda(\hat{w}) \geq \gamma | \hat{w} \in H_0) \quad (22)$$

for $-\infty < \gamma < \infty$ and plotting P_D versus P_F . To compute these expressions in (22) for the deterministic and stochastic test vector setting, we need to characterize the conditional cumulative distribution function (c.d.f.) under H_0 and H_1 for a detector with a test statistic of the form (14). The results in Section V, especially an application of Corollary 5.1, simplify this analysis in the large system limit. The following analysis shows that the conditional distributions are a weighted sum of chi-square random variables. For general D , we use a previous algorithm to compute the c.d.f. of this weighted sum of chi-square random variables necessary in the ROC derivation. However, for the deterministic plug-in and RMT detectors, the theoretical ROC curves may be computed in closed form.

A. Stochastic Testing Setting

In the stochastic setting, the conditional distributions of our test samples under each hypothesis are $\hat{w}|H_0 \sim \mathcal{N}(0, I_{\hat{k}})$ and $\hat{w}|H_1 \sim \mathcal{N}(0, \hat{U}^H U \Sigma U^H \hat{U} + I_{\hat{k}})$. Because the covariance matrix of $\hat{w}|H_0$ is diagonal, for $i = 1, \dots, \hat{k}$, $\hat{w}_i|H_0 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, which implies that $\hat{w}_i^2|H_0 \stackrel{\text{i.i.d.}}{\sim} \chi_1^2$. By Corollary 5.1, the covariance matrix of $\hat{w}|H_1$ is asymptotically diagonal. Therefore for $i = 1, \dots, \hat{k}$, $\hat{w}_i|H_1 \stackrel{\text{i.i.d.}}{\approx} \mathcal{N}(0, \sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 + 1)$ and

$$\frac{\hat{w}_i^2|H_1}{\sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 + 1} \sim \chi_1^2.$$

Using this analysis, for a stochastic detector with the form of (14), the conditional distributions of its test statistic under each hypothesis are

$$\Lambda(\hat{w})|H_0 \sim \sum_{i=1}^{\hat{k}} d_i \chi_{1i}^2, \quad \Lambda(\hat{w})|H_1 \sim \sum_{i=1}^{\hat{k}} d_i (\sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 + 1) \chi_{1i}^2 \quad (23)$$

where χ_{1i}^2 are independent chi-square random variables. The third and fourth columns of Table I use this general analysis to summarize the sample conditional distributions of $\Lambda(\hat{w})$ under each hypothesis for the stochastic plug-in and RMT detectors. An analytical expression for the asymptotic performance in the large matrix limit is obtained by substituting expressions from (17) and Propositions 5.1 and 5.2 for the pertinent quantities in these distributions.

Note that the conditional distributions in (23) are a weighted sum of independent chi-square random variables with one degree of freedom. The c.d.f. of a chi-square random variable is known in closed form. However, the c.d.f. of a weighted sum of independent chi-square random variables is not known in closed form. To evaluate (22), we use a saddlepoint approximation of the conditional c.d.f. of $\Lambda(\hat{w})$ by employing the generalized Lugannani-Rice formula proposed in [31]. To then compute a theoretical ROC curve, we sweep γ over

Detector	Detector Statistic $\Lambda(\hat{w})$	Distribution of ΛH_0	Distribution of ΛH_1
Plug-in	$\sum_{i=1}^{\hat{k}} \hat{w}_i^2$	$\chi_{\hat{k}}^2$	$\chi_{\hat{k}}^2 \left(\sum_{i=1}^{\hat{k}_{\text{eff}}} \sigma_i^2 \langle u_i, \hat{u}_i \rangle ^2 x_i^2 \right)$
RMT	$\sum_{i=1}^{\hat{k}_{\text{eff}}} \hat{w}_i^2$	$\chi_{\hat{k}_{\text{eff}}}^2$	$\chi_{\hat{k}_{\text{eff}}}^2 \left(\sum_{i=1}^{\hat{k}_{\text{eff}}} \sigma_i^2 \langle u_i, \hat{u}_i \rangle ^2 x_i^2 \right)$

TABLE II
SUMMARY OF THE PLUG-IN AND RMT DETERMINISTIC MSDS. SEE SECTIONS III-B AND VI-B FOR DERIVATIONS.

$(0, \infty)$ and for each value of γ , we compute the saddlepoint approximation of the conditional c.d.f. under each hypothesis using this method. This generates a set of points (P_F, P_D) which approximate the (asymptotic) theoretical ROC curve.

B. Deterministic Testing Setting

In the deterministic setting, the conditional distribution of a test sample under H_0 is $\hat{w}|H_0 \sim \mathcal{N}(0, I_{\hat{k}})$. The conditional distribution under H_1 is $\hat{w}|H_1 \sim \mathcal{N}(\hat{U}^H U \Sigma^{1/2} x, I_{\hat{k}})$. By Proposition 5.1 and Claim 5.1, $\hat{U}^H U \xrightarrow{\text{a.s.}} BA$ is asymptotically diagonal with B and A defined in Section VI-B. Therefore, $\hat{w}_i|H_1 \stackrel{\text{i.i.d.}}{\approx} \mathcal{N}(a_i b_i \sigma_i x_i, 1)$ for $i = 1, \dots, \hat{k}$. Using this approximation, for a detector with the form of (14), the conditional distributions of its test statistic are

$$\Lambda(\hat{w})|H_0 \sim \sum_{i=1}^{\hat{k}} d_i \chi_{1i}^2 \quad \text{and} \quad \Lambda(\hat{w})|H_1 \sim \sum_{i=1}^{\hat{k}} d_i \chi_{1i}^2(\delta_i) \quad (24)$$

where $\delta_i = \sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 x_i^2$ is the non-centrality parameter for the noncentral chi-square distribution. The deterministic plug-in and RMT detectors are a special case of these conditional distributions. For the plug-in detector, $d_i = 1$ for $i = 1, \dots, \hat{k}$. For the RMT detector $d_i = 1$ for $i = 1, \dots, \hat{k}_{\text{eff}}$ and $d_i = 0$ for $i = \hat{k}_{\text{eff}} + 1, \dots, \hat{k}$.

For the plug-in and RMT detectors, $\Lambda_{\text{plugin}}(\hat{w})|H_0 \sim \chi_{\hat{k}}^2$ and $\Lambda_{\text{rmt}}(\hat{w})|H_0 \sim \chi_{\hat{k}_{\text{eff}}}^2$. Similarly, $\Lambda_{\text{plugin}}(\hat{w})|H_1 \sim \chi_{\hat{k}}^2(\delta)$ and $\Lambda_{\text{rmt}}(\hat{w})|H_1 \sim \chi_{\hat{k}_{\text{eff}}}^2(\delta)$ where

$$\delta = \sum_{i=1}^{\hat{k}} \sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 x_i^2 = \sum_{i=1}^{\hat{k}_{\text{eff}}} \sigma_i^2 |\langle u_i, \hat{u}_i \rangle|^2 x_i^2. \quad (25)$$

Because $d_i = 1$ for $i = 1, \dots, \hat{k}_{\text{eff}}$ for both the plug-in and RMT detectors, the resulting non-centrality parameter is the sum of all the individual non-centrality parameters. An analytical expression for the asymptotic performance in the large matrix limit is obtained by substituting expressions from Proposition 5.1 in (25). Unlike the stochastic setting, we can obtain a closed form expression for the deterministic plug-in and RMT ROC curves by solving for γ in terms of P_F and substituting this into the expression for P_D in (22). Doing so yields

$$\begin{aligned} P_{D_{\text{plugin}}} &= 1 - Q_{\chi_{\hat{k}}^2(\delta)} \left(Q_{\chi_{\hat{k}}^2}^{-1}(1 - P_F) \right) \\ P_{D_{\text{rmt}}} &= 1 - Q_{\chi_{\hat{k}_{\text{eff}}}^2(\delta)} \left(Q_{\chi_{\hat{k}_{\text{eff}}}^2}^{-1}(1 - P_F) \right) \end{aligned} \quad (26)$$

where Q is the appropriate c.d.f. function.

VIII. DISCUSSION AND INSIGHTS

We use numerical simulations to verify our theoretical ROC curve predictions from Section VII that rely on RMT

approximations presented in Section V. We also demonstrate properties of the new RMT detectors that we derived in Section VI, as described next.

A. Simulation Protocol

To compute an empirical ROC curve, we first generate a random subspace, U , by taking the first k left singular vectors of a random matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Using this U , we generate training samples as described in Section II-A from which we form estimates \hat{U} and $\hat{\Sigma}$ from the eigenvalue decomposition of the sample covariance matrix as described in (1).

We then generate a desired number of test samples from each hypothesis using either (2) or (3). For each test sample, we compute the test statistic for each detector. Using Fawcett's [29] 'Algorithm 2', we compute an empirical ROC curve by first sorting the test statistics. At each statistic, we log a (PF, PD) pair by counting the number of lower scores generated from each hypothesis. This is repeated for multiple realizations of U , generating multiple empirical ROC curves. We refer to a single empirical ROC curve corresponding to a realization of U as a trial. We then average the empirical ROC curves over multiple trials using Fawcett's [29] 'Algorithm 4'. This performs threshold averaging by first uniformly sampling the sorted list of all test scores of ROC curves and then computing (PF, PD) pairs in the same way as 'Algorithm 2'.

B. Convergence and Accuracy of ROC Curve Predictions

The theoretical ROC curve predictions for the plug-in and RMT detectors rely on the asymptotic approximations that ignore finite n and m correction terms. To examine the validity of the asymptotic approximations (Propositions 5.1 and 5.2, Theorem 5.1, and Corollary 5.1) and the rate of convergence, we consider two different settings for the stochastic plug-in detector. Figures 3(a)-3(b) plot three empirical ROC curves for $n = 50, 200, 1000$ as well as the theoretically predicted plug-in ROC curve. Each figure uses different values of k and c but in each case, $\hat{k} = k$.

For both figures, as n increases, the empirical ROC curves approach the theoretical prediction, attesting to the asymptotic convergence of the RMT approximations. Analyzing the rate of convergence (which we conjecture to be $n^{1/2}$ for fixed k and c) is an important open problem that we shall tackle in future work. As evident in Figures 3(a)-3(b) the values of k and c play an important role in the convergence of the empirical ROC curves. For the larger value of k and c (corresponding to the sample starved regime where the amount of training data is smaller than the system dimensionality i.e. $n > m$) the convergence is also slower. We see that for larger k and c , when n is small the empirical ROC curve is not

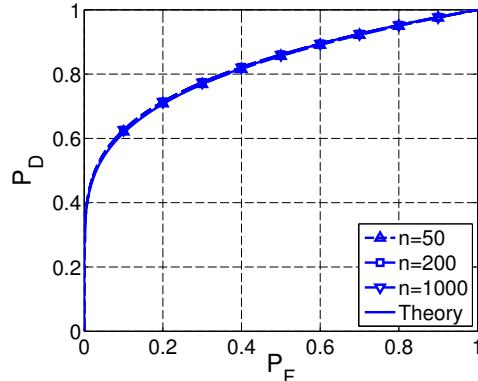
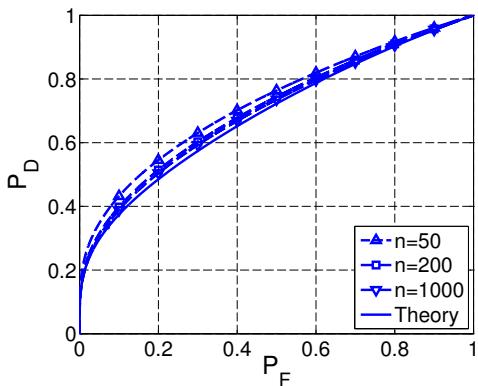
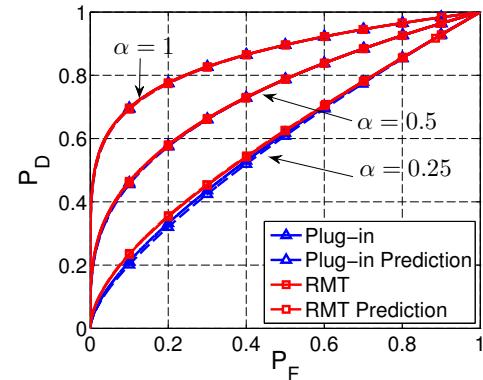
(a) $k = 2, c = 1$ (b) $k = 4, c = 10$

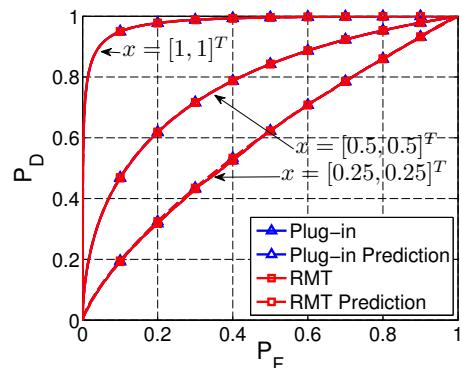
Fig. 3. Empirical and theoretical ROC curves for the stochastic plug-in detector. Empirical ROC curves were simulated using 10000 test samples and averaged over 50 trials using algorithms 2 and 4 of [29]. (a) $\Sigma = \text{diag}(10, 2)$, $c = 1$, $\hat{k} = k = 2$ so that $k_{\text{eff}} = 2$. (b) $\Sigma = \text{diag}(10, 2, 0.5, 0.1)$, $c = 10$, $\hat{k} = k = 4$ so that $k_{\text{eff}} = 1$. Each figure plots empirical ROC curves for $n = 50, 200, 1000$. Theoretical ROC curves were computed as described in Section VII. As n increases, the empirical ROC curves approach the theoretically predicted one. However, this convergence is slower for larger k and c .

well approximated by the asymptotic theoretical predictions. However, as n increases, the deviation of the empirically generated ROC curve from the theoretically predicted one decreases. Claim 5.1 suggests that the off diagonal terms of $\widehat{U}^H U$ asymptotically tend to zero. However, in the finite n and m case these terms are $O(1/\sqrt{n})$ and thus not identically zero. For larger rank systems (increased k), there are more of these non-identically-zero terms that worsen the approximation quality for fixed, relatively small n . As n increases, this bias vanishes.

The ROC predictions developed in Section VII also depend on parameters such as Σ and the deterministic vector x . To test the accuracy of the ROC predictions with respect to these parameters, we consider a setting where $\hat{k} = k = 2$. Figure 4(a) plots empirical and theoretical ROC curves for the plug-in and RMT stochastic detectors for $\Sigma = \alpha \text{diag}(10, 5)$ for three choices of α . As intuition suggests, smaller values of Σ decrease the performance for both the plug-in and RMT detectors. For each choice of α , the empirical ROC curves match the ROC predictions that rely on random matrix theoretic approximations presented in Section V. Using $\alpha = 1$ or $\alpha = 0.5$ results in $k_{\text{eff}} = k = \hat{k} = 2$ but using $\alpha = 0.25$ results in $k_{\text{eff}} = 1$. As $k > k_{\text{eff}}$ for this last case, the plug-



(a) Stochastic



(b) Deterministic

Fig. 4. Empirical and theoretical ROC curves for the plug-in and RMT detectors. Empirical ROC curves were simulated using 10000 test vectors and averaged over 100 trials with $n = 1000$, $m = 500$, and $\Sigma = \alpha \text{diag}(10, 5)$. The theoretical ROC curves were computed as described in Section VII. (a) Stochastic testing setting. Results are plotted for $\alpha = 1, 0.5, 0.25$. For $\alpha = 1$ and $\alpha = 0.5$, $\hat{k} = k = k_{\text{eff}} = 2$ by (16). For $\alpha = 0.25$, $k_{\text{eff}} = 1$. Since $\hat{k} > k_{\text{eff}}$ when $\alpha = 0.25$, we observe a performance gain when using the RMT detector. (b) Deterministic testing setting. Results are plotted for $\alpha = 1$ so that $k_{\text{eff}} = 2$. Three values of the deterministic signal vector were used: $x = [1, 1]^T$, $x = [0.5, 0.5]^T$, and $x = [0.25, 0.25]^T$. The resulting ROC curves depend on the choice of x , however, since $\hat{k} = k_{\text{eff}}$, the plug-in and RMT detector achieve the same performance for all x . For both the stochastic and deterministic detectors, the theoretically predicted ROC curves match the empirical ROC curves, reflecting the accuracy of Corollary 5.1 and the Lugannani-Rice formula.

in detector realizes a performance loss compared to the RMT detector.

In the deterministic setting, x is an additional parameter that affects detector performance. Figure 4(b) plots empirical and theoretical ROC curves for the plug-in and RMT deterministic detectors for $\Sigma = \text{diag}(10, 5)$ for three choices of the deterministic test vector x . Larger values of $|x|$ result in better detector performance but for each choice of x , the theoretically predicted ROC curves match their empirical counterparts. As x does not affect the value of $k_{\text{eff}} = k = \hat{k} = 2$, the plug-in and RMT detectors achieve the same performance because they have identical statistics. For both test vector models, the theoretical ROC curves match the empirical ROC curves thereby validating the accuracy of the random matrix theoretic approximations employed and the accuracy of the saddlepoint approximation to the c.d.f. used in the stochastic derivation.

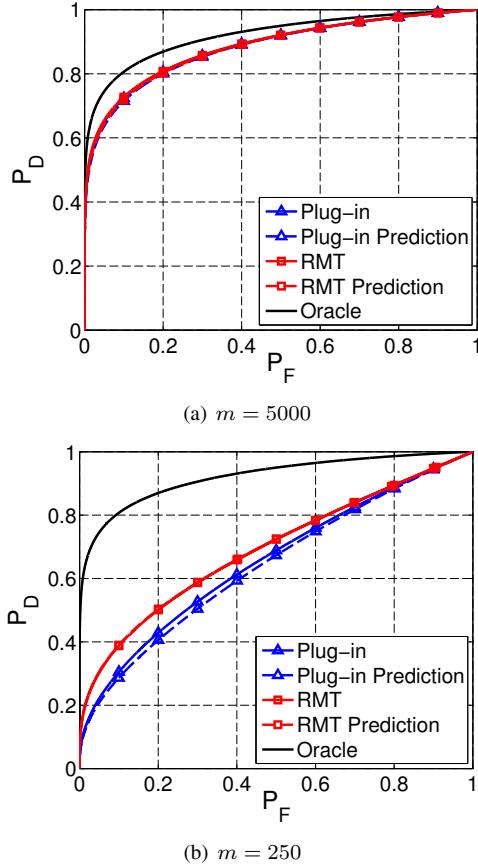


Fig. 5. Empirical and theoretical ROC curves for the plug-in and RMT stochastic detectors. Empirical ROC curves were computed with 10000 test samples and averaged over 100 trials. Here, $n = 5000$, $\hat{k} = k = 4$ and $\Sigma = \text{diag}(10, 3, 2.5, 2)$. The empirical oracle ROC curve is provided for relative comparison purposes. (a) $m = 5000$ so that $c = 1$ and $k_{\text{eff}} = \hat{k} = 4$. The plug-in and RMT detectors achieve relatively the same performance. (b) $m = 250$ so that $c = 20$ and $k_{\text{eff}} = 1 < \hat{k} = 4$. The RMT detector avoids some of the performance loss realized by the plug-in detector. As seen in Section III, limited training samples degrades detector performance. However, the new RMT detector does not suffer as badly as the plug-in detector because it accounts for subspace estimation errors due to finite training data. The disagreement between the theoretical and empirical ROC curves is attributed to finite dimensionality.

C. Effect of the Number of Training Samples

We saw in Section III-C that finite training data degraded the performance of the plug-in detector relative to that of the oracle detector. The analysis of Section V mathematically justifies this observation showing that, for a fixed Σ , the number of training samples, m , directly affects k_{eff} via (16). While the plug-in detector ignores this analysis, we derived a new RMT detector that accounts for subspace estimation errors due to finite training data. By only using the k_{eff} informative signal subspace components, we hope that the RMT detector will avoid some of the performance loss associated with the plug-in detector. To explore how the number of training samples affects the relative performances of the plug-in and RMT detectors, we first consider the setting where $\hat{k} = k = 4$ with $\Sigma = \text{diag}(10, 3, 2.5, 2)$.

Figure 5(a) investigates the performance when $m = n$ so that $c = 1$ for the stochastic setting. This choice of m results in $k_{\text{eff}} = \hat{k} = 4$. As expected, the plug-in and RMT detectors achieve relatively the same performance because $\hat{k} = k_{\text{eff}}$. A similar phenomenon occurs in the deterministic setting. Figure

5(b) chooses $20m = n$ so that $c = 20$ and $k_{\text{eff}} = 1$ for the stochastic settings. This corresponds to the sample starved regime where $m < n$. In this second experiment, the plug-in detector becomes suboptimal because it uses $4 = \hat{k} > k_{\text{eff}} = 1$ subspace components. A similar phenomenon occurs in the deterministic setting. Whenever $k_{\text{eff}} < \hat{k}$ the RMT detectors avoid some of the performance loss (compared to the oracle detectors) realized by the plug-in detectors. We could have observed this same effect by instead varying Σ as both of these quantities drive the value of k_{eff} . The disagreement between the theoretical and empirical stochastic ROC curves for the plug-in detector is attributed to the finite n and m correction terms, which we have discussed previously.

Figure 5 shows that the number of training samples helps to drive the performance of matched subspace detectors. In Section III, we mathematically defined the performance loss of a detector relative to its oracle detector as ϵ in (13) and empirically plotted the number of training samples needed to achieve a desired performance loss for the stochastic plug-in detector in Figure 2. Figures 6(a) and 6(b) theoretically plot this same curve for the plug-in and RMT detectors for each testing setting, respectively.

These figures show that when $k_{\text{eff}} < \hat{k}$, the RMT detector achieves a much smaller performance loss for a fixed number of training samples. Put another way, to achieve the same performance loss, the RMT detectors need a significantly less number of training samples when $k_{\text{eff}} < \hat{k}$. Figure 6(a) shows that the stochastic detectors can achieve an arbitrarily small performance loss given a particularly large number of training samples. However, Figure 6(b) shows that there is a performance loss limit for the deterministic detectors. As discussed in Section III, this arises because the oracle deterministic detector assumes that x is known. As $m \rightarrow \infty$, $\hat{U} \rightarrow U$ and $\hat{\Sigma} \rightarrow \Sigma$, however, the plug-in detector's estimate of \hat{x} still depends on the noisy observed data y . Therefore, unlike the stochastic detectors that can achieve an arbitrarily small performance loss, the deterministic plug-in and RMT detectors can never achieve the same performance as the deterministic oracle detector.

D. Effect of \hat{k}

We discussed in Section II-B that we are given a dimension estimate \hat{k} when deriving our detector. From our perspective, we don't know how \hat{k} was estimated (possibly from the training data or by a domain expert) but simply use it when forming our subspace and signal covariance estimates. Figure 7 empirically examines the performance of the plug-in and RMT detectors as a function of \hat{k} for the stochastic setting. A similar phenomenon arises in the deterministic setting. Here, we relax the constraint that $\hat{k} \geq k$. The figures plot the achieved probability of detection for a constant false alarm rate of 0.01. The result confirms that k_{eff} is the optimal choice for \hat{k} . When the plug-in detectors use $\hat{k} = k_{\text{eff}}$ they achieve an equivalent performance as that of the RMT detector.

Setting $\hat{k} < k_{\text{eff}}$ drastically degrades performance for all detectors. In this regime, the plug-in and RMT detectors realize the same ROC performance, demonstrating that quantification and exploitation of the subspace estimation accuracy

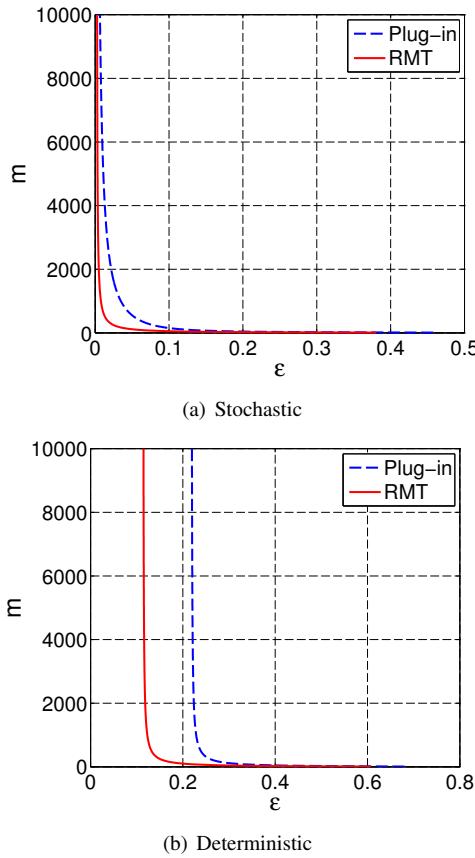


Fig. 6. Theoretically determined number of training samples, m , needed to achieve a desired performance loss, ϵ , as defined in (13). The required false alarm rate is $P_F = 0.1$ with $n = 200$, $\Sigma = \text{diag}(10, 0.1)$, and $\hat{k} = k = 2$. (a) Results for the stochastic detectors. We see that for a given ϵ , the new RMT detector requires less training samples. (b) Results for the deterministic detectors when $x = [0.75, 0.75]^T$. Again, for a given ϵ , the new RMT detector requires less training samples. In the deterministic setting, the limiting performance loss is different (and non-zero) for the plug-in and RMT detectors. This arises in estimation errors of x in the GLRT.

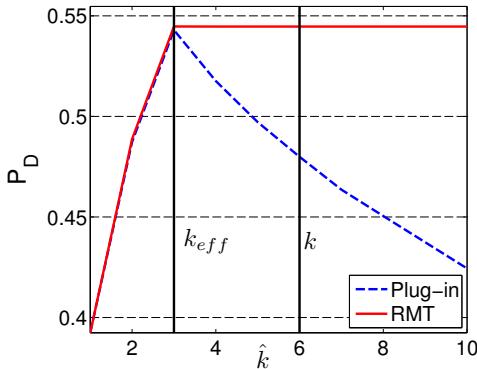


Fig. 7. Empirical exploration of the achieved probability of detection, P_D , for a fixed probability of false alarm, $P_F = 0.01$, for various \hat{k} . Empirical ROC curves were computed using 10000 test samples and averaged over 100 trials with $n = 1000$, $m = 500$, and $\Sigma = \text{diag}(10, 5, 4, 0.75, 0.5, 0.25)$ so that $k_{\text{eff}} = 3$. Results for the stochastic detectors. The optimal \hat{k} resulting in the largest P_D is not the true k , but rather k_{eff} .

($|\langle u_i, \hat{u}_i \rangle|_{\text{rmt}}^2$ and $\sigma_{i_{\text{rmt}}}^2$), while useful in ROC performance prediction, does *not* noticeably enhance detection performance. When $\hat{k} > k_{\text{eff}}$, the performances of the plug-in detectors degrade while those of the RMT detectors are stable as if $\hat{k} = k_{\text{eff}}$. In other words, we do not pay a price for overestimating the subspace dimension with the RMT detectors. This makes sense (and is slightly contrived) because the RMT detectors will only sum to a maximum of k_{eff} indices as evident in (20) and (21). In many applications, practitioners might employ the “play-it-safe” approach and set \hat{k} to be significantly greater than k_{eff} . The performance loss caused by adding each uninformative subspace, as seen in Figure 7, constitutes evidence to the assertion that overestimating the signal subspace dimension is a bad idea. When $k_{\text{eff}} < k$, even perfectly estimating the subspace dimension (i.e. setting $\hat{k} = k$) is suboptimal.

IX. CONCLUSION

In this paper, we considered a matched subspace detection problem where the low-rank signal subspace is unknown and must be estimated from finite, noisy, signal-bearing training data. We considered both a stochastic and deterministic model for the testing data. The subspace estimate is inaccurate due to finite and noisy training samples and therefore degrades the performance of plug-in detectors compared to an oracle detector. We showed how the ROC performance curve can be derived from the RMT-aided quantification of the subspace estimation accuracy.

Armed with this RMT knowledge, we derived a new RMT detector that only uses the effective number of informative subspace components, k_{eff} . Plug-in detectors that use the uninformative components will thus incur a performance degradation, relative to the RMT detector. In settings where a practitioner might play-it-safe and set $\hat{k} > \hat{k}_{\text{eff}}$, the performance loss is significant (see Figures 6(a) and 6(b) for a demonstration of how much training data such a play-it-safe plug-in detector would need to match the performance of a k_{eff} -tuned RMT detector). This highlights the importance of robust techniques [24], [25], [26] for estimating k_{eff} in subspace based detection schemes as opposed to estimating k , particularly in the regime where $k_{\text{eff}} < k$. We showed in Tables I and II that the distributions of the test statistics could be expressed as a weighted sum of independent chi-squared random variables. The associated ROC curves can then be computed using a saddlepoint approximation.

The results in this paper can be extended in several directions. We note that the stochastic detector setting assumed normally distributed training and test data. We can extend the analysis to the Gaussian training data but non-Gaussian test vector setting by ‘integrating-out’ the deterministic detector performance curves with respect to the non-Gaussian distribution of the test-vector. Our results relied on characterization of the quantity $\langle u_j, \hat{u}_i \rangle$. Thus analogous performance curves can be obtained for any alternate training data models for which this quantity can be analytically quantified. To that end, the results in [21] facilitate such an analysis for a broader class of models including the correlated Gaussians training data

setting. An extension to the missing data setting might follow a similar approach and appears within reach. Aspects related to rate of convergence are open and will be the subject of future work.

APPENDIX

Theorem 5.1: Assume the same hypothesis as in Proposition 5.1. Let $\hat{k} = k_{\text{eff}} = k$. For $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, and $i \neq j$, as $n, m \rightarrow \infty$ with $n/m \rightarrow c$, then $\langle u_j, \hat{v}_i \rangle \xrightarrow{\text{a.s.}} 0$.

Proof: Let $U_{n,k}$ be a $n \times k$ real or complex matrix with orthonormal columns, u_i for $1 \leq i \leq k$. Let $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_k^2)$ such that $\sigma_1^2 > \sigma_2^2 > \dots > \sigma_k^2 > 0$ for $k \geq 1$. Define $P_n = U_{n,k} \Sigma U_{n,k}^H$ so that P_n is rank- k . Let Z_n be a $n \times m$ real or complex matrix with independent $\mathcal{CN}(0, 1)$ entries. Let $X_n = \frac{1}{m} Z_n Z_n^H$, which is a random Wishart matrix, have eigenvalues $\lambda_1(X_n) \geq \dots \geq \lambda_n(X_n)$. Let $\hat{X}_n = X_n(I_n + P_n)$. X_n and P_n are independent by assumption. Define the empirical eigenvalue distribution as $\mu_{X_n} = \frac{1}{n} \sum_{j=1}^n \delta_{\lambda_j(X_n)}$. We assume that as $n \rightarrow \infty$, $\mu_{X_n} \xrightarrow{\text{a.s.}} \mu_X$.

For $i = 1, \dots, \hat{k} = k$, let \hat{v}_i be an arbitrary unit eigenvector of \hat{X}_n . By the eigenvalue master equation, $\hat{X}_n \hat{v}_i = \hat{\lambda}_i \hat{v}_i$, it follows that

$$U_{n,k}^H (\hat{\lambda}_i I_n - X_n)^{-1} X_n U_{n,k} \Sigma U_{n,k}^H \hat{v}_i = U_{n,k}^H \hat{v}_i. \quad (27)$$

Let $X_n = V_n \Lambda_n V_n^H$ be the eigenvalue decomposition of X_n such that $\Lambda_n = \text{diag}(\lambda_1(X_n), \dots, \lambda_n(X_n))$ and $\lambda_1(X_n) \geq \dots \geq \lambda_n(X_n)$. Using this decomposition and defining $W_{n,k} = V^H U_{n,k}$, (27) simplifies to

$$W_{n,k}^H (\hat{\lambda}_i I_n - \Lambda_n)^{-1} \Lambda_n W_{n,k} \Sigma U_{n,k}^H \hat{v}_i = U_{n,k}^H \hat{v}_i. \quad (28)$$

Define the columns of $W_{n,k}$ to be $w_j^{(n)} = [w_{1,j}^{(n)}, \dots, w_{n,j}^{(n)}]^T$ for $j = 1, \dots, k$. These columns are orthonormal and isotropically random. We can rewrite (28) as

$$\left[T_{\mu_{r,j}^{(n)}} (\hat{\lambda}_i) \right]_{r,j=1}^k \Sigma U_{n,k}^H \hat{v}_i = U_{n,k}^H \hat{v}_i \quad (29)$$

where for $r = 1, \dots, k$, $j = 1, \dots, k$, $\mu_{r,j}^{(n)} = \sum_{\ell=1}^n \overline{w_{\ell,r}^{(n)}} w_{\ell,j}^{(n)} \delta_{\lambda_\ell(X_n)}$ is a complex measure and $T_{\mu_{r,j}^{(n)}}(\hat{\lambda}_i)$ is the T-transform defined by $T_\mu(z) = \int \frac{t}{z-t} d\mu(t)$ for $z \notin \text{supp } \mu$. We may rewrite (29) as

$$\left(I_k - \left[\sigma_j^2 T_{\mu_{r,j}^{(n)}} (\hat{\lambda}_i) \right]_{r,j=1}^k \right) U_{n,k}^H \hat{v}_i = 0.$$

Therefore, $U_{n,k}^H \hat{v}_i$ must be in the kernel of $M_n(\hat{\lambda}_i) = I_k - \left[\sigma_j^2 T_{\mu_{r,j}^{(n)}} (\hat{\lambda}_i) \right]_{r,j=1}^k$. By Proposition 9.3 of [20]

$$\mu_{r,j}^{(n)} \xrightarrow{\text{a.s.}} \begin{cases} \mu_X & \text{for } i = j \\ \delta_0 & \text{o.w.} \end{cases}$$

where μ_X is the limiting eigenvalue distribution of X_n . Therefore,

$$M_n(\hat{\lambda}_i) \xrightarrow{\text{a.s.}} \text{diag}\left(1 - \sigma_1^2 T_{\mu_X}(\hat{\lambda}_i), \dots, 1 - \sigma_k^2 T_{\mu_X}(\hat{\lambda}_i)\right).$$

As $k_{\text{eff}} = k$, for $i = 1, \dots, k$, $\sigma_i^2 > 1/T_{\mu_X}(b^+)$, where b is the supremum of the support of μ_X . As λ_i is the eigenvalue corresponding to the eigenvector \hat{v}_i , by Theorem 2.6 of [20] $\hat{\lambda}_i \xrightarrow{\text{a.s.}} T_{\mu_X}^{-1}(1/\sigma_i^2)$. Therefore,

$$M_n(\hat{\lambda}_i) \xrightarrow{\text{a.s.}} \text{diag}\left(1 - \frac{\sigma_1^2}{\sigma_i^2}, \dots, 1 - \frac{\sigma_{i-1}^2}{\sigma_i^2}, 0, 1 - \frac{\sigma_{i+1}^2}{\sigma_i^2}, \dots, 1 - \frac{\sigma_k^2}{\sigma_i^2}\right) \quad (30)$$

Recall that $U_{n,k}^H \hat{v}_i$ must be in the kernel of $M_n(\hat{\lambda}_i)$. Therefore, any limit point of $U_{n,k}^H \hat{v}_i$ is in the kernel of the matrix on the right hand side of (30). Therefore, for $i \neq j$, $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, we must have that $\left(1 - \frac{\sigma_j^2}{\sigma_i^2}\right) \langle u_j, \hat{v}_i \rangle = 0$. As $\sigma_i^2 \neq \sigma_j^2$, for this condition to be satisfied we must have that for $j \neq i$, $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, $\langle u_j, \hat{v}_i \rangle \xrightarrow{\text{a.s.}} 0$.

Recall that our observed vectors $y_i \in \mathbb{C}^{n \times 1}$ have covariance matrix $U_{n,k} \Sigma U_{n,k}^H + I_n = P_n + I_n$. Therefore, our observation matrix, Y_n which is a $n \times m$ matrix, may be written $Y_n = (P_n + I_n)^{1/2} Z_n$. The sample covariance matrix, $S_n = \frac{1}{m} Y_n Y_n^H$, may be written $S_n = (I_n + P_n)^{1/2} X_n (I_n + P_n)^{1/2}$. By similarity transform, if \hat{v}_i is a unit-norm eigenvector of \hat{X}_n then $\hat{s}_i = (I_n + P_n)^{1/2} \hat{v}_i$ is an eigenvector of S_n . If $\hat{u}_i = \hat{s}_i / \|\hat{s}_i\|$ is a unit-norm eigenvector of S_n , it follows that

$$\langle u_j, \hat{v}_i \rangle = \frac{\sqrt{\sigma_i^2 + 1} \langle u_j, \hat{v}_i \rangle}{\sqrt{\sigma_i^2} |\langle u_j, \hat{v}_i \rangle|^2 + 1}$$

As $\langle u_j, \hat{v}_i \rangle \xrightarrow{\text{a.s.}} 0$ for all $i \neq j$, $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, it follows that $\langle u_j, \hat{u}_i \rangle \xrightarrow{\text{a.s.}} 0$ for all $i \neq j$, $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$.

Claim 5.1: We conjecture that this result holds for the general case of $i \neq j$, $i = 1, \dots, \hat{k}$, $j = 1, \dots, k$, not just when $k = k_{\text{eff}} = k$. Consider the case when $k = 1$. For $i > 2$, if $\hat{\lambda}_i$ is an eigenvalue of $\hat{X}_n = X_n(I_n + \sigma^2 uu^H)$, then it satisfies $\det(\hat{\lambda}_i I_n - X_n(I_n + \sigma^2 uu^H)) = \det(\hat{\lambda}_i I_n - X_n) \det(I_n - (\hat{\lambda}_i I_n - X_n)^{-1} X_n \sigma^2 uu^H) = 0$. Therefore, if $\hat{\lambda}_i$ is not an eigenvalue of X_n , the corresponding unit norm eigenvector \hat{v}_i is in the kernel of $I_n - (\hat{\lambda}_i I_n - X_n)^{-1} X_n \sigma^2 uu^H$. Therefore

$$|\langle \hat{v}_i, u \rangle|^2 = \frac{1}{\sigma^4 u^H X_n (\hat{\lambda}_i I_n - X_n)^{-2} X_n u}.$$

Recall that Weyl's interlacing lemma for eigenvalues gives $\lambda_i(X_n) \leq \hat{\lambda}_i \leq \lambda_{i-1}(X_n)$. Letting $X_n = V_n \Lambda_n V_n^H$ and $w = V_n^H u$, we see the importance of the asymptotic spacing of eigenvalues of X_n in

$$\begin{aligned} u^H X_n (\hat{\lambda}_i I_n - X_n)^{-2} X_n u &= \sum_{\ell=1}^n \frac{|w_\ell|^2 \lambda_\ell^2(X_n)}{(\hat{\lambda}_i - \lambda_\ell)^2} \\ &\geq \frac{\min_j \lambda_j^2(X_n) \min_j |w_j|^2}{\max_j |\lambda_{j-1} - \lambda_j|^2} \end{aligned}$$

In [32] it is shown that $\min_j \lambda_j^2(X_n) = \lambda_n^2(X_n) \xrightarrow{\text{a.s.}} (1 - \sqrt{c})^2$. The typical spacing between eigenvalues is $O(1/n)$ while the typical magnitude of w_j^2 is $O(1/n)$ [33]. Therefore, the right hand side of the above inequality will typically be $O(n)$ and we get the desired result of $|\langle \hat{v}_i, u \rangle|^2 \xrightarrow{\text{a.s.}} 0$. More generally, it is the behavior of the largest eigenvalue gap

and the smallest element of w_i that drives this convergence. Thus, so long as the eigenvector whose elements are w_i are delocalized (i.e. having elements of $O(1/\sqrt{n})$) and the smallest gap between k successive eigenvalues is at least as large as $O(1/(n^{0.5+\epsilon}))$, the right hand side of the inequality will be unbounded with n . The claim follows after applying a similarity transform as in the proof of Theorem 5.1.

■

ACKNOWLEDGMENT

This work is supported by the ONR Young Investigator Program under Grant N00014-11-1-0660.

REFERENCES

- [1] L. L. Scharf and C. Demeure, *Statistical signal processing: detection, estimation, and time series analysis*. Addison-Wesley Publishing Company, 1991, vol. 1.
- [2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics, 2001, vol. 1.
- [3] O. Besson, L. L. Scharf, and F. Vincent, "Matched direction detectors and estimators for array processing with subspace steering vector uncertainties," *Signal Processing, IEEE Transactions on*, vol. 53, no. 12, pp. 4453–4463, 2005.
- [4] O. Besson and L. L. Scharf, "Cfar matched direction detector," *Signal Processing, IEEE Transactions on*, vol. 54, no. 7, pp. 2840–2844, 2006.
- [5] F. Bandiera, A. De Maio, A. S. Greco, and G. Ricci, "Adaptive radar detection of distributed targets in homogeneous and partially homogeneous noise plus subspace interference," *Signal Processing, IEEE Transactions on*, vol. 55, no. 4, pp. 1223–1237, 2007.
- [6] F. Bandiera, O. Besson, D. Orlando, G. Ricci, and L. L. Scharf, "Glrt-based direction detectors in homogeneous noise and subspace interference," *Signal Processing, IEEE Transactions on*, vol. 55, no. 6, pp. 2386–2394, 2007.
- [7] E. Maris, "A resampling method for estimating the signal subspace of spatio-temporal eeg/meg data," *Biomedical Engineering, IEEE Transactions on*, vol. 50, no. 8, pp. 935–949, 2003.
- [8] A. C. K. Soong and Z. J. Koles, "Principal-component localization of the sources of the background eeg," *Biomedical Engineering, IEEE Transactions on*, vol. 42, no. 1, pp. 59–67, 1995.
- [9] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *Signal Processing, IEEE Trans. on*, vol. 42, no. 8, pp. 2146–2157, 1994.
- [10] F. Vincent, O. Besson, and C. Richard, "Matched subspace detection with hypothesis dependent noise power," *Signal Processing, IEEE Transactions on*, vol. 56, no. 11, pp. 5713–5718, 2008.
- [11] T. McWhorter and L. L. Scharf, *Matched subspace detectors for stochastic signals*. Defense Technical Information Center, 2003.
- [12] Y. Jin and B. Friedlander, "A cfar adaptive subspace detector for second-order gaussian signals," *Signal Processing, IEEE Trans. on*, vol. 53, no. 3, pp. 871–884, 2005.
- [13] L. Elden, "Matrix methods in data mining and pattern recognition," 2007.
- [14] Y. LeCun and C. Cortes, "The MNIST database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, [Online; accessed 10-October-2012].
- [15] B. Thai and G. Healey, "Invariant subpixel material detection in hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 3, pp. 599–608, 2002.
- [16] G. Healey and D. Slater, "Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 6, pp. 2706–2717, 1999.
- [17] H. Kwon and N. M. Nasrabadi, "Kernel matched subspace detectors for hyperspectral target detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 178–194, 2006.
- [18] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *Signal Processing, IEEE Trans. on*, vol. 56, no. 7, pp. 2625–2638, 2008.
- [19] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, p. 1617, 2007.
- [20] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Adv. in Math.*, 2011.
- [21] ———, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Arxiv preprint arXiv:1103.2221*, 2011.
- [22] N. Asendorf and R. R. Nadakuditi, "Improving and characterizing the performance of stochastic matched subspace detectors when using noisy estimated subspaces," in *Proceedings of the Asilomar Conference of Signals and Systems*, November 2011.
- [23] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 918–930, 2006.
- [24] R. R. Nadakuditi and J. W. Silverstein, "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 3, pp. 468–480, 2010.
- [25] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *The Ann. of stat.*, vol. 29, no. 2, pp. 295–327, 2001.
- [26] N. El Karoui, "Tracy-widom limit for the largest eigenvalue of a large class of complex sample covariance matrices," *The Ann. of Prob.*, vol. 35, no. 2, pp. 663–714, 2007.
- [27] R. J. Muirhead, *Aspects of multivariate statistical theory*. Wiley Online Library, 1982, vol. 42.
- [28] H. L. Van Trees, *Detection, estimation, and modulation theory: Detection, estimation, and linear modulation theory*. Wiley, 1968.
- [29] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [30] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.
- [31] A. T. A. Wood, J. G. Booth, and R. W. Butler, "Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions," *Journal of the American Statistical Association*, pp. 680–686, 1993.
- [32] J. Silverstein, "The smallest eigenvalue of a large dimensional wishart matrix," *The Annals of Probability*, vol. 13, no. 4, pp. 1364–1368, 1985.
- [33] A. Barvinok, "Measure concentration lecture notes," See <http://www.math.lsa.umich.edu/barvinok/total710.pdf>, 2005.

Nicholas Asendorf received the B.S. degree in computer engineering from the University of Maryland, College Park, MD, in 2010 and the M.S. degree in electrical engineering:systems from the University of Michigan in 2012.



He is a graduate student in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. His research interests include the need for data driven algorithms in statistical signal processing and machine learning particularly in low SNR and sample starved settings.

Raj Rao Nadakuditi is an Assistant Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. He received his PhD in 2007 from the Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution. He was awarded an Office of Naval Research Young Investigator Award in 2011, an Air Force Office of Scientific Research Young Investigator Award in 2012 and the Signal Processing Society Best Young Author Paper Award in 2012. His research focuses on developing theory for random

matrices for applications such as signal processing, machine learning, queuing theory and scattering theory.

