Learning a Secondary Source From Compressive Measurements for Adaptive Projection Design

Fraser K. Coutts, John Thompson, and Bernard Mulgrew
Institute for Digital Communications, University of Edinburgh, Edinburgh, EH9 3FG, UK
email: fraser.coutts@ed.ac.uk

2021 Sensor Signal Processing for Defence Conference (SSPD) | 978-1-6654-3314-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/SSPD51364.2021.9541447

***Abstract*—Recent work has established that the gradient of the mutual information (MI) in Gaussian channels with input noise can be used for projection design in a compressive sensing (CS) scenario with two independent, Gaussian mixture (GM)-distributed inputs. The resulting CS projection matrices have been shown to be capable of controlling the input-output MI terms for the two inputs. One downside of such information-theoretic strategies is their reliance on access to *a priori* knowledge of the input source statistics. In this paper, we assume that the GM distribution of a primary input is known and that the GM distribution of a secondary input is unknown. We derive a methodology for the online training of the distribution of the secondary input via compressive measurements and illustrate that once the distribution of this secondary source is known, we can use projection design to control the input-output MI of the system. We demonstrate through simulations the various performance trade-offs that exist within the proposed methodology.**

I. Introduction

Dimensionality reduction methods based on linear random projections — i.e., compressive sensing [1] (CS) — have gained significant attention recently; however, random projections may not be the best choice if we know the statistical properties of the source signal [2]. By employing an information-theoretic approach, one can design a linear projection such that the mutual information (MI) between the projected signal and the source signal or its class label is maximised [3], [4]. Intuitively, as the MI increases, the recovery of the source signal or label information improves; indeed, the Bayes classification error is bounded by the MI [3].

The distribution of a non-Gaussian signal can be approximated by a mixture of several Gaussians [5]. Importantly, increasing the number of Gaussians used enables the approximation of general distributions to an arbitrary level of accuracy [6]. Such Gaussian mixture models (GMMs) have been shown to be effective [7] and in some cases superior to sparse signal models in CS scenarios [5]. Recent work [8] utilises MI maximisation within a CS framework to optimise information throughput for a Gaussian mixture (GM) source signal in the presence of GM input noise. In [9], this framework is extended to complex signal models and applied to real radar data containing micro-Doppler (m-D) signatures [10]; subsequent results highlight that the methodology is able to assist in the joint classification of the m-D signatures of a primary, always-present source and a secondary, fleeting source. By modelling two independent inputs via GMMs and treating each as a source of structured input noise for the other, both [8] and [9] employ an iterative gradient-ascent approach to design a linear projection matrix capable of controlling the information throughput for each source. However, these works rely on *a priori* knowledge of the source statistics and are therefore limited to the case of stationary sources.

In [11], Yang *et al.* investigate a scenario in which a desired source in a CS scenario without input noise has an unknown GM distribution. They seek to learn the distribution using only knowledge of the compressive measurements, the projection matrices involved, and the parameters of the Gaussian measurement noise. Their implementation is iterative and related to the expectation-maximisation (EM) algorithm [12].

In this paper, we consider the compressive measurement of two GM-distributed inputs; these perceive each other as additive noise and experience the same linear projection. We assume that the GM distribution of the primary input is known and that the GM distribution of the secondary input is unknown. We extend the work of [11] and derive a novel methodology for the training of the GM distribution of the secondary input from compressive measurements and illustrate that once the distribution of this secondary source is known, we can use the projection design techniques of [8], [9] to control the input-output MI of the system. For this demonstration, we apply the developed adaptive projection design algorithm to real radar data containing two coincident sources of m-D information. Using synthetic data, we also show the various performance trade-offs that exist within the proposed distribution learning methodology.

Below, Sec. II establishes the signal model considered in this paper. In Sec. III, we summarise the optimisation framework from [8], [9] that we use for projection design. In Sec. IV, we introduce the theory and algorithm required to learn the GM distribution of a secondary source from compressive measurements. Sec. V then briefly explains how these projection design and source learning procedures fit together within an adaptive algorithm. Sec. VI provides a practical demonstration of the proposed approach, and conclusions are drawn in Sec. VII.

Notation: Straight bold lowercase and uppercase symbols denote vectors and matrices, respectively, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Italicised uppercase letters such as $\boldsymbol{Y}$ and C denote random vectors and variables; their realisations are lowercase equivalents, such as $\boldsymbol{y}$ and c. Operators $\{\cdot\}^{\mathrm{H}}$, $\{\cdot\}^{*}$, $\mathbb{E}[\cdot]$, $\|\cdot\|_1$, and $\mathrm{tr}\{\cdot\}$ evaluate the Hermitian transpose, complex conjugate, expectation, ℓ_1-norm, and trace, respectively.

II. Signal Model

We consider the following complex-valued signal model with input noise:

$$\boldsymbol{Y} = \boldsymbol{\Phi}(\boldsymbol{X} + \boldsymbol{N}) + \boldsymbol{W}\,. \quad (1)$$

978-1-6654-3314-3/21/$31.00 ©2021 IEEE

Following the CS protocol, we have measurements $\boldsymbol{Y} \in \mathbb{C}^m$ obtained from input signals $\boldsymbol{X} \in \mathbb{C}^n$ and $\boldsymbol{N} \in \mathbb{C}^n$ via a compressive projection matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times n}$, with $m \ll n$. We opt for $\boldsymbol{X}$ and $\boldsymbol{N}$ to have a GM distribution for two reasons. Firstly, GMs are known to be effective in statistical CS scenarios [5], [7]. Secondly, a GM-distributed $\boldsymbol{X}$ permits a natural extension of the work of [11] to the case of estimating $\boldsymbol{N}$ given knowledge of $\boldsymbol{Y}$, $\boldsymbol{\Phi}$, and the distributions of $\boldsymbol{X}$ and the measurement noise $\boldsymbol{W}$.

The signal $\boldsymbol{N}$, which is independent of $\boldsymbol{X}$ and — when attempting to recover the features of $\boldsymbol{X}$ — can be considered as input noise, is distributed according to a complex GM; i.e.,

$$\boldsymbol{N} \sim p_{\boldsymbol{n}}(\boldsymbol{n}) = \sum\nolimits_{g=1}^{J_{\boldsymbol{n}}} r_g \sum\nolimits_{k=1}^{K} s_{g,k}\, \mathcal{CN}(\boldsymbol{n}; \boldsymbol{\mu}_{g,k}, \boldsymbol{\Gamma}_{g,k}) , \tag{2}$$

with mean vectors $\boldsymbol{\mu}_{g,k} \in \mathbb{C}^n$, covariance matrices $\boldsymbol{\Gamma}_{g,k} \in \mathbb{C}^{n \times n}$, and weights $s_{g,k}$ such that $\sum_{k=1}^{K} s_{g,k} = 1$. An instance of $\boldsymbol{N}$ is generated by one of $J_{\boldsymbol{n}}$ classes, which are each characterised by a GM with K components. The classes $g = 1, \dots, J_{\boldsymbol{n}}$ occur with probability r_g such that $\sum_{g=1}^{J_{\boldsymbol{n}}} r_g = 1$. The random vector $\boldsymbol{X}$ represents a signal of interest and is distributed as

$$\boldsymbol{X} \sim p_{\boldsymbol{x}}(\boldsymbol{x}) = \sum\nolimits_{c=1}^{J_{\boldsymbol{x}}} z_c \sum\nolimits_{o=1}^{O} \pi_{c,o}\, \mathcal{CN}(\boldsymbol{x}; \boldsymbol{\chi}_{c,o}, \boldsymbol{\Omega}_{c,o}) . \tag{3}$$

That is, the probability distributions of classes $c = 1, \dots, J_{\boldsymbol{x}}$ of $\boldsymbol{X}$ are each characterised by a GM with O components. The vector $\boldsymbol{W} \sim \mathcal{CN}(\boldsymbol{w}; \boldsymbol{\nu}, \boldsymbol{\Lambda})$ represents additive complex Gaussian noise with mean $\boldsymbol{\nu} \in \mathbb{C}^m$ and covariance $\boldsymbol{\Lambda} \in \mathbb{C}^{m \times m}$.

III. Optimisation Framework

Assume, for now, that — in addition to the distribution of our primary, always present source $\boldsymbol{X}$ — we have access to a predefined GM distribution for $\boldsymbol{N}$, which represents a secondary source that may or may not be present in the system at the time of measurement. As such, in the distribution for $\boldsymbol{N}$, the $J_{\boldsymbol{n}}$th class is characterised as $p_{\boldsymbol{n}|g}(\boldsymbol{n}|g = J_{\boldsymbol{n}}) = \sum_{k=1}^{K} (1/K)\, \mathcal{CN}(\boldsymbol{0}, \sigma \mathbf{I}_n)$ for some arbitrarily small σ; i.e., the value of $\boldsymbol{N}$ for this class is close to zero to represent the scenario in which the secondary source is absent.

We seek the matrix $\boldsymbol{\Phi}$ that maximises the objective function

$$F(\boldsymbol{\Phi}, \boldsymbol{\beta}) = \beta_1 I(\boldsymbol{X};\boldsymbol{Y}) + \beta_2 I(C;\boldsymbol{Y}) + \beta_3 I(\boldsymbol{N};\boldsymbol{Y}) + \beta_4 I(G;\boldsymbol{Y}), \tag{4}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4] \in \mathbb{R}^4$ controls the relative importance of the Shannon MI [13] terms and C and G are random variables that represent the classes of $\boldsymbol{X}$ and $\boldsymbol{N}$. For our purposes, we maintain $\|\boldsymbol{\beta}\|_1 = 1$.

We use the iterative gradient ascent algorithm of [8], [9] to identify the matrix $\boldsymbol{\Phi}$ that locally maximises $F(\boldsymbol{\Phi}, \boldsymbol{\beta})$ by setting $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} + \delta \nabla_{\boldsymbol{\Phi}} F(\boldsymbol{\Phi}, \boldsymbol{\beta})$ and normalising such that $\mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}}\} = m$ at each iteration. The step size $\delta > 0$ controls the rate of change of $\boldsymbol{\Phi}$. When computing $\nabla_{\boldsymbol{\Phi}} F(\boldsymbol{\Phi}, \boldsymbol{\beta})$, we evaluate the gradient terms given in [8], [9] via Monte Carlo (MC) integration and utilise the Bayesian inference model detailed in [9].

IV. Learning the Secondary Source Distribution

Here, we extend the work of [11] such that we are able to learn the distribution of $\boldsymbol{N}$ from compressive measurements. We assume that our compressive measurements have been captured using a block of data that contains instances of only one class of $\boldsymbol{N}$. Therefore, we omit the class parameter g to simplify notation. We first rewrite (1) as $\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{N} + \hat{\boldsymbol{W}}$, where

$$\hat{\boldsymbol{W}} \sim \sum\nolimits_{d=1}^{D} \tau_d\, \mathcal{CN}(\hat{\boldsymbol{w}}; \boldsymbol{\nu}_d, \boldsymbol{\Lambda}_d) , \tag{5}$$

$$\tau_d = z_{c'}\pi_{c',o'}, \quad \boldsymbol{\nu}_d = \boldsymbol{\Phi}\boldsymbol{\chi}_{c',o'} + \boldsymbol{\nu}, \quad \boldsymbol{\Lambda}_d = \boldsymbol{\Phi}\boldsymbol{\Omega}_{c',o'}\boldsymbol{\Phi}^{\mathrm{H}} + \boldsymbol{\Lambda} ,$$
$$D = J_{\boldsymbol{x}}O , \qquad c' = \left\lceil \tfrac{d}{O} \right\rceil , \qquad o' = ((d-1) \bmod O) + 1.$$

We seek the system parameters θ that maximise the log of the marginal probability; i.e., the incomplete log-likelihood:

$$\ell_{\text{inco}}(\theta|\boldsymbol{y}) = \log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta) = \log \sum_{k,d} \int p_{\boldsymbol{y},\boldsymbol{n},k,d|\theta}(\boldsymbol{y}, \boldsymbol{n}, k, d|\theta)\, d\boldsymbol{n}.$$

Since the log of a sum is not easily separable for maximisation purposes, we take a two-stage EM approach [12] and utilise the complete log-likelihood:

$$\ell_{\text{co}}(\theta|\boldsymbol{y}, \boldsymbol{n}, k, d) = \log p_{\boldsymbol{y},\boldsymbol{n},k,d|\theta}(\boldsymbol{y}, \boldsymbol{n}, k, d|\theta) . \tag{6}$$

Specifically, we consider the expected value of $\ell_{\text{co}}(\theta|\boldsymbol{y}, \boldsymbol{n}, k, d)$ under the posterior distribution of the latent variables $(\boldsymbol{n}, k, d)$. In the first stage of iteration $(t+1)$, we use the previous parameters $\theta^{(t)}$ to find the posterior distribution of the latent variables given by $p_{\boldsymbol{n},k,d|\boldsymbol{y},\theta}(\boldsymbol{n}, k, d|\boldsymbol{y}, \theta^{(t)})$. We then use this to find the expectation of the complete log-likelihood

$$\ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)}) = \mathop{\mathbb{E}}_{\boldsymbol{n},k,d|\boldsymbol{y},\theta^{(t)}} \Big[\log p_{\boldsymbol{y},\boldsymbol{n},k,d|\theta}(\boldsymbol{y}, \boldsymbol{n}, k, d|\theta) \Big]$$

with respect to this posterior. In the second stage, we determine the new parameters $\theta^{(t+1)}$ by maximizing $\ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)})$:

$$\theta^{(t+1)} = \arg\max_{\theta}\, \ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)}) . \tag{7}$$

Fortunately, we can show that maximising this function actually maximises the incomplete log-likelihood. We can write

$$p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta) = p_{\boldsymbol{y},\boldsymbol{n},k,d|\theta}(\boldsymbol{y}, \boldsymbol{n}, k, d|\theta) / p_{\boldsymbol{n},k,d|\boldsymbol{y},\theta}(\boldsymbol{n}, k, d|\boldsymbol{y}, \theta) . \tag{8}$$

By taking the expectation of the log of both sides with respect to $p_{\boldsymbol{n},k,d|\boldsymbol{y},\theta}(\boldsymbol{n}, k, d|\boldsymbol{y}, \theta^{(t)})$, we obtain

$$\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta) = \ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)}) + h(\theta|\boldsymbol{y}, \theta^{(t)}) , \tag{9}$$

where $h(\theta|\boldsymbol{y}, \theta^{(t)})$ is a conditional entropy term. The above holds for any θ, including $\theta = \theta^{(t)}$. That is,

$$\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta^{(t)}) = \ell_{\text{ex-co}}(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)}) + h(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)}) . \tag{10}$$

Subtracting this from (9), we obtain

$$\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta) - \log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta^{(t)}) =$$
$$\ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)}) + h(\theta|\boldsymbol{y}, \theta^{(t)}) - \ell_{\text{ex-co}}(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)}) - h(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)}).$$

By Gibbs' inequality [13], we know that for two probability distributions $p_1(\boldsymbol{y})$ and $p_2(\boldsymbol{y})$, we have

$$-\int p_1(\boldsymbol{y}) \log p_1(\boldsymbol{y})\, d\boldsymbol{y} \le -\int p_1(\boldsymbol{y}) \log p_2(\boldsymbol{y})\, d\boldsymbol{y} , \tag{11}$$

with equality only when $p_1(\boldsymbol{y}) = p_2(\boldsymbol{y})$. Thus, we have $h(\theta|\boldsymbol{y}, \theta^{(t)}) \ge h(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)})$ and

$$\begin{aligned} &\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta) - \log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta^{(t)}) \\ &\qquad \ge \ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)}) - \ell_{\text{ex-co}}(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)}) . \end{aligned} \tag{12}$$

That is, choosing θ such that $\ell_{\text{ex-co}}(\theta|\boldsymbol{y}, \theta^{(t)})$ improves upon $\ell_{\text{ex-co}}(\theta^{(t)}|\boldsymbol{y}, \theta^{(t)})$ guarantees that the resulting improvement from $\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta^{(t)})$ to $\log p_{\boldsymbol{y}|\theta}(\boldsymbol{y}|\theta)$ is at least as large.

When evaluating the posterior distribution of the latent variables, we utilise the following Bayesian inference model, where we have omitted the reliance on $\theta^{(t)}$ for brevity:

$$p_{k,d|\boldsymbol{y}}(k,d|\boldsymbol{y}) = \frac{s_k\, \tau_d\, p_{\boldsymbol{y}|k,d}(\boldsymbol{y}|k,d)}{p_{\boldsymbol{y}}(\boldsymbol{y})}\,, \tag{13}$$

$$p_{\boldsymbol{y}|k,d}(\boldsymbol{y}|k,d) = \mathcal{CN}(\boldsymbol{y}; \boldsymbol{\Phi}\boldsymbol{\mu}_k + \boldsymbol{\nu}_d, \boldsymbol{\Phi}\boldsymbol{\Gamma}_k\boldsymbol{\Phi}^{\mathrm{H}} + \boldsymbol{\Lambda}_d)\,, \tag{14}$$

$$p_{\boldsymbol{n}|\boldsymbol{y},k,d}(\boldsymbol{n}|\boldsymbol{y},k,d) = \mathcal{CN}(\boldsymbol{n}; \tilde{\boldsymbol{\mu}}_{k,d}, \mathbf{C}_{k,d})\,, \tag{15}$$

$$\mathbf{C}_{k,d} = \left(\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Lambda}_d^{-1}\boldsymbol{\Phi} + \boldsymbol{\Gamma}_k^{-1}\right)^{-1}\,, \tag{16}$$

$$\tilde{\boldsymbol{\mu}}_{k,d} = \boldsymbol{\mu}_k + \mathbf{C}_{k,d}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Lambda}_d^{-1}\left(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}_k - \boldsymbol{\nu}_d\right). \tag{17}$$

In actuality, we learn the distribution using a set of samples

$$\{\boldsymbol{y}_i = \boldsymbol{\Phi}_i\boldsymbol{n}_i + \hat{\boldsymbol{w}}_i\}\,, \quad i = 1, \ldots, N_s\,. \tag{18}$$

Here, if we consider each $\boldsymbol{\Phi}_i$ as the 'window' through which we observe $\boldsymbol{n}_i$, having a unique projection matrix for each sample allows us to more fully observe the characteristics of the distribution of $\boldsymbol{N}$. However, we will show later that this is not always necessary. For our N_s measurements, we have

$$\ell_{\text{ex-co}}(\theta|\theta^{(t)}) = \sum_{i=1}^{N_s} \mathop{\mathbb{E}}_{\boldsymbol{n},k,d|\boldsymbol{y}_i,\theta^{(t)}}\left[\log p_{\boldsymbol{y},\boldsymbol{n},k,d|\theta}(\boldsymbol{y}_i, \boldsymbol{n}, k, d|\theta)\right].$$

After a number of operations, we are able to expand this expression to obtain (19). Here, $\mathbf{C}_{k,d}^{(i)}$ and $\tilde{\boldsymbol{\mu}}_{k,d}^{(i)}$ are the per-sample equivalents of (16) and (17), respectively, and we have again omitted the reliance on $\theta^{(t)}$. If we set the expressions for the gradient of (19) with respect to s_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Gamma}_k$ to zero — noting that $\sum_k s_k = 1$ — we obtain

$$s_k^{(t+1)} = \frac{\sum_{i=1}^{N_s} p_{k|\boldsymbol{y}}(k|\boldsymbol{y}_i)}{\sum_{i=1}^{N_s}\sum_{k'=1}^{K} p_{k|\boldsymbol{y}}(k'|\boldsymbol{y}_i)} = \frac{\sum_{i=1}^{N_s} p_{k|\boldsymbol{y}}(k|\boldsymbol{y}_i)}{N_s}\,, \tag{20}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N_s}\sum_{d=1}^{D} p_{k,d|\boldsymbol{y}}(k,d|\boldsymbol{y}_i)\tilde{\boldsymbol{\mu}}_{k,d}^{(i)}}{\sum_{i=1}^{N_s} p_{k|\boldsymbol{y}}(k|\boldsymbol{y}_i)}\,, \tag{21}$$

$$\boldsymbol{\Gamma}_k^{(t+1)} = \tag{22}$$

$$\frac{\sum_{i=1}^{N_s}\sum_{d=1}^{D} p_{k,d|\boldsymbol{y}}(k,d|\boldsymbol{y}_i)\left[(\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\mu}_k)(\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\mu}_k)^{\mathrm{H}} + \mathbf{C}_{k,d}^{(i)}\right]}{\sum_{i=1}^{N_s} p_{k|\boldsymbol{y}}(k|\boldsymbol{y}_i)}.$$

Thus, we are able to iteratively move towards parameters for $\boldsymbol{N}$ that better fit our data. The approach ceases after a pre-determined number of iterations or if $\ell_{\text{inco}}(\theta^{(t+1)}|\boldsymbol{y}) - \ell_{\text{inco}}(\theta^{(t)}|\boldsymbol{y})$ falls below a specified threshold.

V. Adaptive Information-Theoretic Algorithm

We combine the pre-existing projection design methodology of Sec. III with our proposed distribution learning approach. A pseudocode representation of the resulting framework is provided in Algorithm 1. The algorithm initialises with a trained $\boldsymbol{\Phi}_{\text{opt}} \in \mathbb{C}^{m_{\text{opt}}\times n}$, which has been designed subject to the objective function of (4) and some prior knowledge of the system parameters θ_{opt}. If there is initially no knowledge of a secondary source, $\boldsymbol{N}$ is considered absent with only one class as defined in Sec. III.

The algorithm captures compressive samples at regular intervals. If the average log-likelihood of the past N_{avg} compressive samples falls below a predefined threshold ζ_{avg}, the algorithm begins a compressive sampling process using random matrices $\boldsymbol{\Phi}_{i_{\text{CS}}} \in \mathbb{C}^{m\times n}$ with elements drawn from $\mathcal{CN}(0,1)$. Each matrix $\boldsymbol{\Phi}_{i_{\text{CS}}}$ is reused N_{rep} times. If the number of consecutive samples with low average log-likelihood exceeds a predefined minimum $N_s^{\min}$, up to $N_s^{\max}$ of these samples will be used to learn the GM distribution of a new class of $\boldsymbol{N}$. This learning process will occur according to the iterative approach described in Sec. IV and will cease after a predefined number of iterations or if the change in log-likelihood across iterations falls below a predefined threshold. When updating the current system parameters θ with the new class of $\boldsymbol{N}$, the class probabilities for $\boldsymbol{N}$ are updated to match their likelihoods in the previous N_T samples:

$$r_g \leftarrow \frac{1}{N_T}\sum_{i'=i-N_T+1}^{i} p_{g|\boldsymbol{y}}(g|\boldsymbol{y}_{i'})\,. \tag{23}$$

Here, N_T is large and defined by the user beforehand. To avoid the retraining of a class with very low likelihood in the future, a threshold ξ can be placed on the class probabilities such that if $r_{g'} < \xi$ for some g', we set $r_{g'} = \xi$ and, for $g \neq g'$,

$$r_g \leftarrow (1-\xi)\, r_g \cdot \left(\textstyle\sum_{g\neq g'} r_g\right)^{-1}. \tag{24}$$

Samples with an average log-likelihood above the threshold are used for classification and signal reconstruction purposes according to the inference model in [9]. If the number of samples taken is a multiple of N_T and the parameters have changed since the last execution of the projection design step, we redesign $\boldsymbol{\Phi}_{\text{opt}}$ subject to the objective function of (4).

Note that, if desired, the random matrices $\boldsymbol{\Phi}_{i_{\text{CS}}}$ can be of a different dimensionality to $\boldsymbol{\Phi}_{\text{opt}}$; i.e., we can have $m_{\text{opt}} \neq m$. Using a high m to generate random projections will provide more information about the statistics of $\boldsymbol{N}$ during the distribution training step; it might therefore be possible to decrease the minimum number of random samples $N_s^{\min}$ that are required to obtain a good estimate of the distribution. However, a low m_{opt} might be sufficient for reconstruction or classification purposes. Using $m \neq m_{\text{opt}}$ will, of course, require an additional (assumed known) distribution for the measurement noise $\boldsymbol{W}_{\text{opt}} \sim \mathcal{CN}(\boldsymbol{w}_{\text{opt}}; \boldsymbol{\nu}_{\text{opt}}, \boldsymbol{\Lambda}_{\text{opt}})$.

VI. Experimental Results

A. Experiments with Synthetic Data

In the following simulations, we use known distributions to generate instances of $\boldsymbol{X}$, $\boldsymbol{N}$, and $\boldsymbol{W}$. We then attempt to recover the distribution of $\boldsymbol{N}$ using compressive measurements under various conditions. Initially, for simplicity, we constrain the number of classes of $\boldsymbol{X}$ and $\boldsymbol{N}$ to $J_{\boldsymbol{x}} = J_{\boldsymbol{n}} = 1$, and consider both sources active at all times. The GM distributions are limited to $O = K = 3$ components. We use $N_s = 1000$ random measurements and 500 iterations during the training of the GM for $\boldsymbol{N}$. Training ceases if the change in the incomplete log-likelihood between iterations drops below a value of one. Our input dimensionality is $n = 32$ and our compressive measurements are of dimensionality $m \in \{4, 8, 12\}$.

$$\ell_{\text{ex}-\text{co}}(\theta|\theta^{(t)}) = \text{constant} - \sum_{i=1}^{N_s}\Bigg\{ \underset{d|\boldsymbol{y}_i}{\mathbb{E}}\Big[\log\det\boldsymbol{\Lambda}_d\Big] + \underset{k|\boldsymbol{y}_i}{\mathbb{E}}\Big[\log\det\boldsymbol{\Gamma}_k\Big] + \underset{k,d|\boldsymbol{y}_i}{\mathbb{E}}\Big[\text{tr}\Big\{\boldsymbol{\Lambda}_d^{-1}\boldsymbol{\Phi}_i\mathbf{C}_{k,d}^{(i)}\boldsymbol{\Phi}_i^{\text{H}}\Big\}\Big] - \underset{k|\boldsymbol{y}_i}{\mathbb{E}}\Big[\log s_k\Big] - \underset{d|\boldsymbol{y}_i}{\mathbb{E}}\Big[\log\tau_d\Big]$$

$$+ \underset{k,d|\boldsymbol{y}_i}{\mathbb{E}}\Big[(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\nu}_d)^{\text{H}}\boldsymbol{\Lambda}_d^{-1}(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\nu}_d)\Big] + \underset{k,d|\boldsymbol{y}_i}{\mathbb{E}}\Big[\text{tr}\Big\{\boldsymbol{\Gamma}_k^{-1}\mathbf{C}_{k,d}^{(i)}\Big\}\Big] + \underset{k,d|\boldsymbol{y}_i}{\mathbb{E}}\Big[(\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\mu}_k)^{\text{H}}\boldsymbol{\Gamma}_k^{-1}(\tilde{\boldsymbol{\mu}}_{k,d}^{(i)} - \boldsymbol{\mu}_k)\Big]\Bigg\} \quad (19)$$

```
Find the $\boldsymbol{\Phi}_{\text{opt}} \in \mathbb{C}^{m_{\text{opt}} \times n}$ that maximises (4)
$i \leftarrow 0$, $i_{\text{CS}} \leftarrow 0$, $\theta \leftarrow \theta_{\text{opt}}$
**repeat**
    $i \leftarrow i+1$, $j \leftarrow 0$
    Store new sample $\boldsymbol{y}_i \leftarrow \boldsymbol{\Phi}_{\text{opt}}(\boldsymbol{x}_i + \boldsymbol{n}_i) + \boldsymbol{w}_i^{\text{opt}}$
    $\zeta_i^{\text{avg}}$ is the average of $N_{\text{avg}}$ last $\zeta_i \leftarrow \log p_{\mathbf{y}|\theta}(\boldsymbol{y}_i|\theta)$
    **if** $\zeta_i^{\text{avg}} < \zeta_{\text{thr}}$ **then** $j \leftarrow 1$
    **if** $j = 1$ **and** $i_{\text{CS}} < N_s^{\max}$ **then**
        $i_{\text{CS}} \leftarrow i_{\text{CS}} + 1$
        **if** $\text{mod}(i_{\text{CS}} - 1, N_{\text{rep}}) = 0$ **then**
            Generate and store random $\boldsymbol{\Phi}_{i_{\text{CS}}} \in \mathbb{C}^{m \times n}$
        **else** $\boldsymbol{\Phi}_{i_{\text{CS}}} \leftarrow \boldsymbol{\Phi}_{i_{\text{CS}}-1}$
        Store random sample $\tilde{\boldsymbol{y}}_{i_{\text{CS}}} \leftarrow \boldsymbol{\Phi}_{i_{\text{CS}}}(\boldsymbol{x}_i + \boldsymbol{n}_i) + \boldsymbol{w}_i$
    **else if** $i_{\text{CS}} > N_s^{\min}$ **then**
        Learn distribution of new class of $\boldsymbol{N}$ using
          random measurements and projection matrices
        Update distribution parameters $\theta$, set $i_{\text{CS}} \leftarrow 0$
    **else**
        Reconstruct/classify $\boldsymbol{x}_i$ and $\boldsymbol{n}_i$, set $i_{\text{CS}} \leftarrow 0$
    **if** $\text{mod}(i-1, N_T) = 0$ **and** $\theta \neq \theta_{\text{opt}}$ **then**
        $\theta_{\text{opt}} \leftarrow \theta$, find the $\boldsymbol{\Phi}_{\text{opt}}$ that maximises (4)
```

Algorithm 1: Adaptive information-theoretic algorithm.

The weights $\pi_{c,o}$ and $s_{g,k}$ are drawn from the standard uniform distribution and normalised. The mean vectors $\boldsymbol{\chi}_{c,o}$ and $\boldsymbol{\mu}_{g,k}$ comprise elements drawn from the complex Gaussian distribution $\mathcal{CN}(0, \sqrt{2}/10)$, and the covariance matrices $\boldsymbol{\Omega}_{c,o}$ and $\boldsymbol{\Gamma}_{g,k}$ are initially equal to instances of the product $\mathbf{QDQ}^{\text{H}}$, where $\mathbf{Q} \in \mathbb{C}^{n \times n}$ is a random unitary matrix and $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with elements drawn from the uniform distribution $\mathcal{U}(10^{-6}, 10^{-2})$. To vary the signal-to-noise ratio (SNR), we adjust the values in the diagonal elements of the matrices $\mathbf{D}$ used to generate $\boldsymbol{\Omega}_{c,o}$. Samples of measurement noise are drawn according to $\boldsymbol{W} \sim \mathcal{CN}(\boldsymbol{w}; \mathbf{0}, 10^{-6}\mathbf{I}_m)$ and $\boldsymbol{W}_{\text{opt}} \sim \mathcal{CN}(\boldsymbol{w}_{\text{opt}}; \mathbf{0}, 10^{-6}\mathbf{I}_{m_{\text{opt}}})$. For now, elements of $\boldsymbol{\Phi}_{\text{opt}}$ are drawn from $\mathcal{CN}(0, 1)$. Results are averaged over 100 instances of the simulation scenario. We use the ground truth parameters for $\boldsymbol{N}$ to generate results for comparison purposes.

Generating and storing a unique random projection matrix for each sample used for the training of the distribution of $\boldsymbol{N}$ provides the best possible insight into the source statistics. However, it may be possible to decrease computational costs and memory requirements by reusing the same projection matrix for multiple samples. In Fig. 1, we show that by decreasing the number of unique projection matrices, we increase the resulting mean-square error of reconstruction for $\boldsymbol{N}$ obtained via the inference model of [9]; i.e., our estimate of the distribution of $\boldsymbol{N}$ becomes increasingly inaccurate. However, it is clear that the reliance of the training process on unique projections depends on m, with smaller m generally requiring more unique matrices to achieve the best-case results. For a large number of measurements, e.g., for $m = 12$, we see that there are no significant disadvantages to using only 1% of the available unique matrices. We can also observe that by increasing the number of measurements m_{opt} that we use for signal recovery, we decrease the reconstruction error.

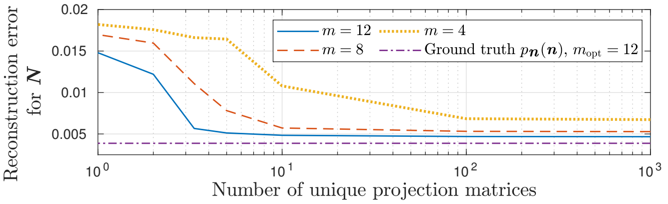

Fig. 1. Mean-square reconstruction error for $\boldsymbol{N}$ versus the number of unique random projection matrices for $m \in \{4, 8, 12\}$ and $m_{\text{opt}} = m$.

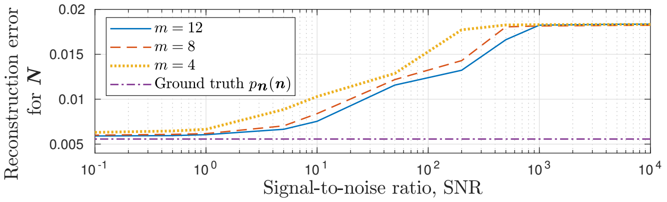

Fig. 2. Mean-square reconstruction error for $\boldsymbol{N}$ versus SNR from (25) when training the distribution of $\boldsymbol{N}$ for $m \in \{4, 8, 12\}$ and $m_{\text{opt}} = 4$.

The SNR — i.e., the ratio of the power in $\boldsymbol{X}$ to the power in $\boldsymbol{N}$ — will impact our ability to learn the distribution of $\boldsymbol{N}$. For the results of Fig. 2, we have used various values of SNR during the training of the distribution of $\boldsymbol{N}$, with

$$\text{SNR} = \text{tr}\{\boldsymbol{\Omega}_{\text{avg}}\} / \text{tr}\{\boldsymbol{\Gamma}_{\text{avg}}\} \; . \quad (25)$$

Here, $\boldsymbol{\Omega}_{\text{avg}}$ and $\boldsymbol{\Gamma}_{\text{avg}}$ are the average covariance matrices for the ground truth GM distributions of $\boldsymbol{X}$ and $\boldsymbol{N}$, respectively. All estimated distributions experienced the same test conditions; i.e., we have attempted to reconstruct $\boldsymbol{N}$ in a scenario with $\boldsymbol{X}$ and $\boldsymbol{N}$ of equal power and $m_{\text{opt}} = 4$. The resulting reconstruction error illustrates the quality of each estimate of the distribution of $\boldsymbol{N}$. We observe that increasing the number of random measurements m improves the distribution estimation in low SNR scenarios. Furthermore, we can see that for very high or very low SNR, all m perform similarly. Significantly, we see that we are unable to estimate the distribution for SNRs of order 10^3 and above, as the reconstruction error is no longer increasing; i.e., our estimates cannot become worse.

B. Experiments with Real Radar Data

Real radar returns from two fixed-location, three-bladed fans were acquired according to the setup in [9]. The fans possessed three rotation speeds, which can be seen in Table I. Acquisitions for each speed were downsampled to 5.5 kHz. Fans 1 and 2 contribute to the primary and secondary sources, $\boldsymbol{X}$ and $\boldsymbol{N}$, respectively. A time series $\mathbf{r}$ is the vectorised output

TABLE I
CLASS DESCRIPTIONS BASED ON FAN SPEEDS IN ROTATIONS PER SECOND

Fan	Input	Class 1	Class 2	Class 3	Class 4
1	$\boldsymbol{X}$	2.63 rps	4.10 rps	5.06 rps	N/A
2	$\boldsymbol{N}$	5.68 rps	6.21 rps	6.78 rps	Absent

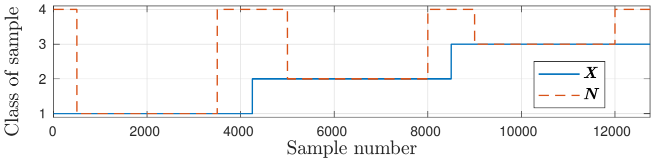

Fig. 3. Ground truth classes of $\boldsymbol{X}$ and $\boldsymbol{N}$ in the examined radar returns.

of the radar receiver system; $\mathbf{r}$ is split into R non-overlapping 'frames', which are segmented into B overlapping 'bursts'. Each burst is Hamming windowed and transformed to the frequency domain via the discrete Fourier transform.

We obtain ground truth distributions for the GMMs of $\boldsymbol{X}, \boldsymbol{N} \in \mathbb{C}^n$ via the EM algorithm [12]. For this, instances of $\boldsymbol{X}$ and $\boldsymbol{N}$ are obtained from transformed bursts when $\mathbf{r}$ contains radar returns from either source in isolation. Training data for this is obtained from 50 frames of data recorded for each fan speed. For feature extraction purposes, we use a frame length of 700. As in [9], we limit the number of frequency coefficients (and therefore the dimensionalities of $\boldsymbol{X}, \boldsymbol{N} \in \mathbb{C}^n$) to $n = 32$. Each burst overlaps its neighbours by 75%.

We initialise Algorithm 1 by designing the matrix $\boldsymbol{\Phi}_{\text{opt}}$ such that $I(C;\boldsymbol{Y})$ is maximised; i.e., we use $\boldsymbol{\beta} = [0,1,0,0]$ in (4). Matrix $\boldsymbol{\Phi}_{\text{opt}}$ is designed over 10^3 iterations using a step size of $\delta = 0.01$ and 500 MC draws to evaluate $\nabla_{\boldsymbol{\Phi}} F(\boldsymbol{\Phi}, \boldsymbol{\beta})$. Our initial system parameters θ_{opt} include $\boldsymbol{W} \sim \mathcal{CN}(\mathbf{0}, 10^{-6}\mathbf{I}_m)$, $\boldsymbol{W}_{\text{opt}} \sim \mathcal{CN}(\mathbf{0}, 10^{-6}\mathbf{I}_{m_{\text{opt}}})$, and the GMM for $\boldsymbol{X}$ with class probabilities $z_c = 1/3 \ \forall \ c$ and $O = 3$ components. We assign $K = 3$ components to each class of $\boldsymbol{N}$ and begin with $\boldsymbol{N} \sim p_{\boldsymbol{n}}^{\text{init}}(\boldsymbol{n}) = \sum_{k=1}^{K} (1/K)\mathcal{CN}(\mathbf{0}, 10^{-6}\mathbf{I}_n)$. The algorithm is applied to a sequence of radar return data of length $N_T = 12750$ in which the ground truth classes of $\boldsymbol{X}$ and $\boldsymbol{N}$ from Table I are changing according to Fig. 3. Note that Fan 2 is absent for class 4 of $\boldsymbol{N}$. We use the following parameters: $m \in \{4, 6, 8, 12, 16\}$, $m_{\text{opt}} \in \{4, 6, 8\}$, $N_{\text{avg}} = 100$, $N_{\text{rep}} = 1$, $N_s^{\max} = 3000$, $N_s^{\min} = 1000$, and $\zeta_{\text{thr}} = 2.5$. When learning GMs for $\boldsymbol{N}$, we use the parameters of Sec. VI-A.

Fig. 4 shows the burst classification accuracies for $\boldsymbol{X}$ obtained after applying Algorithm 1 to the examined radar returns. Note that the retraining of $\boldsymbol{\Phi}_{\text{opt}}$ again used $\boldsymbol{\beta} = [0,1,0,0]$. Clearly, increasing m has improved our ability to classify $\boldsymbol{X}$. This indicates that, as in Sec. VI-A, a large m provides a better estimate of $p_{\boldsymbol{n}}(\boldsymbol{n})$. With a better estimate, we are able to obtain a superior $\boldsymbol{\Phi}_{\text{opt}}$. As in [9], increasing m_{opt} also improves our classification accuracy. Note that classifying on a per-frame basis [9] can further increase performance.

VII. CONCLUSIONS

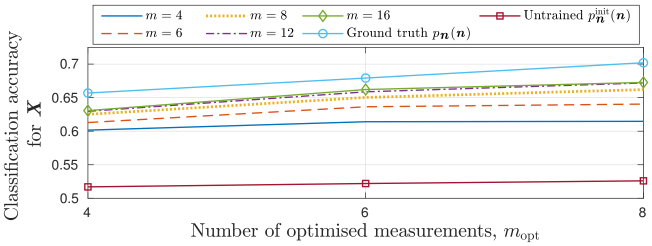

Fig. 4. Burst classification accuracy for $\boldsymbol{X}$ versus the number of optimised measurements m_{opt} for $m \in \{4, 6, 8, 12, 16\}$ random measurements.

In this paper, we have derived a methodology for the training of the GM distribution of a secondary input via compressive measurements. We have shown that well-estimated distributions yield designed projection matrices that are more able to control the input-output MI of a system. Furthermore, we have demonstrated that increasing the number of compressive measurements aids the characterisation of weak secondary sources and can reduce the number of unique projection matrices required for distribution training purposes.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant number EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing. The authors would like to thank Prof. Mike Davies, Prof. Mathini Sellathurai, and Dr João Mota for their valuable input. Gratitude is also extended to Dr Domenico Gaglione, Dr Christos Ilioudis, and Dr Carmine Clemente for capturing the radar data.

REFERENCES

[1] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] J. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, July 2009.

[3] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1394–1407, Aug. 2007.

[4] L. Wang *et al.*, "Information-theoretic compressive measurement design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1150–1164, June 2017.

[5] G. Yu and G. Sapiro, "Statistical compressed sensing of Gaussian mixture models," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5842–5858, Dec. 2011.

[6] B. Paul, C. D. Chapman, A. R. Chiriyath, and D. W. Bliss, "Bridging mixture model estimation and information bounds using I-MMSE," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4821–4832, Sept. 2017.

[7] F. Renna, R. Calderbank, L. Carin, and M. Rodrigues, "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Trans. Signal Process.*, vol. 62, no. 9, pp. 2265–2277, May 2014.

[8] F. K. Coutts, J. Thompson, and B. Mulgrew, "Gradient of mutual information in linear vector Gaussian channels in the presence of input noise," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 2264–2268.

[9] ——, "Information-theoretic compressive measurement design for micro-Doppler signatures," in *Proc. Sens. Signal Process. Defence*, 2020, pp. 1–5.

[10] V. C. Chen, *The micro-Doppler effect in radar*, 1st ed. Norwood, MA: Artech House, 2011.

[11] J. Yang, X. Liao, X. Yuan, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Compressive sensing by learning a Gaussian mixture model from measurements," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 106–119, Jan 2015.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[13] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.