

# Data Acquisition through joint Compressive Sensing and Principal Component Analysis

Riccardo Masiero<sup>§</sup>, Giorgio Quer<sup>§</sup>, Daniele Munaretto<sup>\*</sup>, Michele Rossi<sup>§</sup>, Joerg Widmer<sup>\*</sup>, Michele Zorzi<sup>§</sup>

**Abstract**—In this paper we look at the problem of accurately reconstructing distributed signals through the collection of a small number of samples at a data gathering point. The techniques that we exploit to do so are Compressive Sensing (CS) and Principal Component Analysis (PCA). PCA is used to find transformations that sparsify the signal, which are required for CS to retrieve, with good approximation, the original signal from a small number of samples. Our approach dynamically adapts to non-stationary real world signals through the online estimation of their correlation properties in space and time; these are then exploited by PCA to derive the transformations for CS. The approach is tunable and robust, independent of the specific routing protocol in use and able to substantially outperform standard data collection schemes. The effectiveness of our recovery algorithm, in terms of number of transmissions in the network vs reconstruction error, is demonstrated for synthetic as well as for real world signals which we gathered from an actual wireless sensor network (WSN) deployment. We stress that our solution is not limited to WSNs but it can be readily applied to other types of network infrastructures that require the online approximation of large and distributed data sets.

## I. INTRODUCTION

In this paper we look at the problem of efficiently gathering large amounts of data in Wireless Sensor Networks (WSNs). Our objective is to measure large data sets with high accuracy through the collection of a small number of readings. This entails the design of distributed algorithms for the joint gathering and compression of data and the exploitation, at the sink, of signal processing techniques for the approximation of the signal in space and time.

In our previous paper [1] we targeted this objective through a joint routing and compression scheme based on Compressive Sensing (CS), a technique that effectively exploits the correlation among sensor readings for the recovery of the original signal. The main problems of the approach in [1] are: 1) the scheme needs perfect knowledge of a transformation that makes the signal sparse in some domain, 2) the scheme is not able to adapt to non-stationary signals, i.e., with a time varying correlation structure and 3) the scheme was only designed for grid topologies. In this paper we solve all these issues through the combination of CS with Principal Component Analysis (PCA), a technique that exploits the online estimation of signal statistics. PCA allows to dynamically learn the optimal transformation to be used by CS recovery, effectively accounting for the time varying correlation affecting real signals. While PCA has been used primarily as a measurement basis [2], we exploit it as a transformation basis instead.

The main contributions of this paper are:

- The combination of PCA and CS techniques for the online estimation of signal statistics.
- The design of a technique which iteratively learns optimal transformations through the online estimation of the signal correlation structure.
- The design of a simple protocol, based on the above technique, for the online recovery of large data sets through the collection of a small number of readings. With our scheme the correlation structure of the signal is only estimated and exploited at the sink, whereas data gathering and routing are independent of it.
- We prove the effectiveness of our approach for data gathering and recovery for signals measured from an actual WSN deployment.

The problem of gathering data while jointly performing compression has been receiving increasing attention. One of the first studies addressing this issue is [3], which exploits classical source coding (see e.g., [4], [5]), suitable routing algorithms and re-encoding of information at relay nodes. Along the same lines, subsequent work such as [6], [7] proposes algorithms involving the collaboration among sensors in order to implement classical source coding in a distributed fashion. Further, [8] investigates the relation between routing and location of the aggregation/compression points according to the joint correlation of data among sources. However, in these approaches the required collaboration among nodes impacts the WSN performance in terms of number of transmissions in the network and complexity of the application running on the sensor nodes.

One of the earliest studies that exploit CS in a distributed communication scheme is [9], which targets the energy efficient estimation of sensed data in a WSN. Multi-hop communication and in-network data processing are not considered. Instead, data packets are directly transmitted by each node to the sink. This however requires synchronization among nodes. [10] is a further study exploiting CS. The considered simulation scenario is a network where a small set of nodes fails. The goal is to correctly identify these nodes through the transmission of random projections (i.e., linear combinations) indicating the status of the nodes. However, these random projections are obtained by means of a pre-distribution phase (via simple gossiping algorithms), which is very expensive in terms of number of transmissions. [11] also addresses the problem of gathering data in distributed WSNs through multi-hop routing. The authors of this paper, however, do not investigate the impact of the network topology and that of the routing scheme

<sup>§</sup>DEL, University of Padova, via Gradenigo 6/B – 35131, Padova, Italy.

<sup>\*</sup>DOCOMO Euro-Labs, Landsberger Strasse 312 – 80687, Munich, Germany.

on the compression process. [12] advocates the need to exploit the correlation of the data both temporally and spatially. The projections of the signal measurements are performed at each source node, taking into account the temporal correlation. The spatial correlation is instead exploited at the sink by means of decoders based on sparsity models which aim at describing the different types of signals of interest. We note that all these approaches for signal recovery are not suitable for use in combination with lightweight routing protocols. In contrast, our technique separates routing and recovery so that computationally demanding algorithms are only executed at the sink. This makes it possible to couple our approach with any routing scheme.

The related paper [2] focuses on image recovery and compares classical CS recovery assuming random projections against an alternative method, where the projections are obtained through PCA. Our objective is very different as we use these two techniques in combination, by exploiting PCA to obtain good sparsification bases for the signal and CS to recover the signal given these bases.

The paper is structured as follows. In Section II we give a mathematical overview of the CS and PCA techniques, which is followed by the description of our framework. In Sections III and IV, through comparison with respect to standard approaches, we prove the effectiveness of our combined CS and PCA signal reconstruction technique for synthetic and real signals, respectively. Section V concludes the paper.

## II. JOINT PCA AND CS RECOVERY

In this section we first review basic tools from Principal Component Analysis and Compressive Sensing, which we subsequently combine to achieve an efficient technique for the recovery of large signals from a small subset of measurements.

**Principal Component Analysis [13]:** the Karhunen-Loève expansion is the theoretical basis for PCA. It is a method to represent a generic  $N$ -dimensional signal given that we have full knowledge of its correlation structure. In detail, the signal can be well approximated through a small number of coefficients according to a given basis, which in turn depends on the correlation matrix of the signal itself. In practical cases, such as the ones we are concerned with in this paper, this correlation matrix may not be known a priori. Nevertheless, the Karhunen-Loève expansion can still be achieved thanks to the Principal Component Analysis (PCA) [13], which relies on the online estimation of the signal correlation matrix. In what follows, we describe the PCA along with its practical application to our data gathering problem. The key point of PCA is the Ky Fan theorem.

*Ky Fan theorem [14]:* let  $\Sigma \in \mathbb{R}^{N \times N}$  be a symmetric matrix, let  $\lambda_1 \geq \dots \geq \lambda_N$  be its eigenvalues and  $\mathbf{u}_1, \dots, \mathbf{u}_N$  the corresponding eigenvectors (which are assumed to be orthonormal, without loss of generality). Given  $M$  orthonormal vectors  $\mathbf{b}_1, \dots, \mathbf{b}_M$  in  $\mathbb{R}^N$ , with  $M \leq N$ , it holds that

$$\max_{\mathbf{b}_1, \dots, \mathbf{b}_M} \sum_{j=1}^M \mathbf{b}_j^T \Sigma \mathbf{b}_j = \sum_{j=1}^M \lambda_j, \quad (1)$$

and the maximum is attained for  $\mathbf{b}_i = \mathbf{u}_i, \forall i$ .

*Compression through PCA:* let  $\mathbf{x}_k \in \mathbb{R}^N$  be the vector of measurements from our WSN at a given time instant  $k$ , where the network consists of  $N$  nodes. We collect measurements according to a fixed sampling rate at discrete times  $k = 1, 2, \dots, K$ . From a geometrical point of view, we consider  $\mathbf{x}_k$  as a point in  $\mathbb{R}^N$  and look for the  $M$ -dimensional plane (with  $M \ll N$ ) which best matches the points in  $\mathbf{x}_k$  in terms of minimum Euclidean distance. The sample mean vector  $\bar{\mathbf{x}}$  and the sample covariance matrix  $\hat{\Sigma}$  of  $\mathbf{x}_k$  are given as:

$$\bar{\mathbf{x}} = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_k, \quad \hat{\Sigma} = \frac{1}{K} \sum_{k=1}^K (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T.$$

Note that in the Ky Fan theorem, maximizing  $\sum_{j=1}^M \mathbf{b}_j^T \Sigma \mathbf{b}_j$  corresponds to finding the linear transformation  $\mathcal{T}: \mathbb{R}^N \rightarrow \mathbb{R}^M$  that maximally preserves the information contained in the original signal  $\mathbf{x}_k \in \mathbb{R}^N$ . In fact, in the LHS of (1) we maximize the variance of the  $M$ -term (linear) approximation of  $\mathbf{x}_k$  that, in turn, is strictly related to the information content of the original signal. According to this rationale, we define  $\mathbf{U}_M$  as the matrix whose columns are the first  $M$  eigenvectors of  $\hat{\Sigma}$  (corresponding to the  $M$  largest eigenvalues). Due to the Ky Fan result, the projection of a given measurement  $\mathbf{x}_k$  into its best  $M$ -dimensional fitting plane is given by [13]:

$$\hat{\mathbf{x}}_k = \bar{\mathbf{x}} + \mathbf{U}_M \mathbf{U}_M^T (\mathbf{x}_k - \bar{\mathbf{x}}). \quad (2)$$

In (2),  $\mathbf{U}_M^T (\mathbf{x}_k - \bar{\mathbf{x}})$  is the projection of  $\mathbf{x}_k - \bar{\mathbf{x}}$  into its best fitting  $M$ -dimensional plane. If we define  $\mathbf{s} = \mathbf{s}^{(N)} \stackrel{def}{=} \mathbf{U}_N^T (\mathbf{x}_k - \bar{\mathbf{x}})$ , by construction of the projection matrix  $\mathbf{U}_N^T$  we have that the entries of  $\mathbf{s}$  are ordered as follows:  $s_1 \geq s_2 \geq \dots \geq s_N$ . If for  $i > M$  we have that  $s_i$  is negligible with respect to the previous entries of  $\mathbf{s}$ , i.e.,  $s_i \ll s_j$  with  $j = 1, \dots, M$ , then  $\mathbf{x}_k$  can be very well approximated through (2) by just accounting for  $M \ll N$  coefficients. In summary, the original point  $\mathbf{x}_k \in \mathbb{R}^N$  is transformed into a point  $\mathbf{s}^{(M)} \in \mathbb{R}^M$  as follows:

$$\mathbf{s}^{(M)} \stackrel{def}{=} \mathbf{U}_M^T (\mathbf{x}_k - \bar{\mathbf{x}}). \quad (3)$$

Multiplication of (3) by  $\mathbf{U}_M$  and summation with the sample mean return the best approximation of the original vector.

**Compressive Sensing (CS) [15]:** our goal is to recover a given  $N$ -dimensional signal through the reception of a small number of samples  $L$ , which should be ideally much smaller than  $N$ . CS is the technique that we exploit to achieve this objective.

As above, we consider signals representable through one dimensional vectors  $\mathbf{x} \in \mathbb{R}^N$ , containing the sensor readings of a WSN with  $N$  nodes. We further assume that these vectors are such that there exists a transformation under which they are sparse. In detail, there exists an invertible transformation matrix  $\Psi$  of size  $N \times N$  such that

$$\mathbf{x} = \Psi \mathbf{s} \quad (4)$$

and the  $N$ -dimensional vector  $\mathbf{s}$  is sparse. We say that  $\mathbf{s}$  is  $M$ -sparse if it has at most  $M$  non-zero entries, with  $M < N$ . Assuming to know  $\Psi$ ,  $\mathbf{x}$  can be recovered from  $\mathbf{s}$  by inverting (4). Also,  $\mathbf{s}$  can be obtained through a number  $L$  of random

projections of  $\mathbf{x}$ , namely  $\mathbf{y}$ , with  $M \leq L < N$ , according to the following equation:

$$\mathbf{y} = \Phi \mathbf{x}, \quad (5)$$

where  $\mathbf{y}$  is a vector of size  $L$  and  $\Phi$  is an  $L \times N$  matrix. In our framework,  $\Phi$  is referred to as *routing matrix* as it captures the way in which our sensor data is gathered and transmitted to the sink, which will receive the compressed vector  $\mathbf{y}$  along with the coefficients of matrix  $\Phi$ . Now, using (4) and (5) we can write

$$\mathbf{y} = \Phi \mathbf{x} = \Phi \Psi \mathbf{s} \stackrel{\text{def}}{=} \tilde{\Phi} \mathbf{s}. \quad (6)$$

In general this system is both ill-posed and ill-conditioned as the number of equations  $L$  is smaller than the number of variables  $N$  and small variations of the input signal can produce large variations of the output  $\mathbf{y}$ , respectively. However, if  $\mathbf{s}$  is sparse, it has been shown that (6) can be inverted with high probability through the use of special optimization techniques [16], [17]. These allow to retrieve  $\mathbf{s}$ , whereas the original signal  $\mathbf{x}$  is found through (4).

**Joint CS and PCA:** the main contribution of this paper is the design of a data recovery scheme combining CS and PCA. In particular, CS is exploited to solve the system in (6) after  $L$  data packets have been collected from the WSN and PCA is the technique providing the transformation matrix  $\Psi$ .

In [1] we used CS for the recovery of 2D real signals by considering different transformation matrices. None of them was however sufficiently good in terms of 1) sparsification and 2) incoherence with respect to  $\Phi$ . In summary, while the results<sup>1</sup> obtained for synthetic signals were very promising, those achieved for real signals were unsatisfactory. In this paper we solve this issue showing that the theoretical performance benefits of CS can still be retained if we use PCA to build the transformation matrix  $\Psi$ .

Our joint recovery through CS and PCA works as follows. Following the notation introduced in this section, at each time step  $k$  the sink first gathers  $L$  packets. These correspond to random projections of the input signal  $\mathbf{x}_k$  according to  $\mathbf{y} = \Phi \mathbf{x}_k$ . For the matrix  $\Phi$  we consider a random sampling (RS) scheme, as described in the following Section III.

According to the PCA framework, we can write the sparse vector  $\mathbf{s}_k = \mathbf{s}_k^{(N)}$  at time  $k$  as:

$$\mathbf{s}_k = \mathbf{U}_N^T (\mathbf{x}_k - \bar{\mathbf{x}}).$$

Note that if  $\mathbf{x}$  can be perfectly obtained using  $\mathbf{s}^{(M)}$  of (3) this means that  $\mathbf{s}$  is  $M$ -sparse since  $\mathbf{s} = \begin{bmatrix} \mathbf{s}^{(M)} \\ \mathbf{0}_{N-M} \end{bmatrix}$ , where  $\mathbf{0}_{N-M}$  is the zero column vector of size  $N - M$ .

Moreover, since  $\mathbf{U}_N$  is orthonormal we have that  $\mathbf{U}_N \mathbf{U}_N^T = \mathbf{I}_N$ , where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix. Hence, the above equation can be rewritten as:

$$\mathbf{x}_k - \bar{\mathbf{x}} = \mathbf{U}_N \mathbf{s}_k = \Psi \mathbf{s}_k. \quad (7)$$

where the transformation matrix  $\Psi$  is set equal to  $\mathbf{U}_N$ . Now, using  $\mathbf{y} = \Phi \mathbf{x}_k$  and (7), we can write:

$$\mathbf{y} - \Phi \bar{\mathbf{x}} = \Phi (\mathbf{x}_k - \bar{\mathbf{x}}) = \Phi \mathbf{U}_N \mathbf{s}_k,$$

<sup>1</sup>In terms of reconstruction error vs number of transmissions.

whose form is similar to that of (6) with  $\tilde{\Phi} = \Phi \mathbf{U}_N$ . The original signal  $\mathbf{x}_k$  is approximated as follows: 1) finding a good estimate of  $\mathbf{s}_k$ , namely  $\tilde{\mathbf{s}}_k$ , using the techniques in [16] or [17] and 2) applying the following calculation:

$$\tilde{\mathbf{x}}_k = \bar{\mathbf{x}} + \mathbf{U}_N \tilde{\mathbf{s}}_k.$$

As is clear from the above recovery procedure, our method takes as input the sample mean  $\bar{\mathbf{x}}$  and the covariance matrix  $\hat{\Sigma}$  (from which we obtain  $\mathbf{U}_N$ ). In what follows, we show that an online estimation of these parameters is possible, e.g., by alternating training and monitoring phases, and that this joint CS and PCA recovery leads to substantial performance improvements also for real world signals.

### III. ANALYSIS OF SIGNALS WITH A FIXED SUPPORT

In this section we study the effectiveness of joint CS and PCA recovery when applied to synthetic signals that are measured through the grid network model used in [1].

**Network:** we consider a square area of side  $H$  units, which is split into a grid of  $N$  square cells of side  $H/\sqrt{N}$ . We place each of the  $N$  nodes uniformly within the  $N$  cells so that each cell contains exactly one node. For the transmission range  $R$  of the nodes we adopt a unit disk model, i.e., they can only communicate with all other nodes placed at a distance less than or equal to  $R$ .<sup>2</sup> We set  $R = H\sqrt{5}/\sqrt{N}$  as this guarantees a fully connected structure. A further node, referred to as sink, is placed in the center of the deployment area. For the routing, we use a geographic forwarding technique, where each node considers as its next hop the node within range that provides the largest geographical advancement towards the sink. For our simulations we considered  $N = 400$ .

**Signals:** the input signal is a square matrix  $\mathbf{X}$  with  $N$  elements, where element  $(i, j)$  (referred to as  $x_{i,j}$ ) is the value sampled by the sensor placed in cell  $(i, j)$  of the grid. Here, we consider synthetic signals which are only spatially correlated, i.e., at each time  $k$  we generate a new signal  $\mathbf{X}$  according to the following procedure: 1) we start from a discrete 2D signal  $\mathbf{S}$  with a fixed support in the frequency domain (DCT).  $\mathbf{S}$  is low frequency, i.e., entries in position  $(p, q)$ , where  $p + q \leq \sqrt{N}/2 + 1$  are uniformly picked in  $[0.5, 1.5]$ , while other entries are zero. 2)  $\mathbf{X}$  is obtained from  $\mathbf{S}$  by DCT inversion, 3) in order to verify the robustness of our recovery schemes, we add to each  $x_{i,j}$  of  $\mathbf{X}$  an i.i.d. random gaussian noise component  $\omega_{i,j} \in \mathcal{N}(0, \sigma^2)$ .

Now, we define a  $\text{vec}(\cdot)$  function, transforming a  $\sqrt{N} \times \sqrt{N}$  matrix into a vector of length  $N$  (through a reordering of the matrix elements)

$$\text{vec}(\mathbf{X}) = (x_{1,1}, \dots, x_{k,1}, x_{1,2}, \dots, x_{k,2}, \dots, x_{1,k}, \dots, x_{k,k})^T$$

and we consider the vector representation of  $\mathbf{X}$ , i.e.,  $\mathbf{x} = \text{vec}(\mathbf{X})$ .  $\mathbf{x}$  can be obtained through a linear transformation of  $\mathbf{s} = \text{vec}(\mathbf{S})$  using tools from linear algebra. A more accurate description of this transformation can be found in [1].

<sup>2</sup>The unit disk graph model is used here for simplicity of explanation and topology representation. However, the presented methodology can be readily applied to more realistic propagation models, e.g., fading channels.

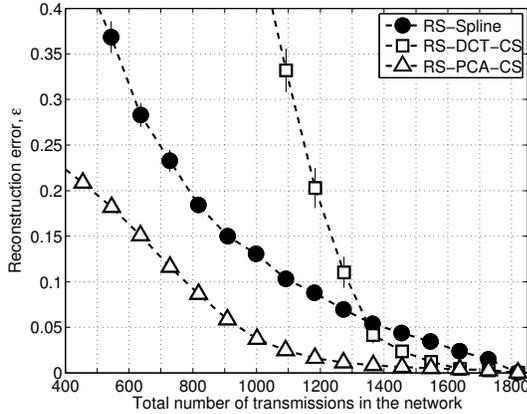


Fig. 1. Performance of three different recovery techniques for a synthetic low-pass signal: number of transmissions per data collection  $vs$   $\epsilon$ .

**Data gathering:** the *data collection* at the generic time  $k$  adopts a simple *random sampling* scheme as follows. Each node becomes a source with probability  $p = L/N$ , which was varied in the simulations to obtain tradeoff curves for increasing transmission overhead. Hence, on average  $L$  nodes transmit a packet containing their own sensor reading. Each packet is routed to the sink via geographical routing. The sink collects incoming data from all transmitting nodes according to  $\mathbf{y} = \Phi\mathbf{x}$ , where  $\mathbf{x}$  is the original signal and  $\Phi$  represents the routing matrix.  $\Phi$  has a single one in each row and at most a single one in each column. In detail, row  $i$  with  $1 \leq i \leq M$  has a one in column  $j$  with  $1 \leq j \leq N$  if the  $i$ -th packet was transmitted by node  $j$ . The cost of delivering a single packet to the sink is given by the number of hops that connect the source node to the sink.<sup>3</sup>

**Recovery:** we consider the following recovery techniques:

- R1. *Random sampling with Spline interpolation (RS-Spline):* the signal is reconstructed by spline interpolation [18] of the values collected through RS.
- R2. *Compressive Sensing (RS-DCT-CS):* we use the CS recovery technique described in Section II, where  $\Phi$  is the RS routing matrix defined above and  $\Psi$  implements the DCT transformation in two dimensions.
- R3. *Compressive Sensing with PCA (RS-PCA-CS):* the original signal is recovered through joint CS and PCA, as described in Section II. The sample mean  $\bar{\mathbf{x}}$  and the covariance matrix  $\hat{\Sigma}$  are calculated from a large enough number of instances of the synthetic signal so as to obtain accurate estimates of these statistics. The regularization parameters of the CS algorithms were set according to the default values suggested in [16], [17].

**Results:** to simplify the investigation and to pinpoint the fundamental performance tradeoffs, we assume a unit cost for each packet transmission. The metrics of interest are the total number of transmissions in the network for any given time  $k$  and the reconstruction quality at the sink, defined as  $\epsilon = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2 / \|\mathbf{x}\|_2$ , where  $\mathbf{x}$  is the original signal and  $\tilde{\mathbf{x}}$  is the signal reconstructed at the sink from the received

<sup>3</sup>Other cost metrics, e.g., energy, could also be used.



Fig. 2. Layout of the WSN testbed.

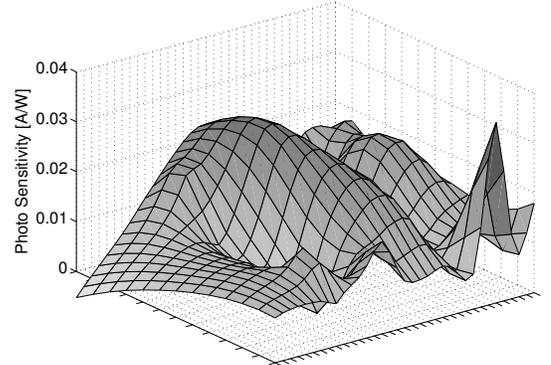


Fig. 3. Signal sample: luminosity in the range 320 – 730 nm.

samples  $\mathbf{y}$ . In Fig. 1 we compare the performance of the above recovery techniques in terms of  $\epsilon$   $vs$  total number of transmissions per data collection for a low-pass signal. RS-DCT-CS outperforms RS-Spline only when  $L$  approaches  $N$ , i.e., when the sink receives nearly all  $N$  packets and the total number of transmissions is close to the maximum (about 1800 for the considered network). In addition, the gain that RS-DCT-CS can provide is very small. Instead, RS-PCA-CS recovery significantly outperforms both RS-Spline and RS-DCT-CS for all values of  $L$  and allows the recovery of  $\tilde{\mathbf{x}}$  with small reconstruction errors. For example, an error requirement of  $\epsilon = 0.05$  is achieved in RS-PCA-CS with about 1000 transmissions, whereas RS-DCT-CS would need 50% more transmissions for the same error performance. We note that the performance of RS-Spline for high-frequency signals would be dramatically impacted, whereas RS-DCT-CS and RS-PCA-CS would undergo minimal performance impairments.

#### IV. ANALYSIS OF REAL SIGNALS FROM A WSN TESTBED

In this section we apply the joint CS and PCA recovery described above to the signals that we gathered from an actual WSN deployment.

**Network:** we consider the WSN testbed of Fig. 2.<sup>4</sup> This experimental network is deployed on the ground floor of the Department of Information Engineering at the University of Padova. The WSN consists of  $N = 68$  TmoteSky

<sup>4</sup>Our framework is flexible and does not depend on a specific topology; the only requirement is to bring the sensor nodes into an order, e.g., based on their IDs.

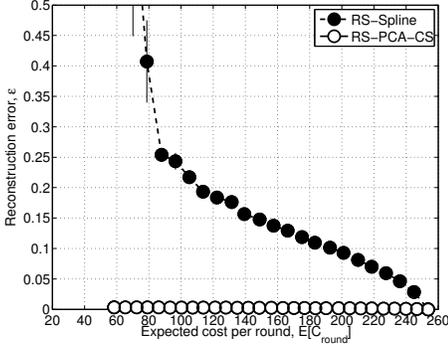


Fig. 4.  $\varepsilon$  vs  $E[C_{\text{round}}]$ : humidity.

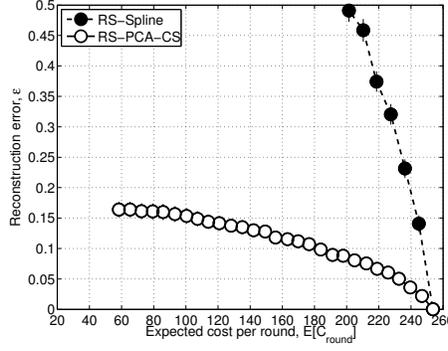


Fig. 5.  $\varepsilon$  vs  $E[C_{\text{round}}]$ : luminosity.

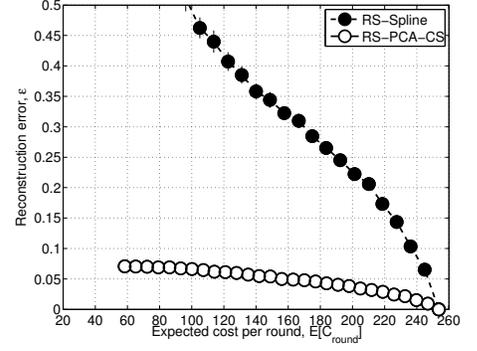


Fig. 6. Average  $\varepsilon$  (signals 1–5) vs  $E[C_{\text{round}}]$ .

wireless nodes equipped with IEEE 802.15.4 compliant radio transceivers.

**Signals:** From the above WSN, we gathered five different types of signals  $\mathbf{X}$ : 1) temperature, 2) humidity, 3) voltage, 4-5) luminosity in two different ranges (320–730 and 320–1100 nm, respectively), collecting measurements from all nodes every 5 minutes for 3 days. We repeated the data collection for three different measurement campaigns, choosing different days of the week. Fig. 3 shows an example signal of type 4, i.e., luminosity in the range 320 – 730 nm.

**Data gathering and Results:** to test the effectiveness of the proposed technique we considered the real data collected through the testbed in Fig. 2 and a data gathering scheme based on geographical routing. We placed the sink in the center of the network, where the signal is reconstructed at each time  $k$  based on our joint CS and PCA technique. Note that the signals in the testbed differ from those we generated in the previous section as they do not necessarily have the well-defined low-frequency representation that was assumed in Section III and are characterized by spatial and temporal correlations that are in general non-stationary. This means that the statistics that we use in our solution (i.e., sample mean and covariance matrix) must be learned at runtime and might not be valid throughout the entire data collection phase. Hence, in order to implement PCA in conjunction with CS for real signals, we alternate the following two phases:

1. a *training phase* of  $K$  data collection rounds, during which the sink collects the readings from all  $N$  sensors and uses this information to compute  $\bar{\mathbf{x}}$  and  $\hat{\Sigma}$ ;
2. a subsequent *monitoring phase* of  $\zeta K$  rounds during which, on average, only  $L \leq N$  nodes become sources according to the random sampling scheme of Section III (each with probability  $p = L/N$ ). The input signal is thus reconstructed using the statistics  $\bar{\mathbf{x}}$  and  $\hat{\Sigma}$  computed in the previous phase.

The ratio  $\zeta$  between the duration of monitoring and training phases should be chosen according to the temporal correlation of the observed phenomena.

In Figs. 4–6 we show the performance in terms of reconstruction error ( $\varepsilon$ ) as a function of the average cost per round, which is given by the number of transmissions for the

collection of a single instance of the signal  $\mathbf{X}_k$ . In these plots each training phase lasts  $K = 2$  rounds and  $\zeta = 4$  (the impact of these parameters is addressed at the end of this section). A training phase entails a cost  $K C_N$ , where  $C_N$  is the total number of transmissions needed to gather the readings from all nodes. The average number of packets sent during the following  $2\zeta = 8$  monitoring phases depends on  $p$ , which is varied from  $1/N$  to 1, and  $\varepsilon$  is computed for each case. For a given  $p = L/N$  each monitoring phase has a cost of  $\zeta K E[C_L]$ , where  $E[C_L]$  is the total number of transmissions needed to collect the readings from the source nodes during a data collection round. Thus, the average cost per round is calculated as:

$$E[C_{\text{round}}] = \frac{C_N + \zeta E[C_L]}{1 + \zeta}. \quad (8)$$

For comparison, in the plots we also show the recovery performance of RS-Spline, see Section III. The cost per round for RS-Spline is  $E[C_L]$ .<sup>5</sup>

In Figs. 4–6 we demonstrate the effectiveness of our recovery technique (“RS-PCA-CS” in the figures). These results prove that PCA is a suitable transformation to be used in conjunction with CS and that, despite the cost incurred in the training phases, the approach still provides substantial benefits with respect to standard data gathering schemes. In Fig. 4  $\varepsilon$  is close to zero as this specific signal varies slowly in time, i.e., its correlation structure is quasi-stationary during a monitoring phase. Also, we note that for those signals showing higher variations over space and time, such as luminosity, RS-Spline has unsatisfactory performance.

In the last two graphs, Figs. 7 and 8, we show the impact of  $K$  and  $\zeta$  on performance. From Fig. 7 (fixed  $K$ ) we see that decreasing  $\zeta$  leads to: 1) a decreased cost per round for a given quality goal and 2) smaller variance for  $\varepsilon$ . On the other hand, the minimum admissible cost per round so that the CS algorithm is able to approximate the input signal becomes smaller for increasing  $\zeta$ . From Fig. 8 (fixed  $\zeta$ ) we see that decreasing  $K$  is beneficial. This means that, for the considered

<sup>5</sup>We do not analyze the performance of RS-DCT-CS. In contrast to the synthetic signals of Section III, the real signals considered here are not sparse in the DCT domain, and thus RS-DCT-CS performs much worse than RS-Spline.

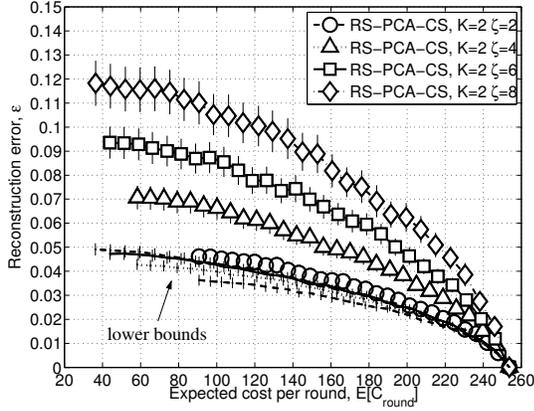


Fig. 7. Average  $\varepsilon$  (signals 1–5) vs  $E[C_{\text{round}}]$ ,  $K = 2$ ,  $\zeta \in \{2, 4, 6, 8\}$ .

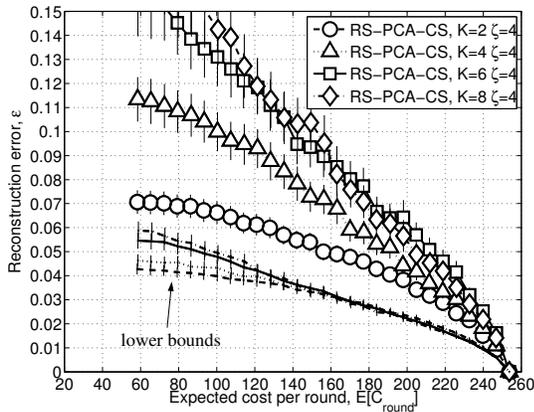


Fig. 8. Average  $\varepsilon$  (signals 1–5) vs  $E[C_{\text{round}}]$ ,  $K \in \{2, 4, 6, 8\}$ ,  $\zeta = 4$ .

signals, a smaller reconstruction error is achievable through more frequent updates of  $\bar{\mathbf{x}}$  and  $\hat{\Sigma}$ . In Figs. 7 and 8 solid and dotted lines without marks represent lower bounds on the error recovery performance, which are obtained as follows. For each  $(K, \zeta)$  pair and for each time  $k$  the input signal  $\mathbf{x}_k$  is reconstructed using RS-PCA-CS with  $\bar{\mathbf{x}}$  and  $\hat{\Sigma}$  calculated assuming perfect knowledge of the previous  $K$  instances of the signal  $\mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-K}$ . As mentioned above, the error so obtained is a lower bound on the reconstruction performance of RS-PCA-CS for a given training set of length  $K$ . The cost associated with the new  $\varepsilon$  is set equal to that of the real RS-PCA-CS scheme for the given  $(K, \zeta)$  pair. These curves reveal the impact of the obsolescence of  $\bar{\mathbf{x}}$  and  $\hat{\Sigma}$  during the monitoring phase for the considered signals. In particular, the recovery performance degrades for either increasing  $\zeta$  (Fig. 7) or  $K$  (Fig. 8).

## V. CONCLUSIONS

In this paper we designed an algorithm, based on Compressive Sensing (CS) and Principal Component Analysis (PCA), for the approximation of large and distributed data sets through the collection of a small number of samples. Our technique adapts, in an online fashion, to the non-stationarity of real

world signals and is independent of the considered routing scheme. This is possible through the online estimation of statistical properties of the signals, which are then used by PCA to derive the optimal transformation for CS. Hence, we showed the effectiveness of this solution and its superiority to standard data gathering approaches by considering real world signals, which were gathered from an actual WSN deployment. Even though a WSN scenario was considered for our performance evaluation, we stress that the technique is general and can be readily applied to the online approximation of distributed data in other types of network, e.g., cellular networks.

## REFERENCES

- [1] G. Quer, R. Masiero, D. Munaretto, M. Rossi, J. Widmer, and M. Zorzi, "On the Interplay Between Routing and Signal Representation for Compressive Sensing in Wireless Sensor Networks," in *Information Theory and Applications Workshop (ITA 2009)*, San Diego, CA, US, Feb. 2009.
- [2] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning Compressed Sensing," in *Forty-Fifth Annual Allerton Conference*, Allerton House, UIUC, IL, USA, Sept. 2007.
- [3] A. Scaglione and S. D. Servetto, "On the Interdependence of Routing and Data Compression in Multi-Hop Sensor Networks," in *ACM MOBICOM*, Atlanta, GA, USA, Sept. 2002.
- [4] D. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, July 1973.
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991.
- [6] S. Pradhan and K. Ramchandran, "Distributed Source Coding Using Syndromes (DISCUS): Design and Construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.
- [7] H. Luo and G. Pottie, "Routing Explicit Side Information for Data Compression in Wireless Sensor Networks," in *Distributed Computing in Sensor Systems (DCOSS)*, Marina del Rey, CA, USA, June 2005.
- [8] S. Patten, B. Krishnamachari, and R. Govindan, "The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks," in *Int. Conf. on Information Processing in Sensor Networks (IPSN)*, Berkeley, CA, USA, Apr. 2004.
- [9] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Joint source-channel communication for distributed estimation in sensor networks," *IEEE Trans. on Information Theory*, vol. 53, no. 10, pp. 3629–3653, Oct. 2007.
- [10] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressive Sensing for Networked Data: a Different Approach to Decentralized Compression," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 92–101, Mar. 2008.
- [11] G. Shen, S. Y. Lee, S. Lee, S. Patten, A. Tu, B. Krishnamachari, A. Ortega, M. Cheng, S. Dolinar, A. Kiely, M. Klimesh, and H. Xie, "Novel distributed wavelet transforms and routing algorithms for efficient data gathering in sensor webs," in *NASA Earth Science Technology Conference (ESTC2008)*, University of Maryland, MD, USA, June 2008.
- [12] M. Duarte, S. Sarvotham, D. Baron, M. Wakin, and R. Baraniuk, "Distributed Compressed Sensing of Jointly Sparse Signals," in *Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on*, Oct. 2005.
- [13] C. R. Rao, "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya: The Indian Journal of Statistics*, vol. 26, pp. 329–358, 1964.
- [14] K. Fan, "On a theorem of Weil concerning eigenvalues of linear transformation I," *Proc. of the National Academy of Sciences*, vol. 35, pp. 652–655, 1949.
- [15] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. on Information Theory*, vol. 52, no. 9, pp. 4036–4048, Sept. 2006.
- [16] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [17] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed L0 norm," *IEEE Trans. on Signal Processing*, 2009, Accepted for publication.
- [18] D. T. Sandwell, "Biharmonic Spline Interpolation of GEOS-3 and SEASAT Altimeter Data," *Geophysical Research Letters*, vol. 14, no. 2, pp. 139–142, 1987.