Inter-sensor Modeling from Multi-Sample/Multi-Sensor: Techniques and Applications

Jennifer L. Wong
CENS, Dept. of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
Email: jwong@cs.ucla.edu

Seaphan Megerian
Dept. of ECE
University of Wisconsin, Madison
Madison, WI 53706
Email: megerian@ece.wisc.edu

Miodrag Potkonjak
CENS, Dept. of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
Email: miodrag@cs.ucla.edu

UCLA Computer Science Department
Technical Report #060005

Abstract

Inter-sensor modeling of data streams is an important problem and an enabler for numerous sensor network tasks such as faulty data detection, missing data recovery, and compression. Using a simple and fast algorithm, we developed a lower bound approach for evaluating the achievable accuracy of regression between the readings at two sensors. We have also developed a symmetric monotonic regression (SMR) technique for predicting data at one sensor using data from another sensor or a set of sensors. SMR often performs very close to the lower bound on a set of collected real-life sensor data, which indicates that more accurate modeling is possible only if prediction is conducted using either data from multiple sensors or by considering information extracted from statistical properties of dynamics (multiple consecutive time samples) of the explanatory stream.

Node assignment is often used for energy efficient data collection in sensor networks. We have developed a new problem formulation and an integer linear programming (ILP) formulation for the node assignment problem which considers all the requirements of the developed prediction models. We first introduce an ILP formulation that generates optimal solutions for medium sized network instances. After that, we present a heuristic that iteratively applies ILP on sub-instances in order to address large instances. We demonstrate the effectiveness of the SMR models on the node assignment problem using collected data in an indoor environment. Our study indicates that the application of the new intersensor models and ILP does not just improve the effectiveness in comparison with the situation where no optimization is performed, but also improves the lifetime of the network by at least an order of magnitude to the situation when traditional prediction from a single sensor is used.

1 Introduction

Modeling is a problem that inherently permeates many tasks in both sensor networks and computational sensing. In terms of complexity, capability to be rigorously defined, conceptual difficulty, and required creativity it spans a wide spectrum from modeling faults and errors in sensor readings to modeling dynamics of the instrumented environment and properties of events of interest. Intersensor modeling aims to predict a reading at a particular sensor at a particular time using one or more readings from other sensors at the same or other times and/or readings from the same sensor at other time moments. Intersensor modeling is probably the single most pervasive prediction task in sensor networks. Our objective is to develop a systematic non-parametric data-driven approach for intersensor modeling. Using data from the Intel Berkeley Lab of readings of 54 sensors of three modalities (temperature, humidity and light) during a period of three weeks, we first derive the conclusion that modeling of a single sensor reading from a single simultaneous reading from another sensor is in many situations is not capable to satisfy the user's requirements. Interestingly, when this prediction is effective we often also have effective prediction from time shifted data from the predicting sensor. These conclusions are obtained using two modeling techniques. The first is the establishment of lower bounds on the accuracy of any arbitrary regression technique that uses single readings from one sensor to calculate the measurement at another sensor. The second tool used to address this problem is symmetric monotonic regression (SMR) that invariably in all of the datasets produces prediction errors that are very close to the lower bounds. SMR benefits from two mechanisms: symmetry and monotonicity. Symmetry ensures that one simultaneously takes into account the errors of predicting sensor A from sensor B and sensor B from sensor A. Monotonicity is a constraint that enforces that for a pair of measurements at sensor A, m'_A and m''_A, such that m'_A is smaller than m''_A readings at sensors B in the corresponding moments m'_B and m''_B satisfy the condition that $m''_B \geq m'_B$. The intuition behind monotonicity is the following. Two sensors are well correlated most often only if they are exposed to similar or identical sets of sources of excitation of the same modality. If one of these sources increase (decreases) its intensity the impact will be the same on both sensors.

Our study of intersensor modeling indicates that there exists at least three conceptually different ways to break the barrier imposed by the lower bound on the error of regression applied to a pair of sensors. The first is to use the implicitly available variable, time, to facilitate more accurate prediction. For example, two light sensors exposed to the sun will have different relationships and, therefore, different regression models during the morning and evening. We found that this option is a special case of the less restrictive paradigm where the accuracy of prediction is improved by considering trends in recent readings from the predicting sensor. Specifically, if we build separate models for predicting sensor A from sensor B when readings of sensor B are increasing and when they are decreasing, we can significantly improve the accuracy of the regression models.

The second way of breaking the lower bounds barrier is to use time shifted versions of the data. In a sense the first method is a special case of the second approach when time shifting is conducted only on the predicting sensor. However, the full power of time shifting-based prediction comes at the moment when we use very few time shifted readings from the predicted sensor to facilitate derivation of accurate regression models. Finally, the third and in a sense the most natural approach for improving prediction beyond the lower bound of two sensor regression is to use multiple sensors for predicting a single sensor. Due to issues related to the curse of dimensionality, computational complexity, and combinatorial explosion [11], we restricted our study to regression from only two sensors. In this case again we used the lower bounds and SMR to show significantly higher potential for prediction from multiple sensors and to deduce a close match between the lower bounds and SMR in multi-sensor prediction. The most challenging conceptual problem in multi-sensor prediction is that the requirements for the number of measurements to facilitate modeling grows very rapidly (exponentially) with an increase in the number of sensor from which prediction is conducted. The standard non-parametric statistical technique to address this issue is smoothing [11]. However, we demonstrate that SMR employs monotonicity as a very effective mechanism to interpolate sparse data points in the prediction space.

The rest of the paper is organized in the following way. We first discuss all modeling techniques and analyze them using data from the Intel Berkeley Lab. Once the models are available, we demonstrate their usefulness by applying these models to formulate an effective node assignment strategy for nodes in sensor networks using time shifted, partitioned, and prediction from multiple sensor regressions. After that we develop an ILP formulation for deriving the node assignment problem using the developed regression models.

While the ILP formulation is simple enough to guarantee derivation of an optimal schedule for all instances of the node assignment problem to the Intel Berkeley Lab dataset, in order to enhance its applicability range we have developed two techniques, time compression and a iterative-improvement heuristic using ILP to further enhance the application range of the node assignment problem. Our study indicates that the application of the new intersensor models and the ILP formulation does not just improve the effectiveness in comparison with the situation where no optimization is performed, but also improves the lifetime of the network by an order of magnitude to the situation when traditional predictions from a single sensor are used.

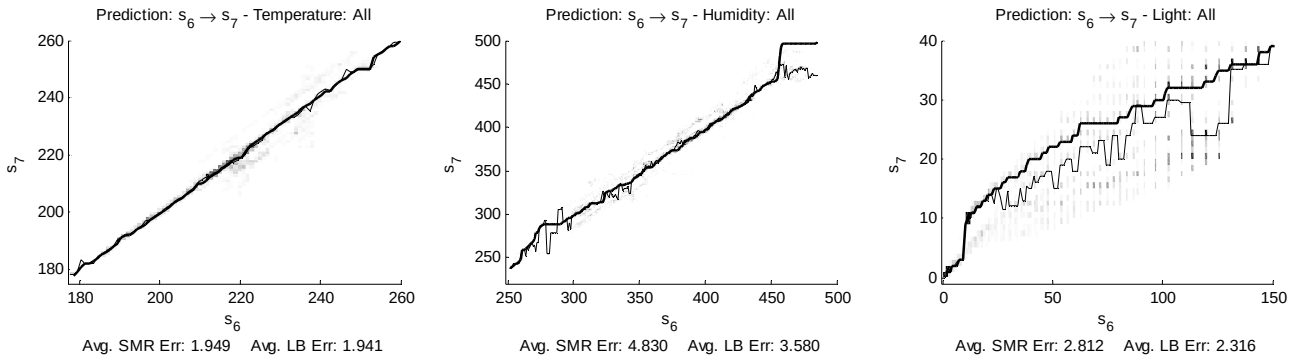


Figure 1: SMR and lower bound models for sensors 6 and 7 for three modality over 8 days.

2 Related Work

We briefly survey the most directly related work in this section. Nowak [17] presented an application of a distributed expectation maximization (EM) algorithm for density estimation in sensor networks. Coates [4] proposed distributed estimation of the current state at multiple sensor nodes via particle filtering. Delouille et al. [7] developed a new iterative distributed algorithm for linear minimum mean-squared error (LMMSE) estimation in sensor networks whose measurements follow a Gaussian hidden Markov graphical model with cycles. In [10], a kernel linear regression approach for in-network modeling is presented. Their goal is to develop linear relationships between correlated sensors in such a way that constraints on the model parameters are communicated instead of the data itself. Paskin et al. [18] developed an architecture for distributed inference in a sensor network using a combination of graphical models and junction trees.

The first monotonic regression approach is Pool Adjacent Violators Algorithm (PAV) proposed by Brunk in 1955 [3]. One popular form of kernel smoothing after applying PAV is proposed by Mukerjee [16], who used the Nadaraya-Watson estimate, for this task. Dette and Pilz [8], proposed a technique that combines density with a regression estimate to obtain a monotone estimate of the inverse regression function. In 1964, Kruskal published a paper that in context of his multidimensional scaling solved the monotonic regression problem [15]. The approach is essentially identical to the PAV algorithm. It has been used and refined by a number of researchers in particular in the signal processing community [6, 5, 2, 21, 20]. Koushanfar et al. [14] have developed a combinatorial isotonic regression (CIR) approach for intersensor modeling. We significantly improved the accuracy of this technique by simultaneously addressing both monotonicity (aka isotonicity) and symmetry.

One of the key issues in wireless sensor networks is power conservation. A number of techniques have been proposed at all levels of the design process from communication protocols [24] to digital signal processing [22]. Willet et al. introduced backcasting where adaptive sampling is applied for efficient field estimation [23]. Jain et al. [12] proposed an adaptive sampling approach which varies the sampling rate at each sensor and therefore adapting to the streaming-data characteristics of the sensor. The use of mobile sensor nodes are used to determine sampling density required in various environmental regions in [1]. Their Fidelity Driven Sampling actively seeks to minimize error without prior knowledge of the variable field.

Kar et al. [13] advocated and theoretically analyzed a technique for dynamic node activation in networks of rechargeable sensors. Koushanfar et al. [14] have developed an ILP formulation that addresses sleeping assuming only synchronous prediction of samples at one node from samples of an another single node.

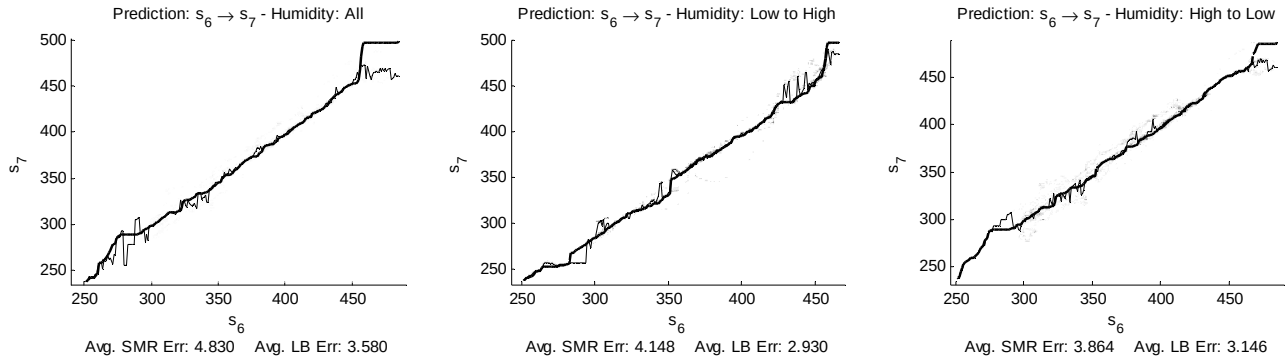


Figure 2: SMR and LBR for sensor 6 and 7 humidity data over 8 days.

Their work can be considered as a very specialized case of our effort both in terms of modeling, number of techniques used to enable efficient sleeping, and a much simpler ILP formulation that is applicable only on smaller instances of the networks. All of the proposed sampling schemes assumed simultaneous sampling at all nodes. Our goal is to demonstrate that by relaxing this requirement and using time-shifted data for data recovery we can improve the lifetime of the network by more than an order of magnitude while maintaining the user specified level of accuracy.

3 Modeling

In this section we address traditional and new techniques for intersensor modeling. We first introduce a lower bound technique and SMR for prediction from a single sensor. After that in the next three subsections we introduce and analyze the three techniques that enable improved modeling accuracy: (i) trend identification, (ii) occasional calibration, and (iii) prediction from multiple sensors.

3.1 Lower bound and Symmetric Monotonic Regression

There is a wide variety of available regression models that predict the value of the target variable of interest (sensor reading at node A) using an observed value of the predicting variable (measurements at node B). They range from simple linear and polynomial regression to complex and computationally intensive techniques such as nonlinear least squares, weighted least squares and Loess regressions [11]. Therefore, identification of the best regression technique for a given data set is a difficult and time consuming problem. In order to speed the identification process and to gain better insight into the structure of the targeted data set, we have developed lower bound regression (LBR). While all other regression technique use one data set (learning) to build model and other (testing data set) to evaluate it, LBR is built using the testing data set. The LBR can be easily calculated in linear time in terms of the available number of samples for any norm, such as L_1, L_2, and L_∞. For example, to calculate the optimal LBR with respect to the L_1 norm, find for each value of variable B the median of corresponding values at variable A. Or, for L_2-optimal LBR to find the average value of values of variable A that occur for each value of variable B. LBR can be used in at least two ways. The first is to evaluate other regression techniques in terms of how close it is to LBR. Small percentage error between the two regressions indicate that the evaluated regression technique is performing well. The other is to terminate the search for two sensor intersensor regression models, if the quality of the LBR prediction is not satisfactory to the user.

In addition to LBR, we have also developed symmetric monotonic regression (SMR). SMR has four important properties which make SMR well-suited for application to modeling of sensor readings. The first is that the symmetry requirement implies simultaneous consideration of predicting both readings of sensor A from sensor B and vice versa. This is important both in order to make the regression more robust and to make it more suitable for prediction using a combination of readings from several sensors. Note that if regression is not symmetric (and a great majority of regressions are not) for a majority of the values, when we attempt simultaneous modeling from two sensors we will have a process that converges toward some value

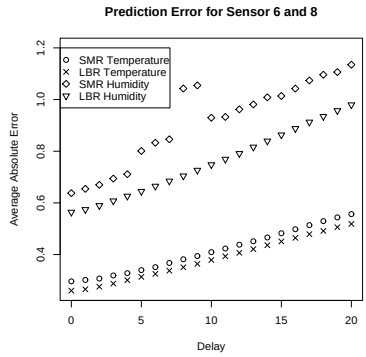


Figure 3: Prediction error for sensor 6 and 8 for temperature and humidity for delays between 0 and 20.

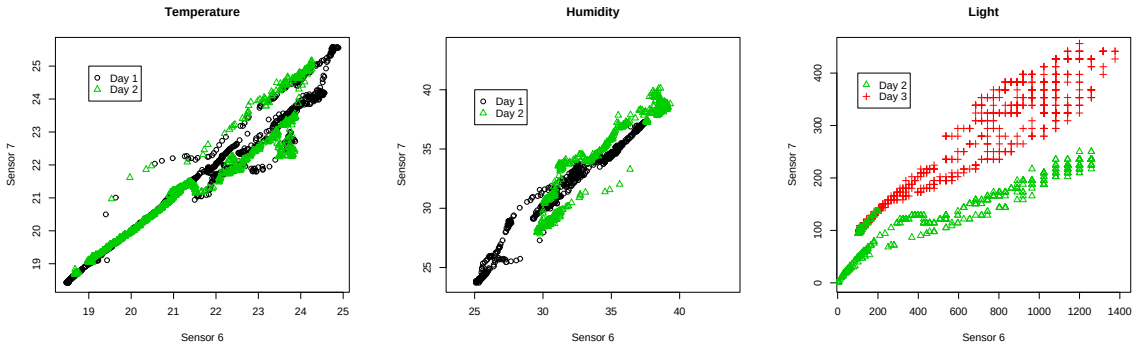
that is the intersection of the regression functions. This is so because for two regression functions $y = F(x)$ and $x = G(y)$, for a majority of values x, $x \neq G(F(x))$, except for the points where functions $y = F(x)$ and $x = G(y)$ intersect.

The second advantage is monotonicity which is discussed in the Introduction. The third advantage is that SMR often performs very close to LBR as shown in the rest of this section. Finally, there is an easy and fast way to compute SMR using combinatorial techniques. Note that readings of digital sensors have a finite number of values. If this is not the case or the number of values is large, binning can be used to translate regression into the graph domain. Each node in the graph domain corresponds to a pair of values v_A, v_B. For each value and therefore the graph node, we can easily calculate an arbitrary error using the same technique as for LBR for both predicting A for B and vice versa. In order to enforce monotonicity constraints, we create edges in the graph in such a way that they are always either horizontal (change in v_A value) or only between the nodes that satisfy monotonicity constraints (see the Introduction section). Now, finding the path between nodes that correspond to minimal value at sensor A and any node that corresponds to the maximal value at sensor A creates SMR. This path can be find using dynamic programming or the generic Dijkstra shortest path algorithm in low complexity polynomial time.

Our implementation of SMR relies on binning by integral values of the data. Thus, in order to specify the proper binning resolutions in our experimentations, we had to apply a scaling factor to all data sets. For temperature and humidity, we scale all data elements by 10, while for light, we scale by 0.1. Consequently, in order to maintain a consistent visual and tabular basis for comparison, all figures, tables, and values presented here reflect results obtained after scaling, unless specified otherwise. Among all data sets, the minimum, mean, and max observed values for each of the modalities is: Temperature (min=14.8, mean=22.4, max=37.7), Humidity (min=14.3, mean=37.6, max=60.9), Light(min=0.0201, mean=303.3, max=1847). Furthermore, the error metric reported corresponds to average absolute error, obtained by dividing the L_1 norm of the differences between predicted values and observed data by the number of data samples.

In order to illustrate the LBR and the close match between the prediction error between the LBR and the SMR technique, Figure 1 shows the prediction error of these two techniques for sensor 6 and 7 for the three modalities. The evaluation is conducted from a dataset that consists of eight days of sensor measurements. Note, that the LB is derived using exactly this dataset while the SMR is developed using a learning dataset that consists of three days of measurements. As we see at the bottom of the figures the percentage difference between error of SMR and LB are 0.41%, 34.9%, and 21.4% for temperature, humidity, and light respectively.

When the cross-correlation between the two sensors is high, the single sensor SMR approach can provide high accuracy prediction. Additionally, this prediction can be delayed in time, we call this time-shifted SMR (TSMR). We have found that a sensor can predict another sensor with high accuracy from data collected 10, 20 or even 40 epochs later. In Figure 3 we illustrate the trends in error for increasing time-shift for humidity and light modalities for the same sensors.



(a) Temperature (b) Humidity (c) Light

Figure 4: Scatterplots of Sensors 6 and 7 on two consecutive days with different trends.

3.2 Trends-based regression (TR)

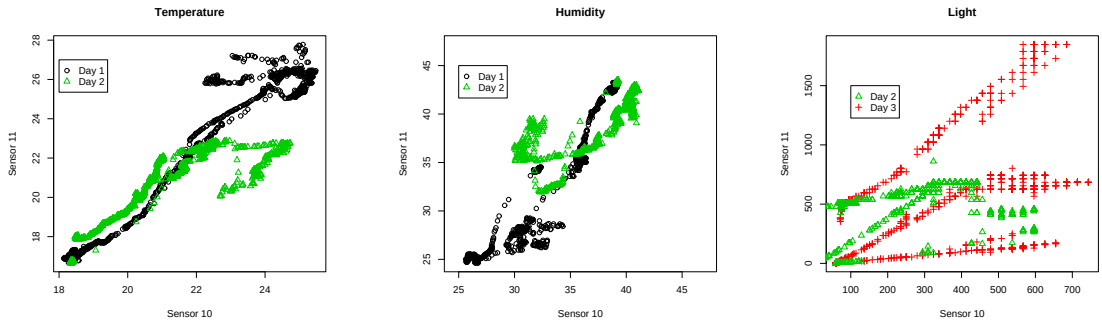
Splitting of data sets into two or more partitions is a well known statistical technique to improve the accuracy of modeling. We evaluated several technique for data splitting, including one where splitting is conducted using time-dependent information, the rate and/or the stability of the change of data samples (signal) at the predicting sensor, and one that models the conditional probabilities of having a specific value at the predicting sensor after τ steps before a particular value was recorded. The best performing technique was the conceptually and computationally the simplest technique where the splitting is done by observing the gradient of the predicting signal. Essentially, we build separate models for the periods of time where predicting variable is increasing and decreasing. In order to increase the robustness, we defined the increasing (decreasing) periods as ones where the current value is larger than the average of the values that were recorded 10, 20, 30, and 50 epochs earlier. Trends-based regression (TR) is defined as SMR applied separately on the increasing and the decreasing trend subsets of data.

Figure 3 shows the results of applying TR on humidity sensors 6 and 7. We see that although we were able to significantly improve both LBR and SMR accuracy, we were not able to break to LBR barrier by splitting the data into increasing and decreasing trends.

3.3 Occasional Calibration Regression (OCR)

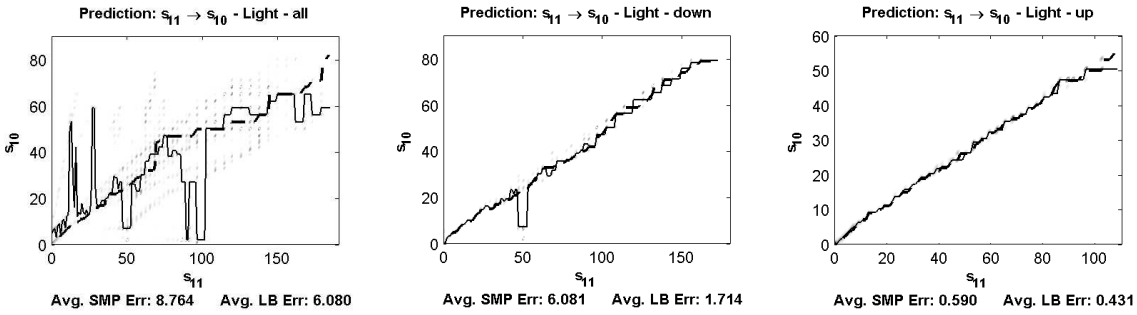
As shown in Figure 4 and 5 there is often very strong autocorrelation of the data samples recorded at the same sensor or cross-correlation for different time lags, in particular after data partitioning using trends. At its straightforward interpretation this observation suggests that values at sensor A can be well modeled from time shifted values of sensor B. However, if we relax the definition of regression and allow occasional use of samples at sensor A for prediction of other samples of the same sensor in addition to the use of time synchronized samples from the samples from sensor B, we can create a more accurate generalized regression technique. In our experimentation, we selected to use for this purposes every 60th sample of the sensor. We named this technique Occasional Calibration regression (OCR) and have developed the following algorithm for the calculation of OCR models.

First, we build a probability density function (PDF) for predicting the readings at sensor A using the readings at sensor B. The PDF provides information not just for what is the most likely value at sensor A for a given value at sensor B, but also the probability that an arbitrary value is observed at sensor A for a particular value at sensor B. The PDF is built using recursive application of SMR. After each application of SMR, we split the data sets into two subsets: one that contains all points with actual values higher than predicted by the SMR and one where all values have lower actual values than predicted by SMR. In such a way, we first build a PDF curves for 50%, then for 25% and 75%, then for 12.5%, 37.5%, 62.5%, and 87.5%, and so on. Once the PDF is available, we use occasional readings to identify the corresponding PDF curve that is used for the prediction of all samples until the next occasional reading value is collected. In order to improve the robustness of the technique, we used an exponentially weighted average of the last 3 readings where the weight factors were 4, 2, and 1 from the most recent to the oldest occasionally collected value at



(a) Temperature (b) Humidity (c) Light

Figure 5: Scatterplots of Sensors 10 and 11 on two consecutive days with different trends.



(a) Temperature (b) Low to High (c) High to Low

Figure 6: SMR and lower bounds of Sensors 10 and 11 on a all days vs. a single day.

sensor A.

In Figure 7 we illustrate the PDF built for the decreasing trend (shown in Figure 6 (c))for humidity prediction for Sensor 10 and 11. Each line from bottom to top in the figure represents the PDF curves 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, and 87.5%.

3.4 2-Dimension Symmetric Monotonic Regression (2D SMR)

Our final model is a 2D SMR where samples from two sensors B and C are used to model the values at sensor A. Note both the 2D SMR and 2D LBR can be directly applied to time-shifted data from one or both predicting sensors B and C. In order to evaluate the accuracy of SMR for this task, we have also developed a 2D LBR that is straight-forward conceptual and algorithmic generalization of the 1D case. However, the development of algorithm and software for SMR is more complex. For this purpose, we have developed the following ILP formulation.

3.4.1 Symmetric Monotonic Regression ILP

In this subsection we introduce an ILP formulation for addressing the SMR modeling problem for prediction of a single sensor reading from two other sensor readings.

When building the symmetric monotonic prediction model for two sensors it is assumed that the training data consists of (x, y, z) triples where each component is a sensor reading for the corresponding sensor. Given three sensors, x, y, and z, the goal is to determine the value for sensor z with minimal L_1 error, given the values of sensors x and y.

The symmetric monotonic prediction model can be addressed using an ILP formulation. Prior to formulating the problem, a preprocessing step is performed where the total L_1 error in the X and Y-axis is calculated for each possible prediction value of z at (x, y). By considering the error in both the X and

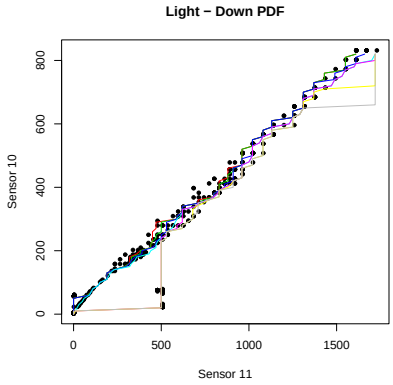


Figure 7: PDF for Sensors 10 and 11, High to Low.

Y directions, we are ensuring symmetry. We denote the preprocessed constant error values by E_{xyz}. The formulation contains one set of variables v_{xyz} which denotes if value z is selected as the SMR prediction model value for point the corresponding (x, y) position.

$$\begin{array}{rcl} E_{xyz} & - & \text{total error in X and Y-axis at position } (x, y, z) \\ v_{xyz} & = & \begin{cases} 1, & \text{if value } z \text{ selected for SMR at } (x, y) \text{ position} \\ 0, & \text{otherwise.} \end{cases} \end{array}$$

There are three types of constraints in the ILP formulation of the problem. The first set of constraints is that for each (x, y) position only a single value for z may be selected. Eq. (1) denotes this constraint. In order to ensure that the model is monotonic to both the X and Y-axis, two sets of constraints are added. The first ensures that for each value selected for z in terms of the X-axis, that the value is always increasing (Eq. (2)). The constraint is formed by multiplying the constant value for the selected z position at the current position (x, y) by the selected position, and ensuring that this value is greater than the value of selected for the position $(x - 1, y)$. An identical set of constraints is created to enforce the same condition in the Y-axis (Eq. (3)). Finally, Eq. (4) states the objective function for the problem which is to minimize the sum of all error in the X and Y-axis for each selected z value.

$$\text{for all } x, y : \sum_z v_{xyz} = 1 \tag{1}$$

$$\text{for all } x, y : \sum_z z v_{xyz} \geq \sum_z z v_{(x-1)yz} \tag{2}$$

$$\text{for all } x, y : \sum_z z v_{xyz} \geq \sum_z z v_{x(y-1)z} \tag{3}$$

$$Y = MIN(\sum_x \sum_y \sum_z E_{xyz} v_{xyz}) \tag{4}$$

Figure 9 shows a plot of the 2D SMR model for predicting temperature readings of sensor 8 from sensors 6 and 7. In Figure 8 the SMR models for prediction of sensor 8 using data from sensor 6 alone and using data from sensor 7 alone are shown. We see significant improvement (2.96 and 4.05 individually, 1.51 combined) in accuracy. In Table 1 we present the SMR, LBR, linear regression (P_1) and a quadratic fit (P_2) for each of the predictions individually for all three modalities (temperature, humidity, light) for sensor 6 or 7 predicting sensor 8. The average absolute SMR and LB errors for 2D prediction of sensor 8 from both sensor 6 and 7 was 1.51, 0.78 for temperature and 1.92, 0.91 for humidity, respectively. In this case, while their was a large improvement for combined prediction for temperature and humidity, for light this was not the case. The error in the LBR for light was 14.85 (no improvement over LBR of individual predictions), signifying that SMR would not perform better than the individual predictions.

Pair	$s_i \rightarrow s_j$				$s_j \rightarrow s_i$			
$i-j$	SMR	LBR	P_1	P_2	SMR	LBR	P_1	P_2
T6-8	2.96	2.60	3.28	3.23	2.96	2.24	2.93	3.32
T7-8	4.05	3.70	4.49	4.70	3.87	3.29	4.14	4.48
H6-8	6.38	5.64	7.83	9.46	5.95	4.78	7.15	6.45
H7-8	10.35	6.91	9.51	10.92	7.23	6.27	9.30	8.55
L6-8	9.69	5.57	12.03	21.5	6.61	5.12	6.49	8.61
L7-8	17.1	10.1	27.6	28.8	2.76	1.97	4.21	4.54

Table 1: Average absolute error in SMR, LBR, and comparison to linear regression (P_1) and a quadratic fit (P_2).

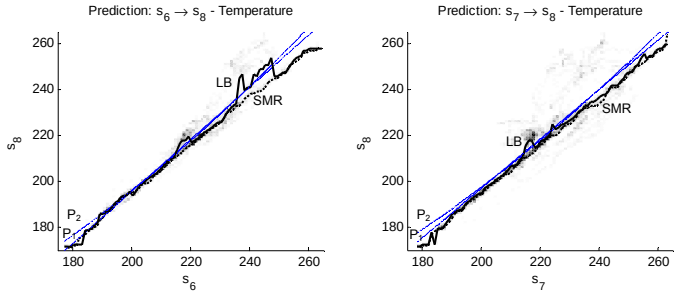


Figure 8: SMR Model for prediction of sensor 8 from sensor 6 and 7 individually.

4 Node Assignment

In this section we address the node assignment problem for energy efficient data collection in sensor networks. The models developed in the previous section are used as enablers for placing a different subset of nodes at each time moment into sleep mode in such a way that their values can be recovered using the available measurements and the regression models. Note that occasional sampling has an important side benefit: it can be used to trigger comprehensive data collection in the situation when discrepancy between the predicted and actually collected data is detected. We have developed several quantitative strategies for this task but we will not present them in this paper due to space limitations and more importantly since we did not detect a single instance when any of the strategies was initiated. Interestingly, the same strategies were regularly triggered when only modeling from a single sensor was used for node assignment [14].

4.1 Problem Formulation

When addressing the node assignment problem it is assumed that all data is collected from the network and processed at a data sink (aka gateway or fusion center). This is the approach employed during‘ collection of the dataset at the Intel Berkeley Lab. The data sink is non-energy constrained and has sufficient processing resources. We assume that the main component of energy consumption in the system is communication. Therefore, any sensor that is not collecting sensor readings or communicating is placed in a low power sleep state where the minimal amount of energy is consumed and does not communicate. Finally, the approach is sensor data-driven, in the sense that it has to be conducted on actually collected data that is at least partially cross-correlated and predictable as is the case for the sensor network at Intel Berkeley Labs and many other densely deployed networks.

It is assumed that a training set is collected and then processed off-line. All of the prediction models (see Section 3) are built for each pair of sensors. To address the trend-based regression (TR) models, it is assumed that a set of models for each trend are built for each of the remaining models (OCR, TSMR, 2D SMR). For each of the built model sets which are to be used to address the node assignment problem it is assumed the following inputs are given. First, all pairs of sensors for which simultaneous sampling leads to user specified prediction ($P\%$) are defined in order to use the occasional calibration regression (OCR) model. For the time shifted single sensor regression (TSMR) model, the maximum allowable phase delay for each sensor prediction pair is determined by examining the error in each phased SMR, the model with the largest

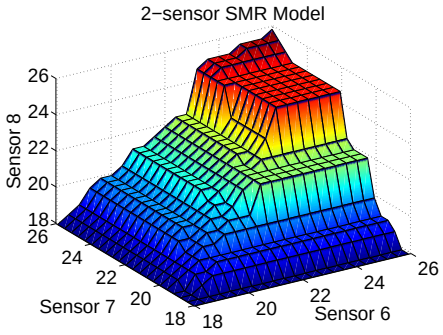


Figure 9: 2D SMR Model for prediction of sensor 8 from sensor 6 and 7 .

phase which still satisfies the specified user accuracy ($P\%$) for prediction is then selected as the maximum delay. Finally, for the 2D SMR model, for all pairs of sensors which can jointly predict a third sensor with the specified user accuracy ($P\%$) a set of phased value pairs are defined. Each phased value pair specifies the rephasing delay for each of the predictor sensors. Using these inputs, we define the node assignment problem formally as follows.

Problem: *Node Assignment Problem*
Instance: *An (ixj) binary matrix B of "simultaneous sampling", an (ixj) integer matrix D of "maximum rephasing", (ixjxk) sets of pairs* $E_{ijk} = \{(v_j, v_k), \ldots\}$ *for "2-sensor rephasing", a positive integer W for "window size", and a positive integer S of "maximum samples".*
Question: *Is there an assignment of each sensor i to at most S time steps in W s.t. each sensor i at each time step* t_i *in W is assigned to* t_i *or either*
(i) at least one sensor j at t_j *where* $B_{ij} = 1$ *and i at* $t_j = 1$*, or*
(ii) at least one sensor j at t_j *where* $(t_j - t_i)\%W \leq D_{ij}$*, or*
(iii) at least one sensor pair j,k where $(t_j - t_i)\%W \leq v_j$ *and* $(t_k - t_i)\%W \leq v_k$*?*

4.2 Complexity

The node assignment problem in NP-complete. An arbitrary instance of the Domatic Number Problem [9] can be transformed into a special case of the node assignment problem where only model the time shifted single sensor regression model is considered (option (ii) in the problem formulation).

4.3 ILP

The node assignment problem can be formulated using any subset of the models presented in Section 3. We introduce a set of constraints for the use of each model in the ILP formulation. The base formulation for the problem assumes the size of the sampling window W is predetermined. The goal is to assign each sensor to collect readings at the minimal number of epochs $t = \{0, \ldots, W\}$ in the window W. Variable s_{it} is used to denote weather sensor i is to be sampled at epoch t. The second set of variables, p_{it} is used to determine if a reading at epoch t for sensor i is predictable using any of the models. Finally, the last variable l is used to determine the maximum number of samples take by any single sensor in the window. The objective function is to minimize this variable (Eq. (5)). Note that uppercase letters are used to refer to constant values, while lowercase letters refer to variables in the ILP formulation. Additionally, we will use B to refer to the OCR model, D to refer to the TSMR model, and E to refer to the 2D SMR model.

$$
\begin{aligned}
W &= \text{size of periodicity window} \\
s_{it} &= \begin{cases} 1, & \text{if sensor } i \text{ is sampled at epoch } t \\ 0, & \text{otherwise.} \end{cases} \\
p_{it} &= \begin{cases} 1, & \text{if sensor } i \text{ is predictable at epoch } t \\ 0, & \text{otherwise.} \end{cases} \\
l &= \text{largest number of samples by any sensor} \\
b_{it}, d_{it}, e_{it} &= \begin{cases} 1, & \text{if sensor } i \text{ is predictable by model} \\ & B, D, \text{ or } E \text{ at epoch } t, \text{ respectively} \\ 0, & \text{otherwise.} \end{cases} \\
q_{ij} &= \begin{cases} 1, & \text{if sensor } i \text{ is simultaneously} \\ & \text{sampled with sensor } j \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

The base ILP formulation includes three sets of constraints. The first set, Eq. (6), specifies that for each sensor the sum of the samples taken by that sample must be less than or equal to l, the largest number of samples taken by any sensor. The second set of constraints (Eq. (7)) specifies that at each epoch, each sensor must be either sampled or/and predicted. Finally, the constraint in Eq. (8) specifies that if a sensor i is to be predicted at epoch t, then the it must be predicted by one of the models (B (OCR), D (TSMR), or E (2D SMR)). Note that the ILP formulation uses the constraint of only three of the five models. The other model (TR) is addressed by introducing separate ILP instances to handle each of the trend cases. Specifically, separate ILP instances for the increasing and decreasing trends must be solved. The formulations are exactly as presented. However, the unique input models are used for each of the trends independently.

$$Y = MIN(l) \tag{5}$$

$$\text{for all } i : \sum_t s_{it} \leq l \tag{6}$$

$$\text{for all } i, t : s_{it} + p_{it} \geq 1 \tag{7}$$

$$\text{for all } i, t : b_{it} + d_{it} + e_{it} \geq p_{it} \tag{8}$$

It is important to note that each model may not be possible or necessary for prediction of a particular sensor. Therefore Eq. (8) must be modified only to include the feasible models for predicting the particular sensor i at each epoch t. We introduce variables b_{it}, d_{it}, and e_{it} in order to formulate the prediction cases for sensor i at epoch t using any of the three models (OCR, TSMR, 2D SMR), respectively.

In order to incorporate the occasional calibration model (B), where two samples are simultaneously sampled for calibration purposes, into the ILP formulation we introduce two sets of constraints. The first set of constraints is created for any pair of sensors i and j which can predict each other with high accuracy. In this case the constraint, Eq (9), specifies that if two sensors are sampled in the same epoch, through the use of a logical AND operation, then denote the relationship by variable q_{ij}. The logical AND operation can be implemented in ILP by introducing the following three constraints: $s_{it} \geq q_{ij}$, $s_{jt} \geq q_{ij}$, $s_{it} + s_{jt} - 1 \leq q_{ij}$. The second set is used to specify that if any sensor j can correct sensor i with an accuracy of $P\%$ and is sampled at epoch t, then i is predictable using model B (ie. $b_{it} = 1$).

$$\text{for all } i, j \text{ where } B_{ij} \geq \beta : \sum_t (s_{it} \wedge s_{jt}) \geq q_{ij} \tag{9}$$

$$\text{for all } i, t : \sum_j (s_{jt} \wedge q_{ij}) \geq b_{it} \tag{10}$$

The constraint for the time-shifted single sensor SMR model (D) ensures that a sensor i is predictable using model D at each epoch (ie. $d_{it} = 1$ if and only if at least one sensor j can predict sensor i within the time-shifted value D_{ij}). For each sensor i at each possible epoch t the predictability variable, d_{it}, is calculated. In the on-line case, which we are considering, the value is only predicted from previously sampled sensor readings. Therefore, for each possible predictor sensor j if j is sampled at any epoch between $t - D_{ij}$ and t then i is predictable at t. To formulate this constraint we calculate the sum of all sampled sensors j which occur within the time period $\tau = (t - D_{ij}), \ldots, t$. If there are no sensor readings which can be used to predict sensor i at epoch t (ie. summation is zero), then d_{it} must be assigned to zero. However, if the

summation is one or more, then d_{it} can be assigned to zero or one. This is acceptable because Eq. (7) will ensure that sensor i is sampled at time k or that it is predictable, and therefore forcing d_{ik} to one if necessary.

Note that in Eq. (11) we denote the calculation of the time period for predictability of each sensor using modulus W. Since W is the duration of the periodicity window, it is acceptable for a sensor to be predicted from a previous window under the assumption that the epoch it is within the specified time-shift. Therefore, the modulus of time position of the sampled sensor reading is taken into account for the time-shift.

$$\begin{array}{rcl} D_{ij} & - & \text{max time-shift for sensor } j \text{ to predict } i \\ E_{ijk} & - & \text{prediction ability of } i \text{ from sensor } j \text{ and } k \end{array}$$

$$\text{for all } i,t : \sum_{j} \sum_{\tau=0}^{D_{ij}} x_{j[(t-\tau)\%W]} \geq d_{it} \tag{11}$$

$$\text{for all } i,j,k,t \text{ where } |E_{ijk}| > 0 :$$

$$s_{j[(t-v_j)\%W]} + s_{k[(t-v_k)\%W]} \geq e_{it} \tag{12}$$

The final constraint is for addressing the use of the 2D SMR model (E). If two sensors j and k can predict sensor i above $P\%$ accuracy, then i is predictable at t. Note that the constraint in Eq. (12) is to be written for any time shifted combination of epochs which are the time phased pairs in the set E_{ijk}.

4.3.1 Time Compression

One of the main issues with using ILP formulations is the trade-off between complexity and size of the formulation with the runtime. We have developed a simple scaling mechanism which addresses the exponential growing runtime with the size of the specified window and number of constraints. When formulating the node assignment problem with time-delays (TSMR and 2D SMR models) the number of possible solutions grows with the amount of allowable time-shift for each sensor. Due to the fact that sensor i can predict j for any time-shift less than D_{ij}, the timing of the entire problem can be reduced by a scaling factor, δ. The main idea is to compress the timing domain by a constant factor δ, and therefore simultaneously reduce the number of possible solutions by the same factor. For example, if δ=2, then each time-shifted value is reduced to $\lfloor \frac{D_{ij}}{\delta} \rfloor$. This results in the ILP formulation constructed on every 2 epochs instead of every epoch. The solution to the problem is re-scaled to gain an approximation of the true solution. Note that the scaling solution is not guaranteed to be optimal due to the approximation of the time-shift values, however it is a close approximation (within $\delta - 1$) to the optimal solution.

4.4 Heuristic

In this section an iterative improvement technique is presented for addressing large instances of the node assignment problem. While scaling of the ILP can be useful for addressing larger instances near optimally, the exponential growth in ILP runtime eventually mandates the need for a heuristic approach. The iterative improvement approach leverages on the ability of the ILP formulation to solve smaller instances fast. Specifically, the ILP formulation is used to solve sub-instances identified as potential iterative improvement moves.

The iterative improvement approach begins with a starting solution of sensors assigned to random sampling epochs in a round-robin fashion. With each sampling placement all corresponding prediction models are considered with the placement and are accounted for. If no benefit is obtained by the placement of a sensor sample at the randomly selected epoch, the sample is discarded. Once readings of all sensors at all epochs can be accounted for by a sample or a prediction model the initial solution is created.

For each iterative improvement move we leverage on two main insights. The first is that the selection of sensors for inclusion in the sub-instance ILP is crucial. If a sensor is overly often sampled, the ILP can find a more effective way to sample the sensor. However, if the sensor is under utilized it should increase its sampling occurrences in order to allow over utilized sensors to reduce its sampling. Secondly, there is a trade-off between the amount of improvement versus the amount of time spent searching for a move.

Initially, the initial solution will not be overly constrained or even well balanced. Therefore, there is room for rapid improvement of the solution. However, as the solution is refined, the amount of rapid improvement decreases, and additional search for moves is required. Therefore, as the iterative improvement approach progresses the number of sensors taken into consideration in the sub-instance ILP are increased.

The approach performs moves in a systematic fashion, first selecting each node individual (ordered by decreasing number of sampling assignments), then in pairs, triples, and so on. When formulating each sub-instance ILP, only the sensor considered for sampling replacement are considered variables of the ILP. All other sampling positions in the ILP become static by placing explicit constraints on their positions. Each move is incorporated by replacement of the sensors according to the new placement specified in the ILP solution. The approach has two types of stopping criteria. The first is the user specified runtime quota and the second is when no improvement is detected in the user specified attempted moves.

5 Experimental Analysis

In our experimental evaluation of the node assignment problem using the monotonic symmetric prediction models we used sensors of three modalities: temperature, humidity, and light taken from the Intel Berkeley dataset [19]. The sensors were sampled at 30 second intervals. All model were built using a of 3 days of data and evaluated using 8 different days.

In order to evaluate the effectiveness of the proposed models, our analysis was performed on ILP instances where dual sensor phase-shifted prediction and occasional calibration were used in situations where no optimization is conducted, where only single sensor prediction is used (the standard base case), where single sensor phase-shifted prediction was employed. The single sensor case is when each sample can only be predicted from other single sensor samples taken at the same time moment. In the single sensor phase-shifted case each sample can be predicted from other single sensor samples taken within the maximum time-shift allowed by the pre-specified prediction error value (TSMR model). In the final case, the addition of two phase-shifted sensor readings are incorporated for prediction (OCR, TSMR, and 2D SMR models).

Each of the cases is formulated using varied input to the ILP formulation. All of the ILP formulations were solved using the CPLEX solver with a maximum runtime of 5 minutes, but almost all instances ran within seconds. Three sets of L_1 errors were considered 2%, 3%, and 5%. In addition, three instances of the dataset were examined: all sensor nodes, a set of approximately two-thirds of the nodes, and a set of nodes from one third of the layout area.

In Table 2 we present the node assignment results for temperature, humidity, and light. In the first column we show the number of nodes, followed by the maximal L_1 prediction error allowed in prediction. Then for temperature and humidity, we shows the improvement for the single sensor simultaneous prediction over no optimization, followed by single time-shifted sensor prediction. The final column for the modality shows the improvement of the 2D SMR prediction models. We see that as the amount of error allowable is increasing, the savings decreases. This is due to the fact that with increased model prediction error a higher number of sensors can predict other sensors over long time-shifts. The final column of the table shows the results for light using the 2D SMR models along with the OCR and TSMR models. For light this case is the only feasible case since the error in prediction using a single sensor higher than allowed.

The most important result is that even for very low error, 2%, the node assignment approach is capable of performing 20 times better than the standard base case sleeping strategy, translating into 20 times longer lifetimes for the network with the single sensor prediction and two orders of magnitude better prediction when considering 2D sensor prediction models. For humidity we see the same patterns, occurring as we increase the allowable error. Humidity overall has lower improvement over the base case, because humidity shows a more complex statistical relationship. Nevertheless, the node assignment approach still was able to achieve 22-120 times more energy savings over the base case. Finally, even for light we were able to achieve a 2-19 times improvement.

6 Conclusion

We have studied techniques for intersensor modeling and their application to energy efficient data collection in sensor networks. We have developed an approach for calculating the lower bounds of prediction errors

		Temperature			Humidity			Light
#	L_1 Err	SMR	TSMR	2D SMR	SMR	TSMR	2D SMR	2D SMR
51	0.02	1	21	41	2	22	84	2
34	0.02	2	34	70	2	22	90	2
17	0.02	6	120	240	4	44	160	4
51	0.03	2	44	260	2	22	240	4
33	0.03	4	44	280	4	34	320	4
17	0.03	8	42	340	4	34	340	6
51	0.05	24	60	150	10	60	140	19
34	0.05	40	200	400	10	120	420	17
17	0.05	16	160	340	4	44	340	16

Table 2: Experimental results for Intel Berkeley dataset.

and use it for analysis of traditional modeling techniques. Motivated by this analysis, we have created a new modeling techniques that significantly improves prediction accuracy and the application domain. The techniques include trend identification, time shifting, occasional calibration, and symmetric monotonic regression for prediction from a single and pair of sensors. The new models are used in an ILP and scalable ILP-based heuristic approach to improve lifetimes of sensor networks by more than an order of magnitude in comparison with the best previously obtained results.

References

[1] M. A. Batalin, et al. Call and response: Experiments in sampling the environment. In *Sensys*, pages 25–38, 2004.

[2] G. Bril, R. Dykstra, C. Pillers, and T. Robertson. Isotonic regression in two independent variables. *Applied Statistics*, 33:352–357, 1984.

[3] H. Brunk. Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics*, 26:607–616, 1955.

[4] M. Coates. Distributed particle filters for sensor networks. In *IPSN*, pages 99–107, 2004.

[5] G. Cran. Amalgamation of the means in the case of simple ordering. *Applied Statistics*, 29:209–211, 1980.

[6] J. de Leeuw. Correctness of Kruskal's algorithms for monotone regression with ties. *Psychometrika*, 42:141–144, 1977.

[7] V. Delouille, R. Neelamani, and R. Baraniuk. Robust distributed estimation in sensor networks using the embedded polygons algorithm. In *IPSN*, pages 405–413, 2004.

[8] H. Dette and K. Pilz. A comparative study of monotone nonparametric kernel estimates. Technical Report 21, Universitt Dortmund, Sonderforschungsbereich, 2004.

[9] M. R. Garey and D. S. Johnson. *Computer and Intractability: A Guide to the theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, 1979.

[10] C. Guestrin, P. Bodi, R. Thibau, M. Paski, and S. Madde. Distributed regression: an efficient framework for modeling sensor network data. In *IPSN*, pages 1–10, 2004.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[12] A. Jain and E. Y. Chang. Adaptive sampling for sensor networks. In *DMSN*, pages 10–16, 2004.

[13] K. Kar, A. Krishnamurthy, and N. Jaggi. Dynamic node activation in networks of rechargeable sensors. In *Infocom*, 2005.

[14] F. Koushanfar, N. Taft, and M. Potkonjak. Sleeping coordination for comprehensive sensing using isotonic regression and domatic partitions. In *Infocom*, 2006.

[15] J. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129, 1964.

[16] H. Mukerjee. Monotone nonparametric regression. *Annals of Statistics*, 16:741–750, 1988.

[17] R. Nowak. Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE Transactions on Signal Processing*, 51(8):2245–2253, 2003.

[18] M. Paskin, C. Guestrin, and J. McFadden. A robust architecture for distributed inference in sensor networks. In *IPSN*, pages 55–62, 2005.

[19] M. Paskin, et al. A robust architecture for distributed inference in sensor networks. In *IPSN*, 2005.

[20] N. Sidiropoulos and R. Bro. Mathematical programming algorithms for regression-based nonlinear filtering in IR. *IEEE Transaction on Signal Processing*, 47(3):771–782, 1999.

[21] Q. Stout. Optimal algorithms for unimodal regression. *Computing Science and Statistics*, 32, 2000.

[22] A. Wang and A. Chandrakasan. Energy-efficient DSPs for wireless sensor networks. *IEEE Sig Proc Magazine*, 43(5):68–78, 2002.

[23] R. Willett, et al. Backcasting: adaptive sampling for sensor networks. In *IPSN*, pages 124–133, 2004.

[24] W. Ye, and et al. An energy-efficient MAC protocol for wireless sensor networks. In *Infocom*, pages 1567–1576, June 2002.