Unsupervised and Supervised Compression in Wireless Sensor Networks

Y. Le Borgne and G. Bontempi

Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

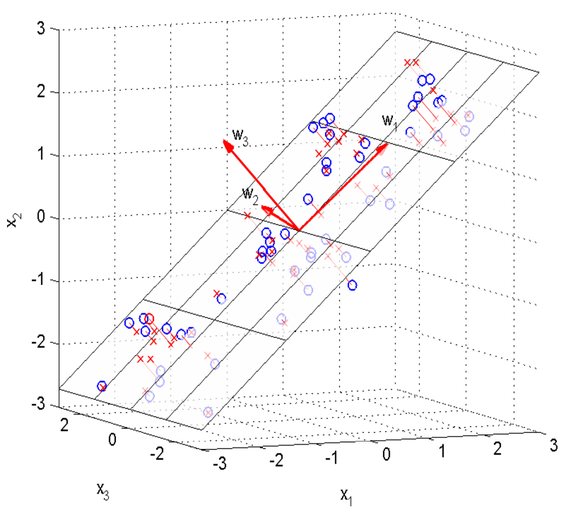
Abstract

*We show that the **Principal Component Analysis**, a compression method widely used in statistical analysis and image processing, can be efficiently implemented in a network of wireless sensors. The proposed scheme proves to be particularly suitable to sensor networks as it allows to reduce the network load while retaining a maximum amount of variance from sensor measurements. We present two operating modes, unsupervised and supervised, allowing (i) to extract a maximum of variance while **keeping the network load bounded**, and (ii) to reduce the network load while **keeping the approximation error bounded**, respectively. We assess the efficiency of the proposed approach in a realistic wireless sensor network deployment for temperature monitoring.*

1. Principal component analysis (PCA) with sensor data

Let a set of p sensors x_i, $1 \leq i \leq p$, sampling measurements $x_i[t] \in \mathbb{R}$ at regular time intervals. Let $\mathbf{x}[t] = (x_1[t], x_2[t], ..., x_p[t]) \in \mathbb{R}^p$ be the vector of measurements collected in the sensor field at every time instant t.

Given that measurements in $\mathbf{x}[t]$ are often correlated, the rationale of the approach is to determine a basis $\{\mathbf{w_k}\}$ for a subspace $\mathbb{R}^q$ of $\mathbb{R}^p$, $q \leq p$, that provides good approximations $\hat{\mathbf{x}}[t] = \sum_{k=1}^{q} \mathbf{w_k}\mathbf{w_k^T}\mathbf{x}[t]$ to the vectors $\mathbf{x}[t]$ over time.



- Three data sources $x_1[t]$, $x_2[t]$ and $x_3[t]$.
- $N = 50$ observations.
- The correlation between $x_1[t]$ and $x_2[t]$ is high.
- $x_3[t]$ measurements are drawn independently.

- Circles give the original observations.
- crosses their approximations on the two-dimensional subspace spanned by the two first principal components.
- The three blue vectors $\{w_1, w_2, w_3\}$ form the PC basis.

The optimization function is

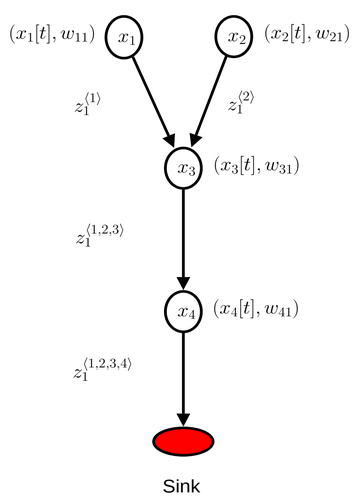
$$J_q(\mathbf{x}[t], \{\mathbf{w_k}\}) = \frac{1}{N}\sum_{t=1}^{N} ||\mathbf{x}[t] - \hat{\mathbf{x}}[\mathbf{t}]||^2 = \frac{1}{N}\sum_{t=1}^{N} ||\mathbf{x}[t] - \sum_{k=1}^{q} \mathbf{w_k}\mathbf{w_k^T}\mathbf{x}[t]||^2$$

The solution is given by the principal component (PC) basis. The set $\{\mathbf{w_k}\}$ is the set of the q eigenvectors of the covariance matrix $C_X = E(\mathbf{x}[t]\mathbf{x}[t]^T)$ whose eigenvalues are the highest [1].

2. Unsupervised and supervised compression

Unsupervised compression

Principal coordinate aggregation



We assume that the sensor nodes are connected by the means of a synchronized routing layer, such as the one proposed by the Tiny AGgregation (TAG) framework [2].

In the **unsupervised mode**, only the q coordinates

$$z_k[t] = w_k^T \mathbf{x}[t]$$

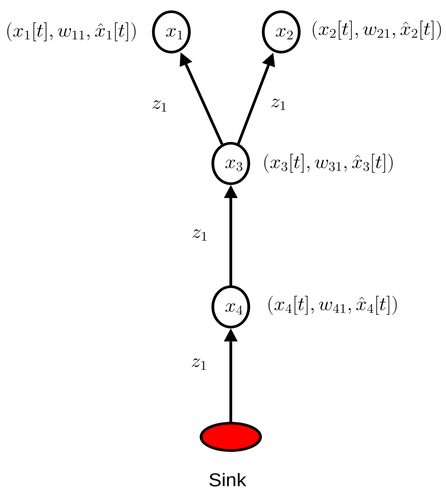
of $\mathbf{x}[t]$ in the PC basis are extracted from the network. The scalar product can be computed along the routing tree if each sensor x_i is aware of the i^{th} element w_{ik} of the q PCs. Approximations to the sensor measurements are then obtained at the sink by computing

$$\hat{\mathbf{x}}[t] = \sum_{k=1}^{q} \mathbf{w_k} z_k[t]$$

The variance retained equals the sum of the eigenvalues of the q eigenvectors.

Supervised compression

Coordinate feedback



The notation $z_k^{\{S\}} = \sum_{i \in S} x_i[t] * w_{ik}$ is used for detailing the progression of the scalar product along the routing tree. The set $\{S\}$ is the set of sensors whose measurements have already been aggregated.

In the **supervised mode**, the sink reinjects the q coordinates $z_k[t]$ in the network. Each sensor can obtain the approximation to its measurement by computing

$$\hat{x}_i[t] = \sum_{k=1}^{q} z_k[t] * w_{ik}$$

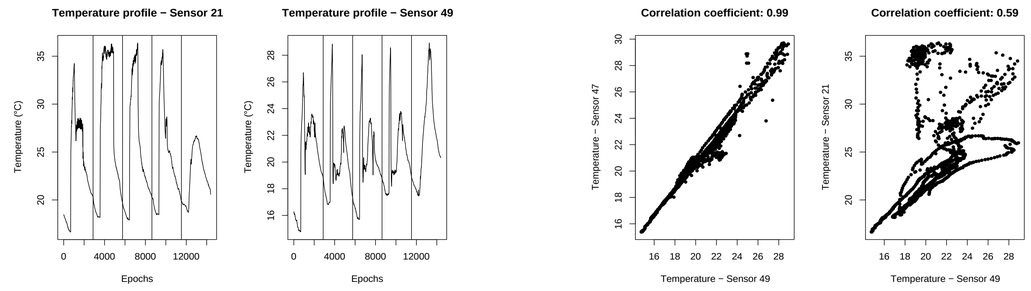
A user defined threshold ϵ can be used to notify the sink if

$$|\hat{x}_i[t] - x_i[t]| > \epsilon$$

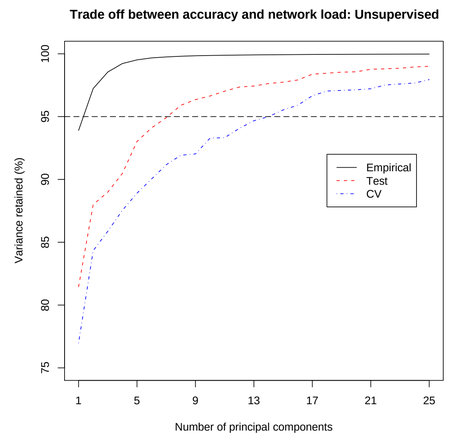
3. Experimental results

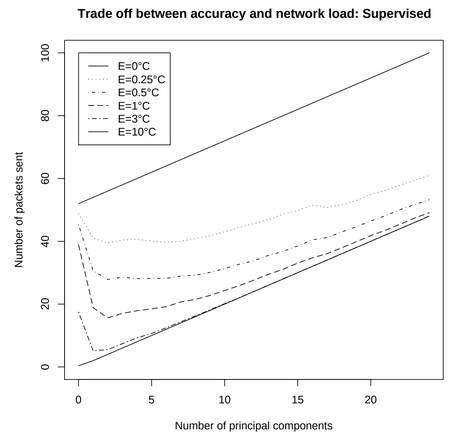
3.1 Data

Simulations were run on a subset of temperature measurements collected during a five day period at the Intel laboratory in Berkeley [3]. Measurements were taken every 31 seconds, at 52 different locations. See below for examples of temperature profiles and dependencies in this dataset. The first day was used for training.



3.2 Tradeoff network load - approximation error





In the **unsupervised mode**, the shared network load N depends on the number of PCs computed.

$$N(q) = q$$

Few PCs can approximate the whole set of measurements with high accuracy (95%).

$$P(q) = 100 * \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

Where λ_k is the eigenvalue of the k-th eigenvector.
Given that the eigenvectors are estimated, cross validation allows to get an estimate of the accuracy expected on new data.

In the **supervised mode**, the shared network load for a given accuracy ϵ depends on the number of PCs computed, and on the number of updates.

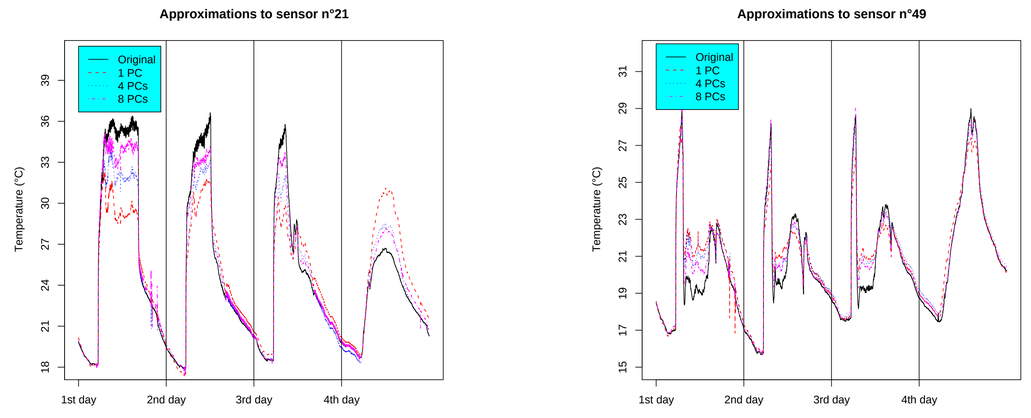
$$N(q, \epsilon, t) = 2q + U(q, \epsilon, t)$$

where $U(q, \epsilon, t)$ is the number of updates sent at time t given the required accuracy ϵ and the number of PC q.

The minimization of the network load implies a tradeoff between the number of PCs relied on and the required accuracy ϵ.

3.3 Accuracy of reconstruction

Examples of approximations obtained at two different locations, for 1, 4 and 8 PCs:



4. Future work

- Distribute the computation of the principal components in the network.
- Extend the compression to the temporal domain.
- Apply the coordinate extraction to event detection and event recognition tasks.

This research was supported by the European Community **COMP^{2}SYS** (MEST-CT-2004-505079) project.

References

[1] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, Germany, 2002.

[2] S.R. Madden, M.J. Franklin, J.M. Hellerstein, and W. Hong. TinyDB: an acquisitional query processing system for sensor networks. *ACM Transactions on Database Systems (TODS)*, 30(1):122–173, 2005.

[3] http://db.csail.mit.edu/labdata/labdata.html. Intel Lab Data webpage.