

Sparse Model Recovery Methods for Enhancing System Identification

Dimitri Kanevsky

— IBM T. J. Watson Research Center —

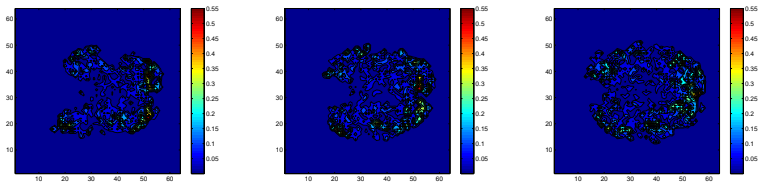
Joint work with: Avishy Carmi (Cambridge, UK), Pini Gurfil
(Technion, Israel)

June 3, 2009

- Introduction
- Brief compressive sensing (CS) overview
- Kalman Filtering Compressive Method
- Extended Compressed Sensing
- Extended Baum-Welch
- Isometric transformation
- Application of our CS methods to fMRI data classification
- Discussion

Introduction

- We describe novel optimization methods that were obtained in an IBM Exploratory Research Project: Pattern Recognition Techniques for High-dimensional and large scale data” .
- Specifically - we consider in our presentation sparse data.
- fMRI is example of sparse data



Overview of Compressive Sensing

- Compressive Sensing (CS) deals with a parameter estimation problem in which the observations are a linear projection onto a lower dimensional space

$$y = Hx + \zeta, \quad H \in \mathbb{R}^{m \times n}$$

- CS theory uses the convex relaxation methods. It is highly probable for some convex relaxation to yield an exact solution to the recovery problem if
 - The signal is sufficiently sparse
 - The sensing matrix obeys the restricted isometry property (RIP) at a certain level

Fundamental result in CS

- If the sensing matrix H' obeys *restricted isometry property* the l -RIP (i.e. subsets of H' are nearly orthogonal)

$$(1 - \delta_s) \|z\|_2 \leq \|H'_s z\|_2 \leq (1 + \delta_s) \|z\|_2$$

where H'_s is composed of $s \leq l$ columns from H' .

- while z is sparse enough possibly with

$$s = \mathcal{O}(m/\log(n/m)) \quad (1)$$

where $s = \#\{\text{supp}(z)\}$,

- then the solution of the combinatorial problem can almost always be obtained by solving the constrained convex optimization

$$\min \|\hat{z}\|_1 \quad \text{s.t.} \quad \sum_{i=1}^k \|y_i - H'_i \hat{z}\|_2^2 \leq \epsilon \quad (2)$$

- The constrained optimization

$$\min_{\hat{z}_k} E_{z_k|y^k} [\|z_k - \hat{z}_k\|_2^2] \quad \text{s.t.} \quad \|\hat{z}_k\|_1 \leq \epsilon' \quad (3)$$

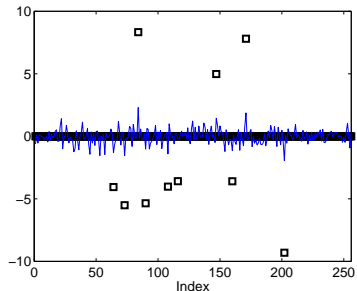
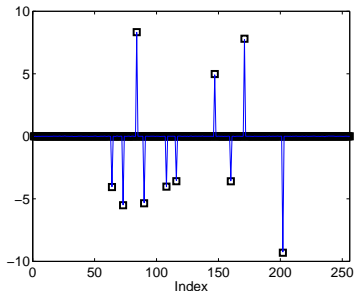
- The classical KF provides an estimate \hat{z}_k that is a solution to the unconstrained l_2 minimization problem

$$\min_{\hat{z}_k} E_{z_k|y^k} [\|z_k - \hat{z}_k\|_2^2]$$

- The constrained optimization problem can be solved in the framework of Kalman filtering using the pseudo-measurement (PM) technique. The inequality constraint $\|z_k\|_1 \leq \epsilon'$ is incorporated into the filtering process using a fictitious measurement $0 = \|z_k\|_1 - \epsilon'$, where ϵ' serves as a measurement noise.

$$0 = \bar{H}z_k - \epsilon', \quad \bar{H} := [\text{sign}(z_k(1)), \dots, \text{sign}(z_k(n))] \quad (4)$$

where $\text{sign}(z_k(i))$ denotes the sign function of the i th element of z_k (i.e., $\text{sign}(z_k(i)) = 1$ if $z_k(i) > 0$ and equals 0



A snapshot at $k = 20$ in a typical run of the CSKF-1 (a) using $N_\tau = 200$ PM iterations and of an ordinary KF (b). Showing the elements of the actual (squares) and estimated (lines) signals. Static case.

- We successfully applied novel compressive sensing KF (with pseudo-measurement) in fMRI analysis: prediction of mental illnesses (schizophrenia) and classification of “state of mind” according different activities using fMRI scan. In 3 different tasks classification methods with cs improved over classification without compression on average by 5%, 11% and 28% (absolutely).
- This work can be found in an IBM Research Report, A. Carmi, G. Cecchi, D. Kanevsky, B. Ramabhadran, I. Rish, “Classification via compressed Random Fields”, February 6, 2009, RC24740.

- Given a sufficiently smooth mapping $h: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m < n$ and some s -sparse vector $z \in \mathbb{R}^n$ that obey the observational relation $y_i = h(z) + \zeta$, $i = 1, \dots, k$, then to what extent and under what conditions can we recover z from y using the l_1 relaxation suggested by CS, i.e., by solving

$$\min \|\hat{z}\|_1 \quad \text{s.t.} \quad \sum_{i=1}^k \|y_i - h(\hat{z})\|_2^2 \leq \epsilon \quad (5)$$

- Now, consider a case where $z = z^* + \Delta z$ with sufficiently small $\|\Delta z\|_2$, then by taking the first-order Taylor expansion of $h(z)$ around z^* the RIP of the Jacobian $[\partial h / \partial z]$ computed locally at z^* , that is

$$(1-\delta) \|\Delta z\|_2 \leq \| [\partial h / \partial z]_{z^*} \Delta z + o(\|\Delta z\|_2^2) \|_2 \leq (1+\delta) \|\Delta z\|_2 \quad (6)$$

Local CS and Nonlinear Estimation: The CS-Embedded EKF

- The local CS idea implies that the l_1 relaxation can improve conventional nonlinear estimation methods that are based on linearization such as the EKF.
- The linearization around some predetermined nominal point, which is usually taken as the best up to date estimate, facilitates the application of a linear estimator (e.g., the KF) for reconstructing the perturbed state.
- The sensing matrix in this case is merely the Jacobian of the sensing function locally computed at the nominal point.
- The sensing function $h(z)$ maps the state onto a lower-dimensional space, then following the preceding argument
- It is expected that local CS will allow better recovery of sufficiently sparse perturbation Δz provided that $h(z)$ obeys the local RIP at a proper level.

Local CS and Nonlinear Estimation: The CS-Embedded EKF

- Implemented the local CS idea by amending the CSKF algorithms for nonlinear estimation. The slight modification consists of replacing the ordinary KF recursion with an EKF one while retaining the desired PM stage (corresponding to the l_1 norm).

- The following theorem is needed to extend EBW methods to sparse constraints.

Theorem

Let $F(z)$ be a function that is defined over $P = \{z_{ij} \geq 0, \sum_j z_{ij} = \sum_{j=1}^{j=m_i} z_{ij} = 1\}$. Let F be differentiable at $z \in P$. Let $c_{ij} = z_{ij} \frac{\partial}{\partial z_{ij}} F(z)$, and let $\hat{z} = \{\hat{z}_{ij}\} \neq z = \{z_{ij}\}$ where

$$\hat{z}_{ij} = \frac{c_{ij} + z_{ij}D}{\sum_i c_{ij} + D} \quad (7)$$

Then $F(\hat{z}) > F(z)$ for sufficiently large positive D and $F(\hat{z}) < F(z)$ for sufficiently small negative D .

- Consider the problem

$$\max_x F(x) \quad \text{s.t.} \quad \|x\|_1 \leq \epsilon \quad (8)$$

Now, using the dummy variable $x_0 \geq 0$ the above problem is rewritten as

$$\max_x F(x) \quad \text{s.t.} \quad \|x\|_1 + x_0 = \epsilon \quad (9)$$

- Further letting $v_i = x_i/\epsilon$, $i = 0, \dots, n$, and $F(x) = F(\{\epsilon v_i\}) = G(v)$ we may write

$$\max_v G(v) \quad \text{s.t.} \quad \|v\|_1 = 1 \quad (10)$$

Recognizing that

$$\|v\|_1 = \sum_i \sigma(v_i) v_i \quad (11)$$

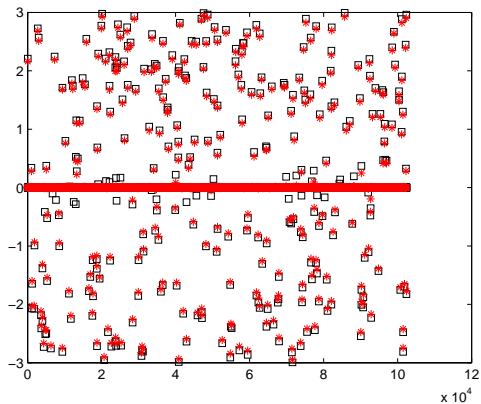
and

$$G(v) = G(\{\sigma(v_i) \sigma(v_i) v_i\}) = G(\sigma(v_i) z_i) = G(z) \quad (12)$$

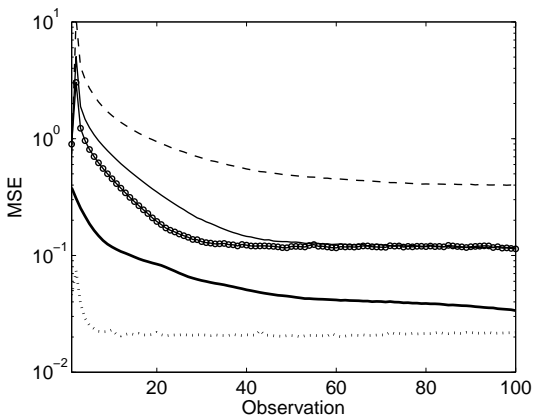
where $\sigma(v_i) = \text{sign}(v_i)$, and $z = \{z_i\} = \{\sigma(v_i) v_i\}$ allows writing an equivalent problem to (8) which takes the form

$$\max_z G(z) \quad \text{s.t.} \quad \sum z_i = 1, \quad z_i \geq 0 \quad (13)$$

Problems of the type (13) have an EBW based solutions that can be obtained by iterating (7).



Reconstruction via EBW a 100000 parameters noise signal



CS-EKF with l_1 norm (marked by circles), CS-EKF with the approximate l_0 norm (solid line), EBW (thick line), ordinary EKF that is unaware (dashed line) and aware (dotted line) of the actual support.

Isometric transformation

Suppose that $H \in \mathbb{R}^{m \times dm}$ for some $d, m \in \mathbb{N}$, and let

$$T = \text{diag}(H_1^{-1}P_1, \dots, H_d^{-1}P_d), \quad \text{Ker}(T) = \emptyset \quad (14)$$

where $H_i \in \mathbb{R}^{m \times m}$ and $P_i \in \mathbb{R}^{m \times m}$, $i = 1, \dots, d$ are the partitions of H and some RIP matrix $P \in \mathbb{R}^{m \times dm}$, respectively. Then there exists an orthogonal transformation $\hat{T} \in \mathbb{R}^{dm \times dm}$ and scalars $\alpha > 0$ and $\delta \in (0, 1)$ for which

$$(1 - \delta) \|x\|_2^2 \leq \|\alpha H \hat{T} x\|_2^2 \leq (1 + \delta) \|x\|_2^2$$

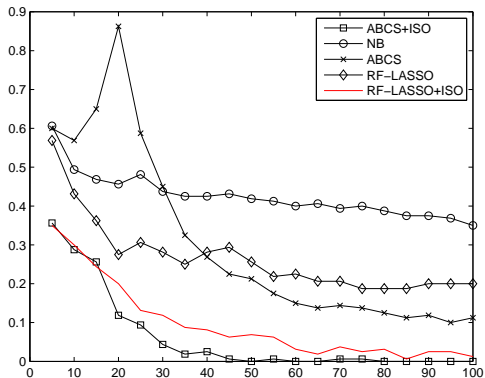
for m -block-sparse x . In particular

$$\hat{T} = \text{diag}(C_1 D_1^T, \dots, C_d D_d^T)$$

where $C_i \Lambda_i D_i^T$, $\Lambda_i = \text{diag}(\lambda_i^1, \dots, \lambda_i^m)$, $\lambda_i^1 \geq \lambda_i^2 \geq \dots \geq \lambda_i^m$ is the singular values decomposition (SVD) of $H_i^{-1}P_i$. Letting $\lambda_{\max} = \arg \max_{i \in [1, d]} \lambda_i^1$ and $\lambda_{\min} = \arg \min_{i \in [1, d]} \lambda_i^m$, the scaling factor can be approximated by

$$\alpha \approx \sqrt{2} (\lambda_{\min}^{-1} + \lambda_{\max}^{-1})^{-1/2}$$

Application of iso transformation to fMRI



ABCS means CSKF and RF means Random Field
Task: 8 objects classification

- How to apply this method in other domains that are not sparse and whose data matrix do not have RIP property
 - Need to identify factors that have significant impact on a process
 - Configuration of factors can be modeled by Markov Random Field
 - Configuration of significant factors are usually sparse
 - Apply isometric transformation to a matrix of configuration of factors to get RIP property

- The random matrix $H' \in \mathbb{R}^{72 \times 256}$ has its entries independently sampled from a zero-mean normal distribution with variance $1/72$.
- The vector z has 10 non-zero elements of which the locations are uniformly sampled over the integers in the interval $[1, 256]$. The values of the elements in the support of z are uniformly sampled between $[0.5, 1.5]$.
- The estimation performance based on 100 Monte Carlo runs of the EKF variants (i.e., with either the l_p , $p = 1, 0.5$ norms or the approximate l_0 norm) is shown in a plot
- For comparison we have depicted the performance of two ordinary EKF's (i.e., without a local CS stage) that were implemented, one of which is aware of the actual support of the signal.
- The local CS stage improves the estimation over regular EKF

Backup: Sparse Signal Recovery and Compressed Sensing (cont.)

- The deterministic case (z is a parameter vector). Can accurately recover z

$$\min \|\hat{z}\|_0 \quad \text{s.t.} \quad \sum_{i=1}^k \|y_i - H'\hat{z}\|_2^2 \leq \epsilon \quad (15)$$

- The stochastic case the sought-after optimal estimator satisfies

$$\min \|\hat{z}_k\|_0 \quad \text{s.t.} \quad E_{z_k|y^k} [\|z_k - \hat{z}_k\|_2^2] \leq \epsilon \quad (16)$$

$$\|v\|_0 := \#\{\text{supp}(v)\}.$$

- The above optimization problems are NP-hard and cannot be solved efficiently

- The property of the sensing function $h(z)$ is termed *Local RIP*. Similarly to the linear case, the level of the local RIP of $h(z)$ at z^* is determined according to the maximal sparseness degree s of the perturbation Δz for which (6) holds. Obviously, when considering the recovery of a sufficiently small and sparse Δz , CS can be applied where the Jacobian $[\partial h / \partial z]_{z^*}$ takes the role of the traditional sensing matrix. The l_1 relaxation would then have the form

$$\min \|\Delta \hat{z}\|_1 \quad \text{s.t.} \quad \sum_{i=1}^k \|y_i - h(z^* + \Delta \hat{z})\|_2^2 \leq \epsilon \quad (17)$$

where the accuracy of recovery would be related to the local RIP constant δ_s of the sensing function $h(z)$.

Backup: Sparse Signal Recovery and Compressed Sensing

- \mathbb{R}^n -valued random discrete-time process $\{x_k\}_{k=1}^{\infty}$
- sparse in some orthonormal sparsity basis $\psi \in \mathbb{R}^{n \times n}$

$$z_k = \psi^T x_k, \quad \#\{\text{supp}(z_k)\} < n \quad (18)$$

- The process x_k is measured by the \mathbb{R}^m -valued random process

$$y_k = Hx_k + \zeta_k = H'z_k + \zeta_k \quad (19)$$

where $\{\zeta_k\}_{k=1}^{\infty}$ is a zero-mean white Gaussian sequence with covariance $R_k > 0$, and $H := H'\psi^T \in \mathbb{R}^{m \times n}$.

- Letting $y^k := [y_1, \dots, y_k]$, our problem is defined as follows. We are interested in finding a y^k -measurable estimator, \hat{x}_k , that is optimal in some sense. Often, the sought after estimator is the one that minimize the mean square error (MSE) $E [\|x_k - \hat{x}_k\|_2^2]$.

- The same nonlinear problem was solved using the EBW. As before, the actual sparse signal is assumed to be normalized, i.e., $\|z\|_1 = 1$. The objective function used by the EBW is given by

$$G(z) = p(y | z) \propto \prod_{i=1}^k \exp \left\{ -\frac{1}{2} (y_i - h(z))^T R_i^{-1} (y_i - h(z)) \right\} \quad (20)$$

where R_i denotes the observation noise covariance.

- The results of this experiment are shown in a plot. EBW outperforms all other methods while exhibiting a rapid convergence towards the EKF that is aware of the signal support. This superiority over the CS-EKF can be related to the guaranteed convergence of the EBW (in this case z is defined over a probability domain), a property that essentially depends on tuning and initial conditions in the case of the EKF.

Backup: Numerical Study: Nonlinear Extensions

- Demonstrate the idea of application of extended CS (or local CS) to nonlinear estimation.
- Consider a recovery problem for the nonlinear observation model

$$y_i = h(z) + \zeta_i, \quad h(z) = H' \left[\text{diag}(z)^{1\frac{1}{2}} \mathbf{1} + az(j) \mathbf{1} \right] \quad (21)$$

where $\mathbf{1}$ denotes a vector of which all entries are 1's, a is some constant, and j is an arbitrary number between 1 and n .

- The Jacobian matrix of the sensing function $h(z)$:

$$\frac{\partial h}{\partial z} = 1 \frac{1}{2} H' \text{diag}(z)^{\frac{1}{2}} + a H' \text{diag}(\mathbf{e}_j) \quad (22)$$

where $\mathbf{e}_j \in \mathbb{R}^n$ has its j th entry equals one while all others are zero.

Backup: Comparative run

- Run 3 experiments using 3 different methods for 256 parameters
 - EBW norm. error 40
 - CS KF norm. error 49
 - dantzig (popular algorithm for lasso problem), norm. error 53