Sparsity via Convex Optimization

Problems and Algorithms

Ying Sun and Daniel P. Palomar

The Hong Kong University of Science and Technology (HKUST)

ELEC 5470 - Convex Optimization
Fall 2018-19, HKUST, Hong Kong

Outline of Lecture

1. Optimization with Sparsity
   - General Formulation
   - A Glance at Applications
2. Algorithms for Sparsity Problems
   - ℓ_1-Norm Heuristic
   - Interpretation of ℓ_1-Norm Heuristic
   - Iterative Reweighted ℓ_1-Norm Heuristic
3. Applications
   - Statistics and Data Analysis
   - Bioinformatics, Image Processing, and Computer Vision
   - Others

Outline

A World with Sparsity

- Many scenarios where sparsity exists:
  - Genetic mutation detection
  - Outlier detection
  - Computer vision
  - Data mining
  - Sudoku
- Question: What can we do with sparsity as a prior information?
- Answer: Enforce sparsity via cardinality proxies, i.e., ℓ_1-norm.

General Formulation

- Problem:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \\ & \text{card}(\mathbf{x}) \leq k \end{array}$$

where cardinality is defined as $\text{card}(\mathbf{x}) = \sum_i 1_{\{x_i \neq 0\}}$, i.e., number of nonzero elements in $\mathbf{x}$, and $\text{supp}(\mathbf{x})$ is defined as the positions with nonzero values.

- Variations:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \text{card}(\mathbf{x}) \\ \text{subject to} & f(\mathbf{x}) \leq \varepsilon \\ & \mathbf{x} \in \mathscr{C} \end{array} \qquad \begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) + \lambda \text{card}(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

General Formulation

- Problem:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \\ & \text{card}(\mathbf{x}) \leq k \end{array}$$

  where cardinality is defined as $\text{card}(\mathbf{x}) = \sum_i 1_{\{x_i \neq 0\}}$, i.e., number of nonzero elements in $\mathbf{x}$, and $\text{supp}(\mathbf{x})$ is defined as the positions with nonzero values.

- Variations:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \text{card}(\mathbf{x}) \\ \text{subject to} & f(\mathbf{x}) \leq \varepsilon \\ & \mathbf{x} \in \mathscr{C} \end{array}$$

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) + \lambda \text{card}(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

A Glance at Applications

- Statistics and data analysis
  - Compressed sensing
  - Estimation with outliers
  - Piecewise constant fitting
  - Piecewise linear fitting
  - Feature selection
- Optimization modeling
  - Minimum number of violations
- Bioinformatics
  - Medical testing design
- Image processing and computer vision
  - Robust face recognition

Combinatorial Nature

- Despite widely applicable areas, solving cardinality constrained problems is not a trivial work.
- Most of cardinality related problems are NP-hard:
  - given $\text{supp}(\mathbf{x})$ we can solve the problem efficiently, but the choice of $\text{supp}(\mathbf{x})$ grows exponentially with $\dim(\mathbf{x})$.
- What can we do?
  - Exhaustive Search: doable only if the variable dimension is small
  - Branch and Bound: in the worst case its complexity is of the same order as exhaustive search
  - Convex Relaxation.

Outline

ℓ_1-Norm Heuristic

- The cardinality operator $\text{card}\,(\mathbf{x})$ is nonnonvex.
- Usually referred to as ℓ_0-norm: $\|\mathbf{x}\|_0$ (although it is not a norm).
- Instead of using the ℓ_0-norm, use ℓ_1-norm, i.e., $\text{card}\,(\mathbf{x}) = \|\mathbf{x}\|_0 \longleftrightarrow \gamma\|\mathbf{x}\|_1$ with γ being a tuning parameter:
  - often called in literature ℓ_1-norm regularization, ℓ_1 penalty, shrinkage, etc.
  - convex relaxation of cardinality constraint
  - convex envelope of ℓ_0-norm
  - in some cases, relaxation is not tight, but works well in practice.

Polishing After Application of ℓ_1-Norm Heuristic

- After the approximation of the cardinality operator with the ℓ_1-norm $\gamma\|\mathbf{x}\|_1$, we will obtain a solution where some elements are very small, almost zero.
- Fix the sparsity pattern by setting the very small elements to zero.
- Re-solve the (now convex) optimization problem with the fixed sparsity pattern to obtain the final (heuristic) solution.

Variations of ℓ_1-Norm

- The ℓ_1-norm proxy of ℓ_0-norm seeks a trade-off between sparsity and problem tractability.
- More sophisticated versions include:
  - Weighted ℓ_1-norm: $\sum_i w_i |x_i|$
  - Asymmetric weighted ℓ_1-norm: $\sum_i w_i (x_i)^+ + \sum_i v_i (x_i)^-$, where $\mathbf{w}$, $\mathbf{v}$ are positive weights.

Interpretation of ℓ_1-Norm Heuristic as Convex Relaxation

- Start with the original formulation (and a bound on $\mathbf{x}$)

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \text{card}\left(\mathbf{x}\right) \\ \text{subject to} & \mathbf{x} \in \mathscr{C}, \qquad \|\mathbf{x}\|_\infty \leq R. \end{array}$$

- Rewrite it as the mixed Boolean convex problem

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & \mathbf{1}^T\mathbf{z} \\ \text{subject to} & |x_i| \leq Rz_i, \quad z_i \in \{0,1\}, \quad i = 1,\cdots,n \\ & \mathbf{x} \in \mathscr{C}. \end{array}$$

Interpretation of ℓ_1-Norm Heuristic as Convex Relaxation

- Start with the original formulation (and a bound on $\mathbf{x}$)

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \text{card}(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{C}, \qquad \|\mathbf{x}\|_\infty \leq R. \end{array}$$

- Rewrite it as the mixed Boolean convex problem

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & \mathbf{1}^T\mathbf{z} \\ \text{subject to} & |x_i| \leq Rz_i, \quad z_i \in \{0,1\}, \quad i = 1, \cdots, n \\ & \mathbf{x} \in \mathcal{C}. \end{array}$$

Interpretation of ℓ_1-Norm Heuristic as Convex Relaxation

- Now relax $z_i \in \{0,1\}$ to $z_i \in [0,1]$ to obtain

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & \mathbf{1}^T\mathbf{z} \\ \text{subject to} & |x_i| \leq Rz_i, \quad 0 \leq z_i \leq 1, \quad i = 1,\ldots,n \\ & \mathbf{x} \in \mathscr{C}. \end{array}$$

- Since the optimal solution of the problem above satisfies $|x_i| = Rz_i$, the problem is equivalent to

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & (1/R)\,\|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

which is the ℓ_1-norm heuristic and provides a lower bound on the original problem.

Interpretation of ℓ_1-Norm Heuristic as Convex Relaxation

- Now relax $z_i \in \{0,1\}$ to $z_i \in [0,1]$ to obtain

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & \mathbf{1}^T\mathbf{z} \\ \text{subject to} & |x_i| \leq Rz_i, \quad 0 \leq z_i \leq 1, \quad i = 1,\ldots,n \\ & \mathbf{x} \in \mathscr{C}. \end{array}$$

- Since the optimal solution of the problem above satisfies $|x_i| = Rz_i$, the problem is equivalent to

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & (1/R)\,\|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

which is the ℓ_1-norm heuristic and provides a lower bound on the original problem.

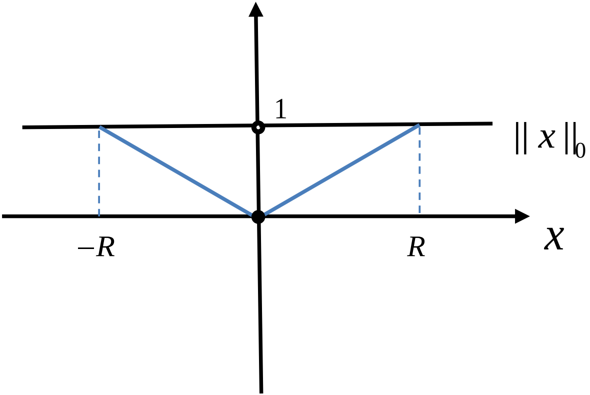
Interpretation of ℓ_1-Norm Heuristic via Convex Envelope

- The convex envelope of a function f on set $\mathscr{C}$ is the largest convex function that is an underestimator of f on $\mathscr{C}$.
- For x scalar, $|x|$ is the convex envelope of $\operatorname{card}(x)$ on $[-1, 1]$.
- For $\mathbf{x} \in \mathrm{R}^m$, $(1/R)\|\mathbf{x}\|_1$ is the convex envelope of $\operatorname{card}(\mathbf{x})$ on $\{\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq R\}$.
- Now suppose we know lower and upper bounds on x_i over $\mathscr{C}$, $l_i \leq x_i \leq u_i$ (can be found by solving $2n$ convex problems). Then, assuming $l_i < 0$, $u_i > 0$ (otherwise $\operatorname{card}(x_i) = 1$), the convex envelope is

$$\sum_{i=1}^{n} \left(\frac{(x_i)^+}{u_i} + \frac{(x_i)^-}{-l_i} \right).$$

Interpretation of ℓ_1-Norm Heuristic via Convex Envelope

- Convex envelope of ℓ_0-norm on interval $[-R, R]$:



Iterative Reweighted ℓ_1-Norm Heuristic

Algorithm

set $\mathbf{w} = \mathbf{1}$ **repeat**

minimize$_{\mathbf{x}}$ $\|\text{diag}(\mathbf{w})\,\mathbf{x}\|_1$ subject to $\mathbf{x} \in \mathscr{C}$

$w_i = 1/(\varepsilon + |x_i|)$

until convergence to local point

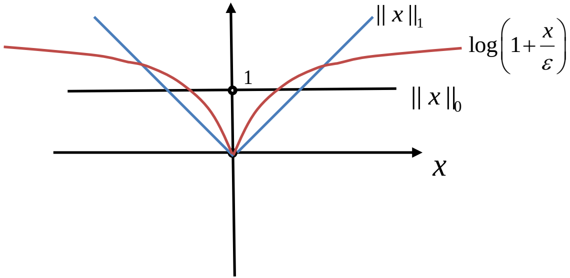
- Interpretation:
  - the first iteration is the basic ℓ_1-norm heuristic
  - then, for the next iteration:
    - for small $|x_i|$, the weight increases (enforcing even smaller $|x_i|$)
    - for large $|x_i|$, the weight decreases (allowing it to be larger if necessary)
- Typically, it converges in 5 of fewer steps with some modest improvement.

Derivation of Iterative Reweighted ℓ_1-Norm Heuristic

- First of all, "w.l.o.g.", we can assume $\mathbf{x} \geq \mathbf{0}$ (if not, just write $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$ with $\mathbf{x}^+, \mathbf{x}^- \geq \mathbf{0}$ and use $\tilde{\mathbf{x}} = (\mathbf{x}^+, \mathbf{x}^-)$).
- Then, we can use the (nonconvex) approximation

$$\text{card}\,(z) \approx \log\,(1 + z/\varepsilon)$$

where $\varepsilon > 0$ and $z \geq 0$.



Derivation of Iterative Reweighted ℓ_1-Norm Heuristic

- Using this approximation, we get the nonconvex problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{i=1}^{n} \log\left(1 + x_i/\varepsilon\right) \\ \text{subject to} & \mathbf{x} \in \mathscr{C}, \quad \mathbf{x} \geq \mathbf{0}. \end{array}$$

- This problem is then solved by an iterative convex approximation:
  - approximate nonconvex problem around current point $\mathbf{x}^{(k)}$ with a convex problem (which in this case will be a linear approximation of the log function)
  - solve approximated convex problem to get next point $\mathbf{x}^{(k+1)}$
  - repeat until convergence to get a local solution.

Derivation of Iterative Reweighted ℓ_1-Norm Heuristic

- To approximate the nonconvex problem, linearize the objective at current point $\mathbf{x}^{(k)}$

$$\sum_{i=1}^{n} \log\left(1 + x_i/\varepsilon\right) \approx \sum_{i=1}^{n} \log\left(1 + x_i^{(k)}/\varepsilon\right) + \sum_{i=1}^{n} \frac{x_i - x_i^{(k)}}{\varepsilon + x_i^{(k)}}$$

- Solve the resulting convex problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{i=1}^{n} \frac{x_i - x_i^{(k)}}{\varepsilon + x_i^{(k)}} \\ \text{subject to} & \mathbf{x} \in \mathscr{C}, \quad \mathbf{x} \geq \mathbf{0} \end{array}$$

or, equivalently,

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{i=1}^{n} w_i x_i \\ \text{subject to} & \mathbf{x} \in \mathscr{C}, \quad \mathbf{x} \geq \mathbf{0} \end{array}$$

where $w_i = 1/(\varepsilon + x_i^{(k)})$.

Interpretation by Majorization-Minimization

- Consider the objective function $f(\mathbf{x})$ that we want to minimize
- The Majorization Minimization algorithm [1, 2]:
  - finds a function g that majorizes f in the kth step in the following sense:
    - $g\left(\mathbf{x}^{k-1}|\mathbf{x}^{k-1}\right) = f\left(\mathbf{x}^{k-1}\right)$;
    - $\nabla g\left(\mathbf{x}^{k-1}|\mathbf{x}^{k-1}\right) = \nabla f\left(\mathbf{x}^{k-1}\right)$;
    - $g\left(\mathbf{x}|\mathbf{x}^{k-1}\right) \geq f(\mathbf{x})$;
  - then solves the majorized problem: $\mathbf{x}^k = \arg\min g\left(\mathbf{x}|\mathbf{x}^{k-1}\right)$.
- In our particular problem, since the log function is concave monotone increasing, the linearized objective majorizes f.

Outline

Compressed Sensing I

- Consider the following linear equations: $\mathbf{y} = \mathbf{A}\mathbf{x}$, with $\mathbf{A} \in \mathrm{R}^{m \times n}$. By fundamental linear algebra:
  - if $m \geq n$ and $\mathbf{A}$ is full rank, the system admits a unique solution or has no solution
  - if $m < n$, the problem is ill-posed and have infinitely many solutions $\hat{\mathbf{x}}$.
- Classical solution: $\hat{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_2$, closed form solution $\hat{\mathbf{x}} = \mathbf{A}^{\dagger}\mathbf{y}$.
- However in many applications, $\hat{\mathbf{x}}$ is not good and $\mathbf{x}$ is required to be sparse.

Compressed Sensing II

- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^\star = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing II

- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^{\star} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing II

- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^{\star} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing II

- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^{\star} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing II

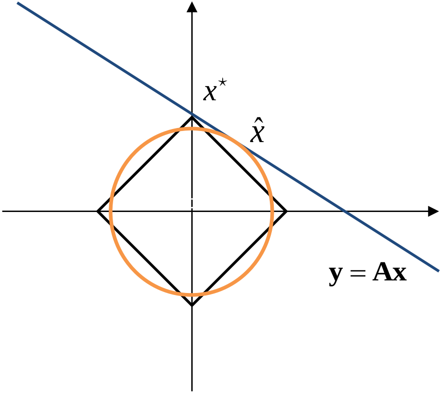
- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^{\star} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing II

- Question: How to incorporate sparsity as prior information?
- Answer: $\mathbf{x}^\star = \arg\min_{\mathbf{y}=\mathbf{Ax}} \|\mathbf{x}\|_0$.
- Question: Any efficient algorithm for ℓ_0-norm minimization problem?
- Answer: Relax ℓ_0-norm by its convex envelope, i.e., $\tilde{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{Ax}} \|\mathbf{x}\|_1$.
- Question: Under what condition is the relaxation tight?
- Answer: Roughly speaking, measurement matrix $\mathbf{A}$ is required to be sufficiently "incoherent" (i.e., number of measurements $(\dim(\mathbf{y}))$ greater than certain threshold). Not going to be covered in this course, refer to compressed sensing literature.

Compressed Sensing III

- Illustration in two dimensions with exact recovery:



Estimation with Outliers

- Consider measurements $y_i = \mathbf{a}_i^T \mathbf{x} + v_i, \quad i = 1, \ldots, m$ under Gaussian noise $v_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

- In practice, however, we have *outliers*: incorrect measurements for some unknown and expected reasons. This can be modeled as

$$y_i = \mathbf{a}_i^T \mathbf{x} + v_i + w_i, \quad i = 1, \ldots, m$$

where the only assumption on the outlier error $\mathbf{w}$ is sparsity: $\text{card}\left(\mathbf{w}\right) \leq k$.

- Problem formulation that takes into account k possible outliers:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{w}}{\text{minimize}} & \left\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{w}\right\|_2 \\ \text{subject to} & \text{card}\left(\mathbf{w}\right) \leq k \end{array}.$$

Estimation with Outliers

- Consider measurements $y_i = \mathbf{a}_i^T \mathbf{x} + v_i, \quad i = 1, \ldots, m$ under Gaussian noise $v_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

- In practice, however, we have *outliers*: incorrect measurements for some unknown and expected reasons. This can be modeled as

$$y_i = \mathbf{a}_i^T \mathbf{x} + v_i + w_i, \quad i = 1, \ldots, m$$

where the only assumption on the outlier error $\mathbf{w}$ is sparsity: $\mathrm{card}\left(\mathbf{w}\right) \leq k$.

- Problem formulation that takes into account k possible outliers:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{w}}{\text{minimize}} & \left\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{w}\right\|_2 \\ \text{subject to} & \mathrm{card}\left(\mathbf{w}\right) \leq k \end{array}.$$

Estimation with Outliers

- Consider measurements $y_i = \mathbf{a}_i^T \mathbf{x} + v_i, \quad i = 1, \ldots, m$ under Gaussian noise $v_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

- In practice, however, we have *outliers*: incorrect measurements for some unknown and expected reasons. This can be modeled as

$$y_i = \mathbf{a}_i^T \mathbf{x} + v_i + w_i, \quad i = 1, \ldots, m$$

  where the only assumption on the outlier error $\mathbf{w}$ is sparsity: $\mathsf{card}\left(\mathbf{w}\right) \leq k$.

- Problem formulation that takes into account k possible outliers:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{w}}{\text{minimize}} & \|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{w}\|_2 \\ \text{subject to} & \mathsf{card}\left(\mathbf{w}\right) \leq k \end{array} .$$

Piecewise Constant Fitting

- Problem: fit corrupted $\mathbf{x}_{\text{cor}}$ by a piecewise constant signal $\hat{\mathbf{x}}$ with k or fewer jumps.
- Convex if jump locations are known, but not otherwise.
- Property: $\hat{\mathbf{x}}$ piecewise constant with $\leq k$ jumps $\Longleftrightarrow$ card $(\mathbf{D}\hat{\mathbf{x}}) \leq k$, where

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathrm{R}^{(n-1)\times n}.$$

- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\|_2 \\ \text{subject to} & \text{card}(\mathbf{D}\hat{\mathbf{x}}) \leq k. \end{array}$$

Piecewise Constant Fitting

- Problem: fit corrupted $\mathbf{x}_{\text{cor}}$ by a piecewise constant signal $\hat{\mathbf{x}}$ with k or fewer jumps.
- Convex if jump locations are known, but not otherwise.

- Property: $\hat{\mathbf{x}}$ piecewise constant with $\leq k$ jumps $\iff$ $\text{card}\left(\mathbf{D}\hat{\mathbf{x}}\right) \leq k$, where

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathsf{R}^{(n-1)\times n}.$$

- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \left\|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\right\|_2 \\ \text{subject to} & \text{card}\left(\mathbf{D}\hat{\mathbf{x}}\right) \leq k. \end{array}$$

Piecewise Constant Fitting

- Problem: fit corrupted $\mathbf{x}_{\text{cor}}$ by a piecewise constant signal $\hat{\mathbf{x}}$ with k or fewer jumps.
- Convex if jump locations are known, but not otherwise.

- Property: $\hat{\mathbf{x}}$ piecewise constant with $\leq k$ jumps $\Longleftrightarrow$ card$\left(\mathbf{D}\hat{\mathbf{x}}\right) \leq k$, where

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathsf{R}^{(n-1)\times n}.$$

- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \left\|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\right\|_2 \\ \text{subject to} & \text{card}\left(\mathbf{D}\hat{\mathbf{x}}\right) \leq k. \end{array}$$

Total Variation Reconstruction

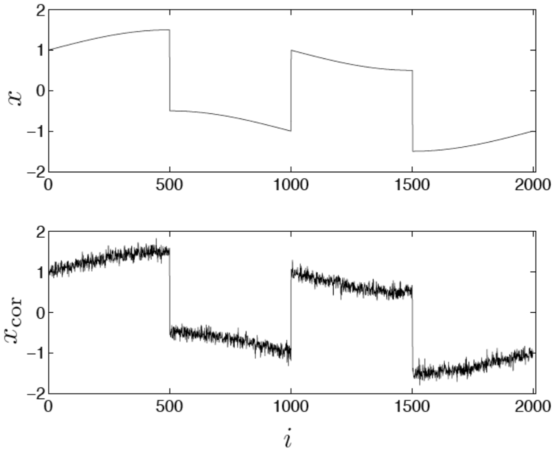
- The total variation (TV) reconstruction is just another name for the piecewise constant fitting.
- Problem: given a corrupted signal $\mathbf{x}_{\text{cor}} = \mathbf{x} + \mathbf{n}$, recover the original one $\mathbf{x}$.
- The trick is the assumption that original signal $\mathbf{x}$ is smooth (except some occasional jumps), whereas noise $\mathbf{n}$ is not smooth.
- Problem formulation:

$$\underset{\hat{\mathbf{x}}}{\text{minimize}} \quad \|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\|_2 + \gamma \|\mathbf{D}\hat{\mathbf{x}}\|_1$$

- Widely used in signal processing and image processing.
- The term $\|\mathbf{D}\hat{\mathbf{x}}\|_1$ is called total variation of signal $\hat{\mathbf{x}}$.

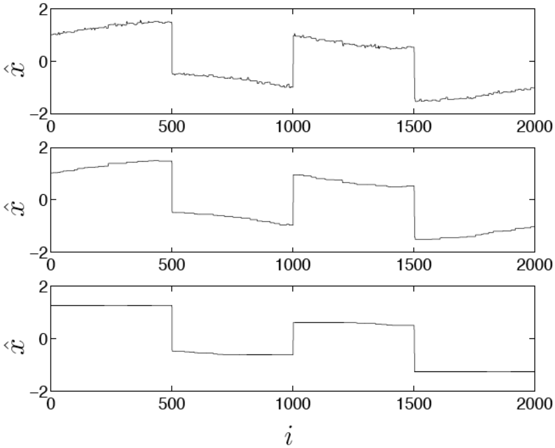
Total Variation Reconstruction: Numerical Example

- Consider the original and corrupted signals ($n = 2000$):



Total Variation Reconstruction: Numerical Example

- The total variation reconstruction is (for three values of γ)



Piecewise Linear Fitting

- Problem: fit corrupted $\mathbf{x}_{\text{cor}}$ by a piecewise linear signal $\hat{\mathbf{x}}$ with k or fewer kinks.

- The derivative of a piecewise linear signal $\mathbf{D}\hat{\mathbf{x}}$ is piecewise constant, so the second derivative $\nabla\hat{\mathbf{x}}$ is sparse.

- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\|_2 \\ \text{subject to} & \text{card}\left(\nabla\hat{\mathbf{x}}\right) \leq k \end{array}$$

where

$$\nabla = \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix}.$$

Piecewise Linear Fitting

- Problem: fit corrupted $\mathbf{x}_{\mathsf{cor}}$ by a piecewise linear signal $\hat{\mathbf{x}}$ with k or fewer kinks.
- The derivative of a piecewise linear signal $\mathbf{D}\hat{\mathbf{x}}$ is piecewise constant, so the second derivative $\nabla\hat{\mathbf{x}}$ is sparse.
- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \|\hat{\mathbf{x}} - \mathbf{x}_{\mathsf{cor}}\|_2 \\ \text{subject to} & \operatorname{card}(\nabla\hat{\mathbf{x}}) \le k \end{array}$$

where

$$\nabla = \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix}.$$

Piecewise Linear Fitting

- Problem: fit corrupted $\mathbf{x}_{\text{cor}}$ by a piecewise linear signal $\hat{\mathbf{x}}$ with k or fewer kinks.
- The derivative of a piecewise linear signal $\mathbf{D}\hat{\mathbf{x}}$ is piecewise constant, so the second derivative $\nabla\hat{\mathbf{x}}$ is sparse.
- Problem formulation:

$$\begin{array}{ll} \underset{\hat{\mathbf{x}}}{\text{minimize}} & \|\hat{\mathbf{x}} - \mathbf{x}_{\text{cor}}\|_2 \\ \text{subject to} & \text{card}\left(\nabla\hat{\mathbf{x}}\right) \leq k \end{array}$$

where

$$\nabla = \begin{bmatrix} -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \end{bmatrix}.$$

Feature Selection

- Problem: fit vector $y \in \mathbb{R}$ as a linear combination of k regressors (chosen from p possible regressors):

$$\begin{array}{ll} \underset{\beta}{\text{minimize}} & \left\|\mathbf{y} - \mathbf{X}^T\beta\right\|_2^2 \\ \text{subject to} & \text{card}\left(\beta\right) \leq k. \end{array}$$

- The solution chooses subset of k regressors that best fit y (role of expert).
- In principle, this could be solved by trying all $\binom{p}{k}$ choices, but not practical for large n.
- Variations:
  - minimize $\text{card}\left(\beta\right)$ subject to $\left\|\mathbf{y} - \mathbf{X}^T\beta\right\|_2^2$
  - minimize $\left\|\mathbf{y} - \mathbf{X}^T\beta\right\|_2^2 + \lambda \text{card}\left(\beta\right)$.

LASSO

- Relaxing the cardinality constraint in the objective, we get the famous LASSO regression (least absolute shrinkage and selection operator) [Tibshirani'96]:
  - $\hat{\beta}_{LASSO} = \arg\min \left\|\mathbf{y} - \mathbf{X}^T \beta\right\|_2^2 + \gamma \|\beta\|_1$
  - biased but more stable estimator (bias variance tradeoff)
  - results in sparse β since ℓ_1-norm ball is pointy
  - interpretable parsimonious model, variable selection.
- Extensions:
  - Fused LASSO [Tibshirani-etal'2005]
  - Group LASSO [Yuan-Lin'2006].

Coordinate Descent Algorithm for LASSO

- LASSO is a QP and can be solved efficiently with a QP solver.
- Problem: when N is extremely large, a universally applicable convex programming algorithm is no longer satisfactory.
- Solution: Seeking problem specific structure to speed up and beat the Newton type method [Friedman-etal'07].
- Consider LASSO with univariate predictor, i.e., x is a scalar. It has the closed-form solution:
  Threshold least square: $\hat{\beta}_{LASSO} = \operatorname{sign}\left(\hat{\beta}_{OLS}\right)\left(\left|\hat{\beta}_{OLS}\right| - 2\gamma\right)^{+}$.

Coordinate Descent Algorithm for LASSO

Coordinate Descent for LASSO

Initialize β_0, set $k, r = 1$ **repeat**

repeat

$$\beta_r^k = \arg\min \left\| \mathbf{y} - \mathbf{X}_{-r}^T \beta_{-r}^k - \mathbf{X}_r^T \beta_r \right\|_2^2 + \gamma \|\beta_r\|_1$$

$$r = r+1,\ \beta^k = \left(\beta_1^k, \ldots, \beta_r^k, \beta_{r+1}^{k-1}, \ldots, \beta_p^{k-1}\right)$$

until $r = p$

$k = k+1,\ r = 1$

until convergence

- Faster than calling off-the-shelf convex problem solver.

Minimum Number of Violations

- Consider a set of convex inequalities

$$f_1(\mathbf{x}) \le 0, \ldots, f_m(\mathbf{x}) \le 0, \qquad \mathbf{x} \in \mathscr{C}.$$

- Determining whether they are feasible or not is easy: convex feasibility problem. But what if they are infeasible?

- Problem formulation to find the minimum number of violated inequalities:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{t}}{\text{minimize}} & \text{card}(\mathbf{t}) \\ \text{subject to} & f_i(\mathbf{x}) \le t_i, \qquad i = 1, \ldots, m \\ & \mathbf{x} \in \mathscr{C}, \quad \mathbf{t} \ge \mathbf{0}. \end{array}$$

Minimum Number of Violations

- Consider a set of convex inequalities

$$f_1(\mathbf{x}) \leq 0, \ldots, f_m(\mathbf{x}) \leq 0, \qquad \mathbf{x} \in \mathscr{C}.$$

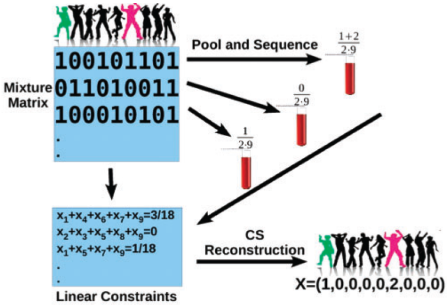
- Determining whether they are feasible or not is easy: convex feasibility problem. But what if they are infeasible?

- Problem formulation to find the minimum number of violated inequalities:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{t}}{\text{minimize}} & \text{card}(\mathbf{t}) \\ \text{subject to} & f_i(\mathbf{x}) \leq t_i, \qquad i = 1, \ldots, m \\ & \mathbf{x} \in \mathscr{C}, \quad \mathbf{t} \geq \mathbf{0}. \end{array}$$

Rare Allele Identification in Medical Testing I

- Problem: reconstruct the genotypes of N individuals at a specific locus. N is a large number and DNA sequencing is expensive.
- Solution: pool blood sample of multiple individuals in a single DNA sequencing experiment [7].



Rare Allele Identification in Medical Testing II

- Test procedure:
  - Sequence DNA fragments of sample pools instead of each individual.
  - Reads of the fragments of DNA of each sample pool are mapped back to the reference genome.
- Genotype vector $\mathbf{x} \in \{0, 1, 2\}^N$, x_i for the genotype of the ith individual at a specific locus:
  - Reference allele AA is coded as 0;
  - Heterozygous allele Aa is coded as 1;
  - Homozygous alternative allele aa is coded as 2.
- Genetic mutation is rare $\Longleftrightarrow$ $\mathbf{x}$ is a sparse vector.

Rare Allele Identification in Medical Testing III

- Bernoulli sensing matrix $\mathbf{M}$:
  - $M_{ij} \in \{0, 1\}$: whether individual j's blood sample is included in the ith experiment or not
  - $\mathbf{M}_{i,:}\mathbf{x}$ is the number of a alleles (rare alleles)
  - $2\sum_{j=1}^{N} M_{ij}$ is the number of alleles (each person has two)
  - normalized sensing matrix (by the number of people in a test) $\hat{\mathbf{M}}$: $\hat{M}_{ij} = \frac{M_{ij}}{\sum_{j=1}^{N} M_{ij}}$
  - proportion of rare alleles: $\mathbf{M}_{i,:}\mathbf{x} / \left(2\sum_{j=1}^{N} M_{ij}\right) = \frac{1}{2}\hat{\mathbf{M}}_{i,:}\mathbf{x}$
- Test output:
  - $\mathbf{z}$: number of reads containing rare allele a.
  - r: total number of reads covering locus of interest in each pool.

Rare Allele Identification in Medical Testing IV

- Problem formulation:

$$\begin{array}{ll} \underset{\mathbf{x}\in\{0,1,2\}^N}{\text{minimize}} & \|\mathbf{x}\|_0 \\ \text{subject to} & \left\|\frac{1}{2}\hat{\mathbf{M}}\mathbf{x} - \frac{\mathbf{z}}{r}\right\|_2 \leq \varepsilon \end{array}$$

- Relaxation:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_1 \\ \text{subject to} & \left\|\frac{1}{2}\hat{\mathbf{M}}\mathbf{x} - \frac{\mathbf{z}}{r}\right\|_2 \leq \varepsilon \end{array}$$

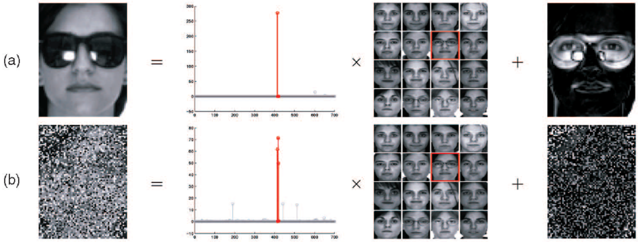
- Heuristic post-processing: rounding $\mathbf{x}$ to integer value.

Rare Allele Identification in Medical Testing V

- The obteined result $\hat{\mathbf{x}}$ is real-valued.
- Straingtforward heuristic:
  - rounding to the nearest integer in $\{0, 1, 2\}$.
- What the paper does:
  - rank all non-zero values of $\hat{\mathbf{x}}$,
  - round the largest s non-zero values to $\{0, 1, 2\}$, set all other remaining values to 0 to get $\mathbf{x}^s$.
  - compute error $e_s = \left\| \frac{1}{2}\hat{\mathbf{M}}\mathbf{x}^s - \frac{\mathbf{z}}{r} \right\|_2$.
  - select s such that $\mathbf{x}^s$ minimizes e_s.

Robust Face Recognition I

- Problem: given n_i face pictures of the ith individual with k individuals in total as training set, figure out the class a test image belongs to.
- Difficulties: noise, occlusion.
- Solution: Robust face recognition via ℓ_1-norm [Wright-etal'09].



Robust Face Recognition II

- Construct matrix $\mathbf{A}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{in_i}) \in \mathbb{R}^{m \times n_i}$ for the ith individual, each $\mathbf{v}_{ij}$ represents the jth training image of individual i (stack all the pixel values of the image into a single vector).
- Group all the $\mathbf{A}_i$'s to get $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_k)$.
- For the testing image $\mathbf{y}$, solve:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{y} = \mathbf{A}\mathbf{x} \end{array}$$

- Interpretation: use the minimum number of linear combination of images from the traing set to express the testing image.
- The non-zero entry of $\mathbf{x}$ indicates the class that the testing image belongs to.

Robust Face Recognition III

- Given $\hat{\mathbf{x}} = \arg\min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1$, we need to identify which class (person) $\mathbf{y}$ belongs to by the following steps:
  - Reconstruct image by $\hat{\mathbf{x}}$.
    - For the ith class, define vector $\delta_i(\hat{\mathbf{x}})$ that keeps coefficients corresponding to the ith class unchanged and maps the other entries to 0.
    - Reconstructed image $\hat{\mathbf{y}} = \mathbf{A}\delta_i(\hat{\mathbf{x}})$.
    - Residual $r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}})\|_2$.
  - Identify the class as $i^\star = \arg\min_i r_i(\mathbf{y})$.

Robust Face Recognition IV

- Small dense noise:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_1 \\ \text{subject to} & \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \leq \varepsilon \end{array}$$

- Occlusion or corruption:
  - Assumption: Sparse error w.r.t. some basis $\mathbf{A}_\varepsilon$.
  - Test image: $\mathbf{y} = \mathbf{y}_0 + \mathbf{e}_0 = \mathbf{A}\mathbf{x}_0 + \mathbf{e}_0$.
  - Define matrix $\mathbf{B} = (\mathbf{A}, \mathbf{A}_\varepsilon)$, solve

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \|\mathbf{w}\|_1 \\ \text{subject to} & \mathbf{y} = \mathbf{B}\mathbf{w} \end{array}$$

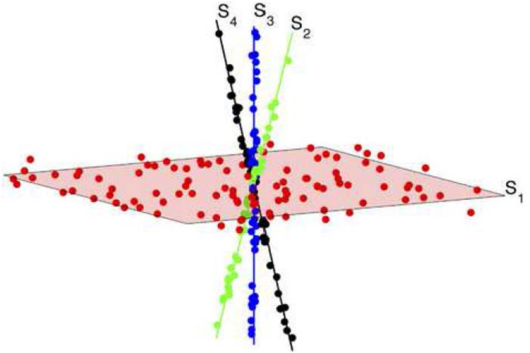
  - $\mathbf{w}$ reveals both the class testing image $\mathbf{y}$ belongs to and the error.
- Similar technique in speech recognition [Gemmeke-etal'10].

What's Else Can Be Done with Sparsity?

- We have discussed classical sparsity problems in different applications, as well as resolution techniques.
- The story always begins with: find something that is sparse...
- A rich literature on this kind of problems, what is next?
- Some seemingly unrelated problems can be formulated via sparsity.

Subspace Clustering Problem I

- Problem: given data points $\mathbf{x}_i$, $i = 1, \ldots, N$, figure out the subspaces that data lies in.
- Solution: ℓ_1-norm minimization [Soltanolkotabi-Candes'12].



Subspace Clustering Problem II

- Observation: data in the same subspace $\Longleftrightarrow$ can be expressed as linear combination of others.
- Solution: define $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix}$
  - for each $\mathbf{x}_i$, solve

$$\begin{array}{ll} \underset{\mathbf{z}}{\text{minimize}} & \left\|\mathbf{z}^{(i)}\right\|_1 \\ \text{subject to} & \mathbf{X}\mathbf{z}^{(i)} = \mathbf{x}_i \\ & \mathbf{z}_i^{(i)} = 0 \end{array}$$

  - construct matrix $\mathbf{Z} = \left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}\right]$;
  - form affinity graph G with nodes representing N data points and edge weights given by $\mathbf{W} = |\mathbf{Z}| + |\mathbf{Z}|^T$;
  - apply a spectral clustering technique to G.
- Flexible model for error and missing data.
- Tolerable of large quantity of outliers and can detect them.

Sudoku: Let's Play a Game

- Rules for Sudoku: fill in the blanks such that digits $1, \ldots, 9$ occur only once in each row, each column, each 3×3 box.
- Example of a 9×9 Sudoku puzzle:

	1		7		8	9		
3	8							
		9			5	6		
	9			7				
	3	1					2	
			4	5			8	
	5			6	2	4	9	
6	7	3		4	9		5	1
	4							3

Solving Sudoku by ℓ_1-Norm

- For cell n, define the content as $S_n \in \{1, 2, \ldots, 9\}$ and the indication vector $\mathbf{i}_n = \left(1_{\{S_n=1\}}, \ldots, 1_{\{S_n=9\}}\right)^T$.
- Stack indicator vector of all cells in row order, denote as $\mathbf{x}$.
- Objective: Find sparse $\mathbf{x}$ satisfies game rules.
- Equivalence between Sudoku and Optimization Problem [Babu-Pelckmans-Stoica'2010]:

Game:	Programming:
Objective: Solve the puzzle.	Objective: Minimize $\|\mathbf{x}\|_0$
Rules:	Constraints:
digits $1, \ldots, 9$ occur only once	
each row	$\mathbf{A}_{\text{row}}\mathbf{x} = \mathbf{1}$
each column	$\mathbf{A}_{\text{col}}\mathbf{x} = \mathbf{1}$
each box	$\mathbf{A}_{\text{box}}\mathbf{x} = \mathbf{1}$
each cell needs to be filled	$\mathbf{A}_{\text{cell}}\mathbf{x} = \mathbf{1}$
some given clue	$\mathbf{A}_{\text{clue}}\mathbf{x} = \mathbf{1}$

Summary

- What have we done?
  - Introduced cardinality constrained problems.
  - Given algorithms to solve this kind of problems via ℓ_1-norm minimization.
  - Shown many examples related to sparsity that can be nicely solved.
- Attention:
  - "All models are wrong, but some are useful", be cautious with the assumptions.
  - ℓ_1-norm relaxation is not supposed to work in all cases, it depends on the problem.
  - Examples provided in the slides are just a sketch, for details please refer to the references.

References

- D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.
- Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Trans. on Signal Processing*, vol. 65, no. 3, pp. 794-816, Feb. 2017.
- R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.
- M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.
- N. Shental, A. Amir, and O. Zuk, "Identification of rare alleles and their carriers using compressed se (que) nsing," *Nucleic acids research*, vol. 38, no. 19, pp. e179–e179, 2010.

References

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272–287, 2010.

M. Soltanolkotabi and E. J. Candes, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

P. Babu, K. Pelckmans, and P. Stoica, "Linear Systems, Sparse Solutions, and Sudoku," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 40–42, Jan. 2010.

Thanks

For more information visit:

http://www.danielppalomar.com