A totally unimodular view of structured sparsity

Volkan Cevher
volkan.cevher@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

DISCML (NIPS)
[December 13, 2014]

Joint work with
Marwa El Halabi, Luca Baldassarre and Baran Gözcü @ LIONS
Anastasios Kyrillidis and Bubacarr Bah @ UT Austin
Nirav Bhan @ MIT

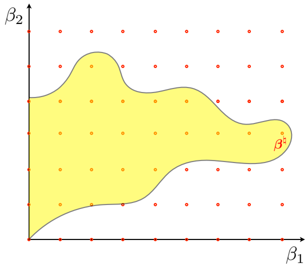
lions@epfl
MARIE CURIE ACTIONS
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
erc
EPFL

Outline

Integer linear programming

Discrete optimization

Search for an optimum object within a finite collection of objects.



Integer linear programming

Discrete optimization

Search for an optimum object within a finite collection of objects.

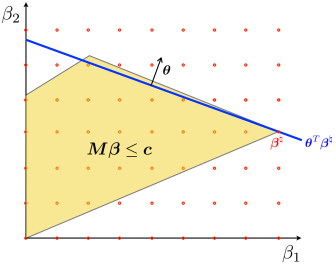
Integer linear program

Many important discrete optimization problems can be formulated as an integer linear program

$$\beta^\natural \in \arg\max_{\beta \in \mathbb{Z}^m} \{\theta^T \beta : M\beta \leq c, \beta \geq 0\} \quad \text{(ILP)}$$

NP-Hard (in general)

- vertex cover, set packing, maximum flow, traveling salesman, boolean satisfiability.



Polyhedra & Polytopes

$$\mathcal{P} = \{\beta | M\beta \leq c, \beta \geq 0\}$$

$$(\beta \in \mathbb{R}^m, c \in \mathbb{R}^m)$$

Polytope: A bounded polyhedron

Integer linear programming

Discrete optimization

Search for an optimum object within a finite collection of objects.

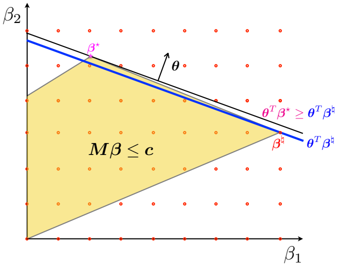
Integer linear program

Many important discrete optimization problems can be formulated as an integer linear program

$$\beta^\natural \in \arg\max_{\beta \in \mathbb{Z}^m} \{\theta^T \beta : M\beta \leq c, \beta \geq 0\} \quad \text{(ILP)}$$

NP-Hard (in general)

- vertex cover, set packing, maximum flow, traveling salesman, boolean satisfiability.



A general approach

Attempt the following convex relaxation

$$\beta^\star \in \arg\max_{\beta \in \mathbb{R}^m} \{\theta^T \beta : M\beta \leq c, \beta \geq 0\} \quad \text{(LP)}$$

Obtains an upperbound

Polyhedra & Polytopes

$$\mathcal{P} = \{\beta | M\beta \leq c, \beta \geq 0\}$$

$$(\beta \in \mathbb{R}^m, c \in \mathbb{R}^m)$$

Polytope: A bounded polyhedron

Integer linear programming

Discrete optimization

Search for an optimum object within a finite collection of objects.

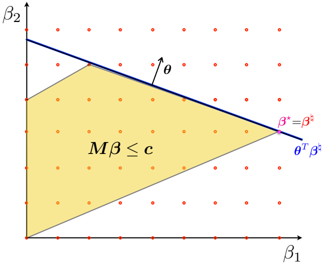
Integer linear program

Many important discrete optimization problems can be formulated as an integer linear program

$$\beta^\natural \in \arg\max_{\beta \in \mathbb{Z}^m} \{\theta^T \beta : M\beta \leq c, \beta \geq 0\} \quad \text{(ILP)}$$

NP-Hard (in general)

- vertex cover, set packing, maximum flow, traveling salesman, boolean satisfiability.



A general approach

Attempt the following convex relaxation

$$\beta^\star \in \arg\max_{\beta \in \mathbb{R}^m} \{\theta^T \beta : M\beta \leq c, \beta \geq 0\} \quad \text{(LP)}$$

Obtains an upperbound

Polyhedra & Polytopes

$$\mathcal{P} = \{\beta | M\beta \leq c, \beta \geq 0\}$$

Observation:

When every vertex of $\mathcal{P}$ is integer,

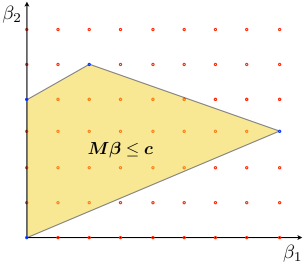
LP is a "correct" relaxation.

A sufficient condition

Polyhedra $\mathcal{P} = \{\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$ has integer vertices when $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer

Definition (Total unimodularity)

A matrix $\boldsymbol{M} \in \mathbb{R}^{l \times m}$ is totally unimodular (TU) iff the determinant of every square submatrix of $\boldsymbol{M}$ is 0, or ± 1.



Correctness of LP [23]

When $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer, then the LP

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^m} \{\boldsymbol{\theta}^T \boldsymbol{\beta} : \boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$$

has integer optimal solutions (i.e., ILP $\subseteq$ LP).

Verifying if a matrix is TU is in P [31]

TU matrices are not rare!

- Regular matroids have TU representations [29]
- Network flow problems & interval constraints involve TU matrices [23]
- Incidence matrices of undirected bipartite graphs are TU [23]

A sufficient condition

Polyhedra $\mathcal{P} = \{\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$ has integer vertices when $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer

Definition (Total unimodularity)

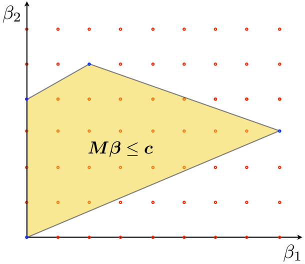
A matrix $\boldsymbol{M} \in \mathbb{R}^{l \times m}$ is totally unimodular (TU) iff the determinant of every square submatrix of $\boldsymbol{M}$ is 0, or ± 1.

Correctness of LP [23]

When $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer, then the LP

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^m} \{\boldsymbol{\theta}^T \boldsymbol{\beta} : \boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$$

has integer optimal solutions (i.e., ILP $\subseteq$ LP).



Verifying if a matrix is TU is in P [31]

Computational complexity of LP

- Polynomial time in l (i.e., number of constraints) and m (i.e., ambient dimension)
- IPM performs $\mathcal{O}\left(\sqrt{l} \log \frac{l}{\epsilon}\right)$ iterations ($l > m$) with up to $\mathcal{O}(m^2 l)$ operations, where ϵ is the absolute solution accuracy

A sufficient condition

Polyhedra $\mathcal{P} = \{\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$ has integer vertices when $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer

Definition (Total unimodularity)

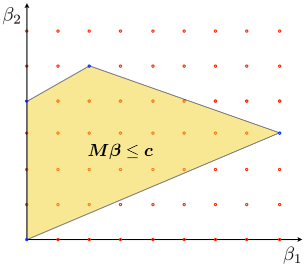
A matrix $\boldsymbol{M} \in \mathbb{R}^{l \times m}$ is totally unimodular (TU) iff the determinant of every square submatrix of $\boldsymbol{M}$ is 0, or ± 1.

Correctness of LP [23]

When $\boldsymbol{M}$ is TU and $\boldsymbol{c}$ is integer, then the LP

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^m} \{\boldsymbol{\theta}^T \boldsymbol{\beta} : \boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, \boldsymbol{\beta} \geq 0\}$$

has integer optimal solutions (i.e., ILP $\subseteq$ LP).



Verifying if a matrix is TU is in P [31]

Computational complexity of LP

- Polynomial time in l (i.e., number of constraints) and m (i.e., ambient dimension)
- **What if l is exponentially large?**

A *weaker* sufficient condition

Submodularity & submodular polyhedron [15]

$F : 2^{\mathcal{V}} \to \mathbb{R}$ is submodular iff it has the following diminishing returns property:

$$F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}),$$

$\forall \mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{V}, \forall e \in \mathcal{V} \setminus \mathcal{T}$. The submodular polyhedron is defined as

$$\mathcal{P}(F) := \{\boldsymbol{\beta} \in \mathbb{R}^m \mid \forall \mathcal{S} \subseteq \mathcal{V}, \boldsymbol{\beta}^T \mathbb{1}_{\mathcal{S}} \leq F(\mathcal{S})\}$$

where $\mathbb{1}_{\mathcal{S}}$ is the support indicator vector, i.e., $(\mathbb{1}_{\mathcal{S}})_i = 1$ if $i \in \mathcal{S}$, 0 otherwise.

- We cannot verify submodularity in polynomial time [28].
- Submodular polyhedron is TDI: LP is a "**correct**" relaxation of ILP.

Total dual integrality (TDI) [17]

A system $\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}$ is called TDI when primal objective is finite and the dual problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \left\{ \boldsymbol{\alpha}^T \boldsymbol{c} : \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha}^T \boldsymbol{M} = \boldsymbol{\theta}^T \right\}$$

has integer optimum solutions for all rational $\boldsymbol{M}$ and $\boldsymbol{c}$, and for each **integer** $\boldsymbol{\theta}$.

- A polynomial time (in l and m) algorithm can verify if $\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}$ is TDI [12].

A *weaker* sufficient condition

Submodularity & submodular polyhedron [15]

$F : 2^{\mathcal{V}} \to \mathbb{R}$ is submodular iff it has the following diminishing returns property:

$$F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}),$$

$\forall \mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{V}, \forall e \in \mathcal{V} \setminus \mathcal{T}$. The submodular polyhedron is defined as

$$\mathcal{P}(F) := \{\boldsymbol{\beta} \in \mathbb{R}^m \mid \forall \mathcal{S} \subseteq \mathcal{V}, \boldsymbol{\beta}^T \mathbb{1}_{\mathcal{S}} \leq F(\mathcal{S})\}$$

where $\mathbb{1}_{\mathcal{S}}$ is the support indicator vector, i.e., $(\mathbb{1}_{\mathcal{S}})_i = 1$ if $i \in \mathcal{S}$, 0 otherwise.

- We cannot verify submodularity in polynomial time [28].
- Submodular polyhedron is TDI: LP is a "**correct**" relaxation of ILP.

Total dual integrality (TDI) [17]

A system $\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}$ is called TDI when primal objective is finite and the dual problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \left\{ \boldsymbol{\alpha}^T \boldsymbol{c} : \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha}^T \boldsymbol{M} = \boldsymbol{\theta}^T \right\}$$

has integer optimum solutions for all rational $\boldsymbol{M}$ and $\boldsymbol{c}$, and for each **integer** $\boldsymbol{\theta}$.

- A polynomial time (in l and m) algorithm can verify if $\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}$ is TDI [12].

Structure matters! LP is ***efficiently*** solvable on the submodular polyhedra $\mathcal{P}(F)$.

In the rest of the talk...

We can use these concepts in obtaining

- tight convex relaxations
- efficient nonconvex projections

for supervised learning and inverse problems

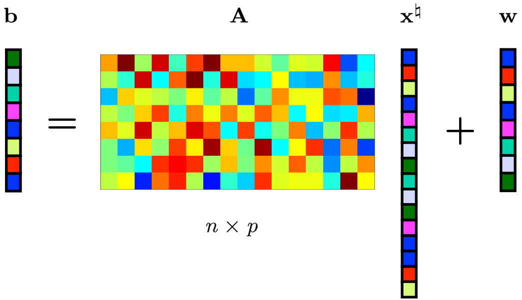
In the rest of the talk...

We can use these concepts in obtaining

- tight convex relaxations
- efficient nonconvex projections

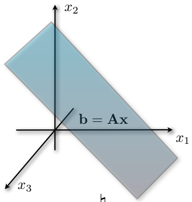
for supervised learning and inverse problems

Running example:



Applications: **Machine learning, signal processing, theoretical computer science...**

In the rest of the talk...

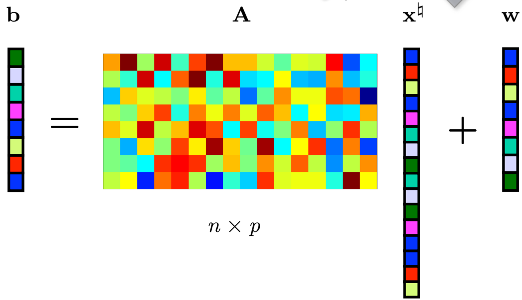


We can use these concepts in obtaining

- tight convex relaxations
- efficient nonconvex projections

for supervised learning and inverse problems

Running example:



A difficult estimation challenge when $n < p$:

Nullspace (null) of $\mathbf{A}$: $\quad \mathbf{x}^\natural + \boldsymbol{\delta} \to \mathbf{b}, \quad \forall \boldsymbol{\delta} \in \text{null}(\mathbf{A})$

- **Needle in a haystack:** ***We need additional information on $\mathbf{x}^\natural$!***

Outline

Total unimodularity in discrete optimization

From sparsity to *structured sparsity*

Convex relaxations for structured sparse recovery

Enter nonconvexity

Conclusions

Three key insights

1. Sparse or compressible $\mathbf{x}^\natural$

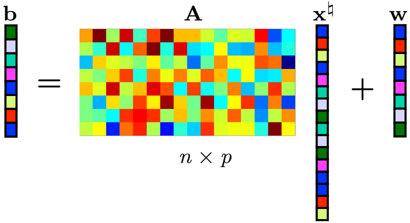
   not sufficient alone

2. Recovery

   tractable & stable

3. Projection $\mathbf{A}$

   information preserving



Typical goals:

1. Find $\mathbf{x}^\star$ to minimize $\|\mathbf{x}^\star - \mathbf{x}^\natural\|$

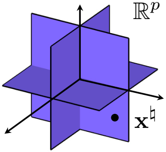
2. Find $\mathbf{x}^\star$ to minimize $\mathcal{L}\left(\mathbf{x}^\star(\mathbf{a}), \mathbf{x}^\natural(\mathbf{a}) + \mathbf{w}\right)$

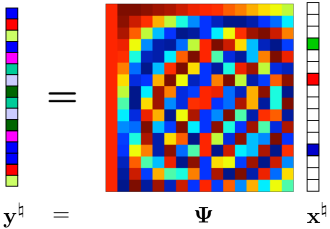
Swiss army knife of signal models

Definition (s-sparse vector)

A vector $\mathbf{x} \in \mathbb{R}^p$ is s-sparse, i.e., $\mathbf{x} \in \Sigma_s$, if it has at most s non-zero entries.

$$\left\|\mathbf{x}^\natural\right\|_0 := |\{i : x_i^\natural \neq 0\}| = s$$

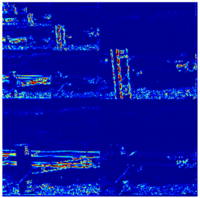




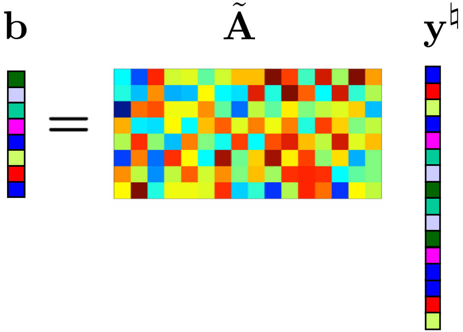
Sparse representations:

$\mathbf{y}^\natural$ has *sparse* transform coefficients $\mathbf{x}^\natural$

- Basis representations $\Psi \in \mathbb{R}^{p \times p}$
  - *Wavelets*, DCT, ...
- Frame representations $\Psi \in \mathbb{R}^{m \times p}$, $m > p$
  - Gabor, curvelets, shearlets, ...
- Other *dictionary* representations...

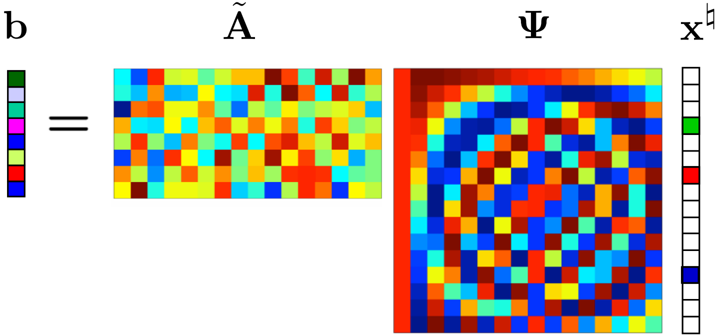


Sparse representations strike back!



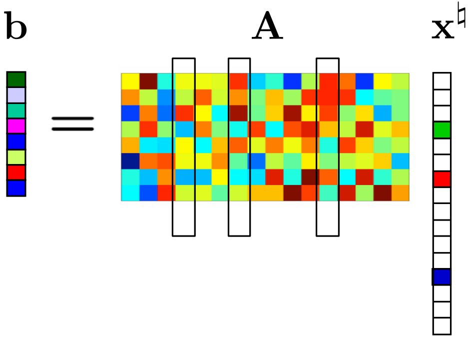
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



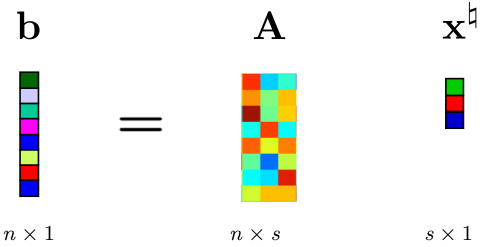
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \Sigma_s$, and $s < n < p$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, $\mathbf{x}^\natural \in \Sigma_s$, and $s < n < p$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, $\mathbf{x}^\natural \in \Sigma_s$, and $s < n < p$

Impact: Support restricted columns of $\mathbf{A}$ leads to an *overcomplete* system.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

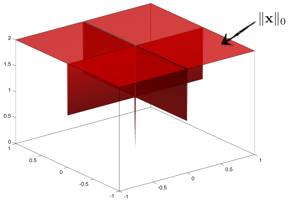
$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

$\|\mathbf{x}\|_0$ over the unit ℓ_∞-ball



Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

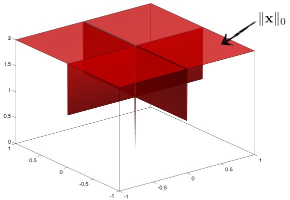
- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Tightest convex relaxation:

$\|\mathbf{x}\|_0^{**}$ is the biconjugate (Fenchel conjugate of Fenchel conjugate)

Fenchel conjugate:
$f^*(\mathbf{y}) := \sup_{\mathbf{x}:\text{dom}(f)} \mathbf{x}^T\mathbf{y} - f(\mathbf{x})$.

$\|\mathbf{x}\|_0$ over the unit ℓ_∞-ball



A technicality: Restrict $\mathbf{x}^\natural \in [-1, 1]^p$.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

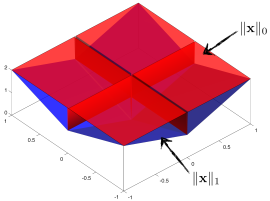
- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Tightest convex relaxation:

$\|\mathbf{x}\|_0^{**}$ is the biconjugate (Fenchel conjugate of Fenchel conjugate)

Fenchel conjugate:
$f^*(\mathbf{y}) := \sup_{\mathbf{x}:\mathrm{dom}(f)} \mathbf{x}^T\mathbf{y} - f(\mathbf{x})$.

$\|\mathbf{x}\|_1$ is the convex envelope of $\|\mathbf{x}\|_0$



A technicality: Restrict $\mathbf{x}^\natural \in [-1, 1]^p$.

The role of convexity: Tractable & stable recovery

A convex candidate solution for $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \leq 1 \right\}. \qquad \text{(SOCP)}$$

Theorem (A **model** recovery guarantee [27])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. Gaussian random variables with zero mean and variances $1/n$. For any $t > 0$ with probability at least $1 - 6\exp\left(-t^2/26\right)$, we have

$$\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2 \leq \left[\frac{2\sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s}}{\sqrt{n} - \sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s} - t}\right]\|\mathbf{w}\|_2 := \varepsilon, \quad \textit{when } \|\mathbf{x}^\natural\|_0 \leq s.$$

Observations:

- perfect recovery (i.e., $\varepsilon = 0$) with $n \geq 2s\log(\frac{p}{s}) + \frac{5}{4}s$ whp when $\mathbf{w} = 0$.
- ϵ-accurate solution in $k = \mathcal{O}\left(\sqrt{2p+1}\log(\frac{1}{\epsilon})\right)$ iterations via IPM[1]
  with each iteration requiring the solution of a structured $n \times 2p$ linear system.[2]
- robust to noise.

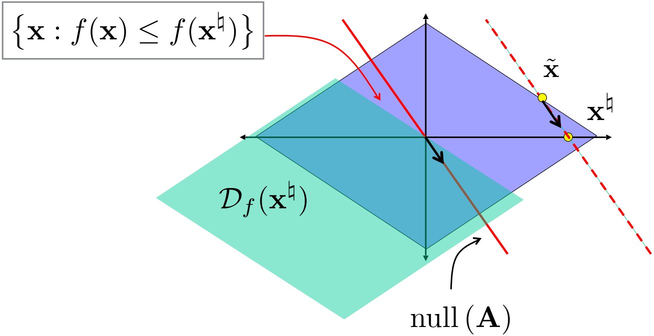
[1] For a rigorous primal-dual algorithm for this class of problems, see my NIPS 2014 paper [30].

[2] When $\mathbf{w} = 0$, the IPM complexity (# of iterations $\times$ cost per iteration) amounts to $\mathcal{O}(n^2p^{1.5}\log(\frac{1}{\epsilon}))$.

The role of the matrix $\mathbf{A}$: Preserving information

Proposition (Condition for exact recovery in the noiseless case)

We have successful recovery with $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}, \|\mathbf{x}\|_\infty \leq 1\}$, *i.e.,* $\boldsymbol{\delta} := \mathbf{x}^\star - \mathbf{x}^\natural = 0$, *if and only if* $\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}$.



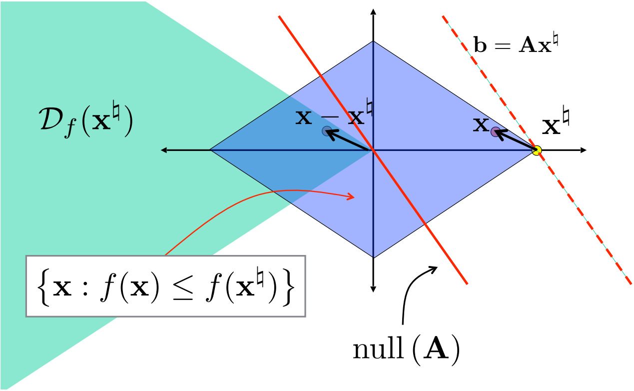
Assume that the constraint $\|\mathbf{x}\|_\infty \leq 1$ is inactive.

Descent cone: $$\mathcal{D}_f(\mathbf{x}^\natural) := \mathrm{cone}\left(\left\{\mathbf{x} : f(\mathbf{x}^\natural + \mathbf{x}) \leq f(\mathbf{x}^\natural)\right\}\right).$$

The role of the matrix $\mathbf{A}$: Preserving information

Proposition (Condition for exact recovery in the noiseless case)

We have successful recovery with $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}, \|\mathbf{x}\|_\infty \leq 1\}$, *i.e.,* $\boldsymbol{\delta} := \mathbf{x}^\star - \mathbf{x}^\natural = 0$, *if and only if* $\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}$.



The role of the matrix $\mathbf{A}$: Preserving information

$$\boxed{\mathbb{P}\left\{\mathbf{x}^\star = \mathbf{x}^\natural\right\} = \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}\right\}}$$

Definition (Statistical dimension [2][3])

Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a closed convex cone. The statistical dimension of $\mathcal{C}$ is defined as

$$d(\mathcal{C}) := \mathbb{E}\left[\|\mathrm{proj}_{\mathcal{C}}(\mathbf{g})\|_2^2\right].$$

Theorem (Approximate kinematic formula [2])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$, $n < p$, be a matrix of i.i.d. standard Gaussian random variables, and let $\mathcal{C} \subseteq \mathbb{R}^p$ be a closed convex cone. Let $\eta \in (0,1)$, then we have

$$\begin{aligned} n \geq d(\mathcal{C}) + c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{C} = \{0\}\right\} \geq 1-\eta; \\ n \leq d(\mathcal{C}) - c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{C} = \{0\}\right\} \leq \eta, \end{aligned}$$

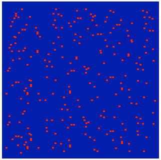
where $c_\eta := \sqrt{8\log(4/\eta)}$.

We can compute $d(\mathcal{C}) \lesssim 2s\log(\frac{p}{s}) + \frac{5}{4}s$ for $\mathcal{C} = \mathcal{D}_{\|\cdot\|_1}(\mathbf{x}^\natural)$ when $\mathbf{x}^\natural \in \Sigma_s$.

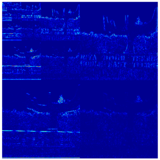
[3] The statistical dimension is closely related to the Gaussian complexity [7], Gaussian width [10], mean width [32], and Gaussian squared complexity [9].

Beyond sparsity towards model-based or *structured* sparsity

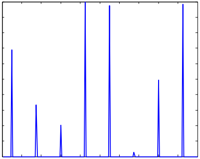
- The following signals can look the **same** from a **sparsity** perspective!



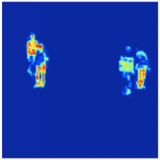
Sparse image



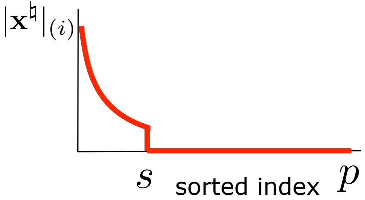
Wavelet coefficients of a natural image



Spike train

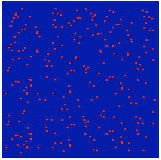


Background substracted image

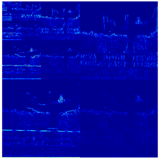


Beyond sparsity towards model-based or *structured* sparsity

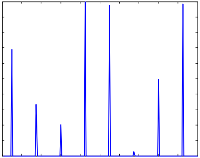
- The following signals can look the **same** from a **sparsity** perspective!



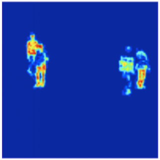
Sparse image



Wavelet coefficients of a natural image

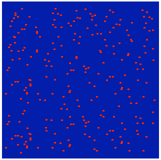


Spike train

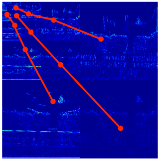


Background substracted image

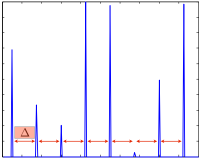
- In reality, these signals have additional structures beyond the simple sparsity



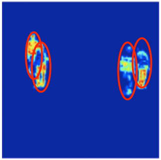
Sparse image



Wavelet coefficients of a natural image



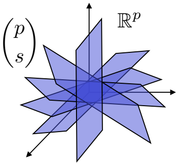
Spike train

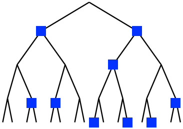


Background substracted image

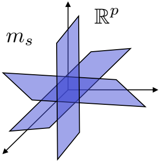
Beyond sparsity towards model-based or *structured* sparsity

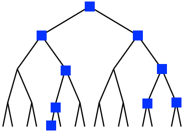
Sparsity model: Union of all s-dimensional canonical subspaces.





Structured sparsity model: A **particular** union of m_s s-dimensional canonical subspaces.





Model-based or *structured* sparsity

Structured sparsity models are discrete structures describing the interdependency between the non-zero coefficients of a vector.

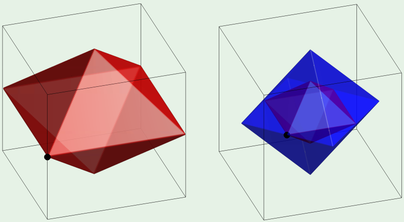
Three upshots of structured sparsity

Key properties of the statistical dimension [2]

- The statistical dimension is invariant under unitary transformations (rotations).
- Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be closed convex cones. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $d(\mathcal{C}_1) \leq d(\mathcal{C}_2)$.

1. The smaller the statistical dimension is, the less we need to sample

Example (If $\mathcal{D}_{f_1}(\mathbf{x}^\natural) \subseteq \mathcal{D}_{f_2}(\mathbf{x}^\natural) \subseteq \mathbb{R}^n$, then $d(\mathcal{D}_{f_1}(\mathbf{x}^\natural)) \leq d(\mathcal{D}_{f_2}(\mathbf{x}^\natural))$.)



$f_1(\mathbf{x}) := \max\{|x_1|, |x_2|\} + |x_3|$

$f_2(\mathbf{x}) := \|\mathbf{x}\|_1$

$\mathbf{x}^\natural = [1, -1, 0]^T$

Observations:

1. $n_1 < n_2$ for $\mathbf{x}^\natural$
2. $n_1 > n_2$ for $\mathbf{z}^\natural = [0, 0, 1]^T$

- Reduced sample complexity: *phase transition* at the statistical dimension

Three upshots of structured sparsity

Key properties of the statistical dimension [2]

- The statistical dimension is invariant under unitary transformations (rotations).
- Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be closed convex cones. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $d(\mathcal{C}_1) \leq d(\mathcal{C}_2)$.

2. The smaller the statistical dimension is, the better we can denoise

- Reduced sample complexity: *phase transition* at the statistical dimension
- Better noise robustness: denoising capabilities depend on the statistical dimension

$$\max_{\sigma>0} \frac{\mathbb{E}\left[\|\mathsf{prox}_f(\mathbf{x}^\natural + \sigma\mathbf{w}, \sigma\lambda) - \mathbf{x}^\natural\|^2\right]}{\sigma^2} \leq d(\lambda\mathcal{D}_f(\mathbf{x}^\natural))$$

Minimize a bound to the minimax risk via the regularization parameter λ [27]

Three upshots of structured sparsity

Key properties of the statistical dimension [2]

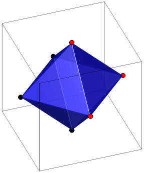
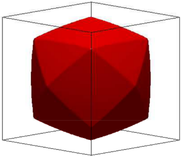
- The statistical dimension is invariant under unitary transformations (rotations).
- Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be closed convex cones. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $d(\mathcal{C}_1) \leq d(\mathcal{C}_2)$.

3. The smaller the statistical dimension is, the better we can enforce structure

- Reduced sample complexity: *phase transition* at the statistical dimension
- Better noise robustness: denoising capabilities depend on the statistical dimension

$$\max_{\sigma > 0} \frac{\mathbb{E}\left[\|\text{prox}_f(\mathbf{x}^\natural + \sigma\mathbf{w}, \sigma\lambda) - \mathbf{x}^\natural\|^2\right]}{\sigma^2} \leq d(\lambda\mathcal{D}_f(\mathbf{x}^\natural))$$

  Minimize a bound to the minimax risk via the regularization parameter λ [27]
- Better interpretability: geometry can enhance interpretability

ℓ_1-ball:  TV-ball: 

Influence the recovered support via customized convex geometry

Outline

A simple template for linear inverse problems

Find the "***sparsest***" $\mathbf{x}$ subject to ***structure*** and ***data***.

- **Sparsity**

  We can generalize this desideratum to other notions of simplicity

- **Structure**

  We only allow certain sparsity patterns

- **Data fidelity**

  We have many choices of convex constraints & losses to represent data; e.g.,

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa$$

A *convex* proto-problem for *structured* sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the sparsest estimator or its surrogate with a valid sparsity pattern:

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\boldsymbol{s}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq 1 \right\} \qquad (\mathcal{P}_s)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then the structured sparse $\mathbf{x}^\natural$ is a feasible solution.

Sparsity and structure together [14]

Given some weights $\boldsymbol{d} \in \mathbb{R}^d$, $\boldsymbol{e} \in \mathbb{R}^p$ and an integer input $\boldsymbol{c} \in \mathbb{Z}^l$, we define

$$\|\mathbf{x}\|_{\boldsymbol{s}} := \min_{\boldsymbol{\omega}} \{ \boldsymbol{d}^T\boldsymbol{\omega} + \boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\begin{bmatrix}\boldsymbol{\omega}\\ \boldsymbol{s}\end{bmatrix} \leq \boldsymbol{c}, \mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}, \boldsymbol{\omega} \in \{0,1\}^d \}$$

for all feasible $\mathbf{x}$, ∞ otherwise. The parameter $\boldsymbol{\omega}$ is useful for latent modeling.

A *convex* proto-problem for *structured* sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the sparsest estimator or its surrogate with a valid sparsity pattern:

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\boldsymbol{s}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq 1 \right\} \qquad (\mathcal{P}_s)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then the structured sparse $\mathbf{x}^\natural$ is a feasible solution.

Sparsity and structure together [14]

Given some weights $\boldsymbol{d} \in \mathbb{R}^d$, $\boldsymbol{e} \in \mathbb{R}^p$ and an integer input $\boldsymbol{c} \in \mathbb{Z}^l$, we define

$$\|\mathbf{x}\|_{\boldsymbol{s}} := \min_{\boldsymbol{\omega}} \{ \boldsymbol{d}^T \boldsymbol{\omega} + \boldsymbol{e}^T \boldsymbol{s} : \boldsymbol{M} \begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{s} \end{bmatrix} \leq \boldsymbol{c}, \mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}, \boldsymbol{\omega} \in \{0,1\}^d \}$$

for all feasible $\mathbf{x}$, ∞ otherwise. The parameter $\boldsymbol{\omega}$ is useful for latent modeling.

A convex candidate solution for $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We use the convex estimator based on the tightest convex relaxation of $\|\mathbf{x}\|_{\boldsymbol{s}}$:

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathrm{dom}(\|\cdot\|_{\boldsymbol{s}})} \left\{ \|\mathbf{x}\|_{\boldsymbol{s}}^{**} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

with some $\kappa \geq 0$, $\mathrm{dom}(\|\cdot\|_{\boldsymbol{s}}) := \{\mathbf{x} : \|\mathbf{x}\|_{\boldsymbol{s}} < \infty\}$.

Tractability & tightness of biconjugation

Proposition (Hardness of conjugation)

Let $F(\boldsymbol{s}) : 2^{\mathfrak{P}} \to \mathbb{R} \cup \{+\infty\}$ *be a set function defined on the support* $\boldsymbol{s} = supp(\mathbf{x})$. *Conjugate of* F *over the unit infinity ball* $\|\mathbf{x}\|_\infty \leq 1$ *is given by*

$$g^*(\mathbf{y}) = \sup_{\boldsymbol{s} \in \{0,1\}^p} |\mathbf{y}|^T \boldsymbol{s} - F(\boldsymbol{s}).$$

Observations:

- $F(\boldsymbol{s})$ is general set function

  Computation: NP-Hard

- $F(\boldsymbol{s}) = \|\mathbf{x}\|_{\boldsymbol{s}}$

  Computation: ILP in general. However, if
  - $\boldsymbol{M}$ is TU
  - $(\boldsymbol{M}, \boldsymbol{c})$ is TDI

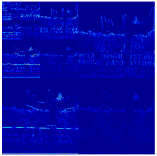
  then tight convex relaxations with an LP ("usually" tractable)

  Otherwise, relax to LP anyway!

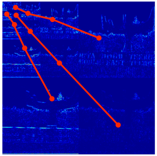
- $F(\boldsymbol{s})$ is submodular

  Computation: Polynomial-time

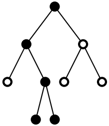
Tree sparsity [21, 13, 6, 33]



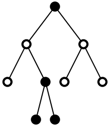
Wavelet coefficients



Wavelet tree



Valid selection of nodes



Invalid selection of nodes

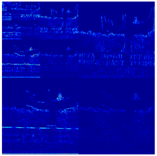
Structure: *We seek the sparsest signal with a rooted connected subtree support.*

Linear description: A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

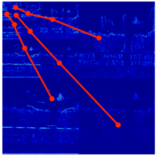
$$\boxed{\boldsymbol{T}\mathbb{1}_{\mathrm{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0}$$

where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

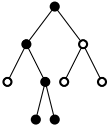
Tree sparsity [21, 13, 6, 33]



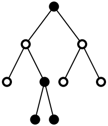
Wavelet coefficients



Wavelet tree



Valid selection of nodes



Invalid selection of nodes

Structure: *We seek the sparsest signal with a rooted connected subtree support.*

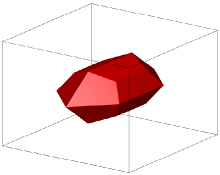
Linear description: A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

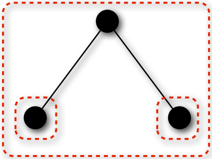
$$\boxed{\boldsymbol{T}\mathbb{1}_{\text{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0}$$

where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

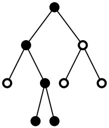
Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \min_{\boldsymbol{s}\in[0,1]^p}\{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{T}\boldsymbol{s} \geq 0, |\mathbf{x}| \leq \boldsymbol{s}\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Tree sparsity [21, 13, 6, 33]





$\mathfrak{G}_H = \{\{1,2,3\},\{2\},\{3\}\}$



valid selection of nodes

Structure: *We seek the sparsest signal with a rooted connected subtree support.*

Linear description: A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

$$\boxed{\boldsymbol{T}\mathbb{1}_{\text{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0}$$

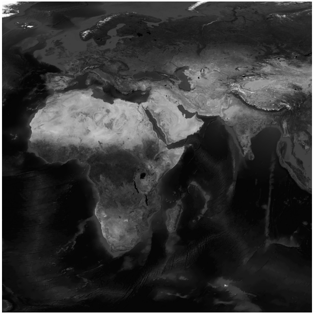
where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \min_{\boldsymbol{s}\in[0,1]^p}\{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{T}\boldsymbol{s} \geq 0, |\mathbf{x}| \leq \boldsymbol{s}\} \overset{\star}{=} \sum_{\mathcal{G}\in\mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

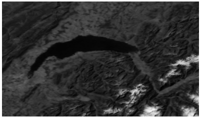
The set $\mathcal{G} \in \mathfrak{G}_H$ are defined as each node and all its descendants.

Tree sparsity example: 1:100-compressive sensing [30, 1]

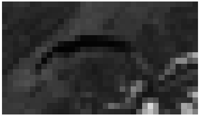
World [1Gpix]



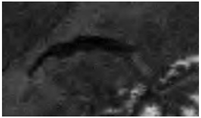
Lac Léman



World [10Mpix]

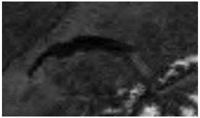


sparse



PNSR $= 31.83$db

tree-sparse



PNSR $= 32.48$db

Tree sparsity example: TV & TU-relax 1:15-compression [30, 1]

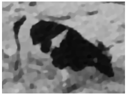
Original

Original

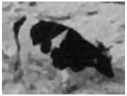
BP

TU-relax

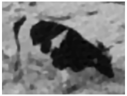
TV



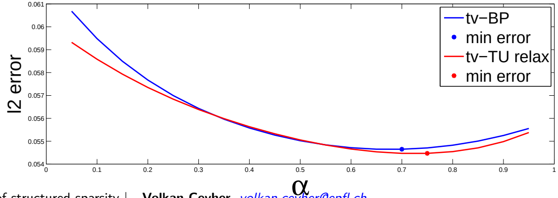
TV with BP



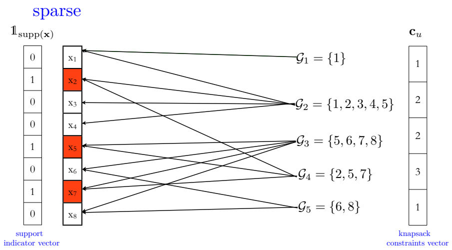
TV with TU-relax



Regularization:



Group knapsack sparsity [35, 18, 16]



Structure: *We seek the sparsest signal with group allocation constraints.*

Linear description: A valid support obeys budget constraints over $\mathfrak{G}$

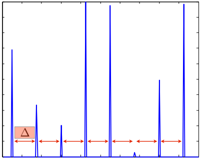
$$\boxed{\boldsymbol{B}^T\boldsymbol{s} \leq \boldsymbol{c}_u}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

When $\boldsymbol{B}$ is an interval matrix or $\mathfrak{G}$ has a *loopless* group intersection graph, it is TU.

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \boldsymbol{B}^T\boldsymbol{s} \leq \boldsymbol{c}_u$.

Group knapsack sparsity [35, 18, 16]



$$\boldsymbol{B}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ & & & & \ddots & & & & \\ 0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}_{(p-\Delta+1)\times p}$$

Structure: *We seek the sparsest signal with group allocation constraints.*

Linear description: A valid support obeys budget constraints over $\mathfrak{G}$

$$\boxed{\boldsymbol{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

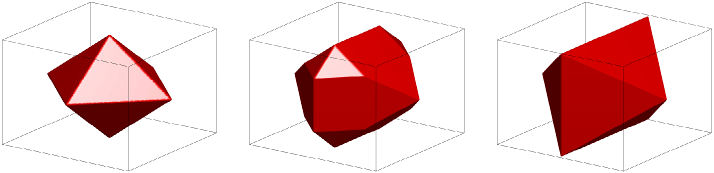
When $\boldsymbol{B}$ is an interval matrix or $\mathfrak{G}$ has a *loopless* group intersection graph, it is TU.

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \boldsymbol{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \begin{cases} \|\mathbf{x}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \boldsymbol{B}^T|\mathbf{x}| \leq \boldsymbol{c}_u, \\ \infty & \text{otherwise} \end{cases}$

For the neuronal spike example, we have $\boldsymbol{c}_u = \mathbb{1}$.

Group knapsack sparsity [35, 18, 16]



(left) $\|\mathbf{x}\|_s^{**} \leq 1$ (middle) $\|\mathbf{x}\|_s^{**} \leq 1.5$ (right) $\|\mathbf{x}\|_s^{**} \leq 2$ for $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$

Structure: *We seek the sparsest signal with group allocation constraints.*

Linear description: A valid support obeys budget constraints over $\mathfrak{G}$

$$\boxed{\boldsymbol{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

When $\boldsymbol{B}$ is an interval matrix or $\mathfrak{G}$ has a *loopless* group intersection graph, it is TU.

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \boldsymbol{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \begin{cases} \|\mathbf{x}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \boldsymbol{B}^T|\mathbf{x}| \leq \boldsymbol{c}_u, \\ \infty & \text{otherwise} \end{cases}$

For the neuronal spike example, we have $\boldsymbol{c}_u = \mathbb{1}$.

Group knapsack sparsity example: A stylized spike train

- Basis pursuit (BP): $\|\mathbf{x}\|_1$
- TU-relax (TU):

$$\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \begin{cases} \|\mathbf{x}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \boldsymbol{B}^T|\mathbf{x}| \leq \boldsymbol{c}_u, \\ \infty & \text{otherwise} \end{cases}$$

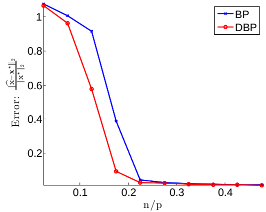
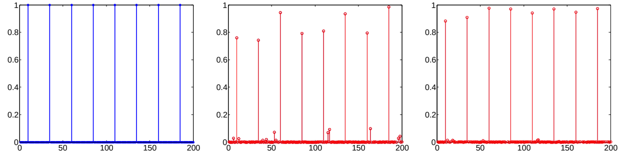
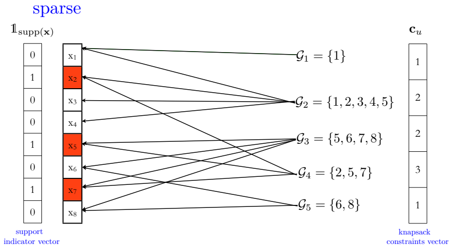


Figure: *Recovery for $n = 0.18p$.*



	$\mathbf{x}^\natural$	$\mathbf{x}_{\text{BP}}$ solution	$\mathbf{x}_{\text{TU}}$ solution
relative errors:		$\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{BP}}\|_2}{\|\mathbf{x}^\natural\|_2} = .200$	$\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{TU}}\|_2}{\|\mathbf{x}^\natural\|_2} = .067$

Group knapsack sparsity: A simple variation



Structure: *We seek the signal with the minimal overall group allocation.*

$$\text{Objective: } \mathbb{1}^T \boldsymbol{s} \to \|\mathbf{x}\|_{\boldsymbol{\omega}} = \begin{cases} \min_{\omega \in \mathbb{Z}_{++}} \omega & \text{if } \mathbf{x} \in [-1,1]^p, \boldsymbol{B}^T|\mathbf{x}| \leq \omega \mathbb{1}, \\ \infty & \text{otherwise} \end{cases}$$

Linear description: A valid support obeys budget constraints over $\mathfrak{G}$

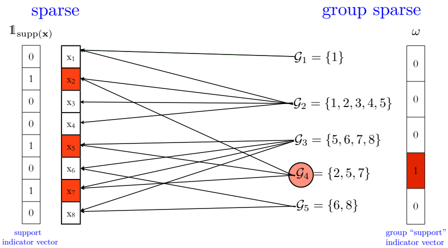
$$\boxed{\boldsymbol{B}^T \boldsymbol{s} \leq \omega \mathbb{1}}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.
When $\boldsymbol{B}$ is an interval matrix or $\mathfrak{G}$ has a *loopless* group intersection graph, it is TU.

$$\textbf{Biconjugate: } \|\mathbf{x}\|_s^{**} = \begin{cases} \max_{\mathcal{G} \in \mathfrak{G}} \|\mathbf{x}_{\mathcal{G}}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \\ \infty & \text{otherwise} \end{cases}$$

Remark: The regularizer is known as *exclusive Lasso* [35, 26].

Group cover sparsity: Minimal group cover [5, 25, 19]



Structure: *We seek the signal covered by a minimal number of groups.*

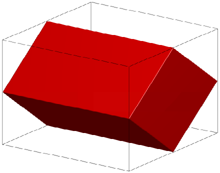
$$\text{Objective: } \mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$$

Linear description: At least one group containing a sparse coefficient is selected

$$\boxed{\boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.
When $\boldsymbol{B}$ is an interval matrix, or $\mathfrak{G}$ has a *loopless* group intersection graph it is TU.

Group cover sparsity: Minimal group cover [5, 25, 19]



$\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

Structure: *We seek the signal covered by a minimal number of groups.*

$$\text{Objective: } \mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$$

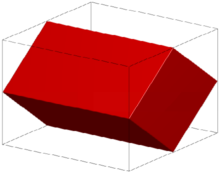
Linear description: At least one group containing a sparse coefficient is selected

$$\boxed{\boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.
When $\boldsymbol{B}$ is an interval matrix, or $\mathfrak{G}$ has a *loopless* group intersection graph it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{B}\boldsymbol{\omega} \geq |\mathbf{x}|\}$ for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise

Group cover sparsity: Minimal group cover [5, 25, 19]



$\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

Structure: *We seek the signal covered by a minimal number of groups.*

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

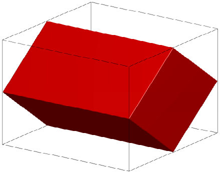
Linear description: At least one group containing a sparse coefficient is selected

$$\boxed{\boldsymbol{B\omega} \geq \boldsymbol{s}}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.
When $\boldsymbol{B}$ is an interval matrix, or $\mathfrak{G}$ has a *loopless* group intersection graph it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\boldsymbol{d}^T\boldsymbol{\omega} : \boldsymbol{B\omega} \geq |\mathbf{x}|\}$ for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise
$\overset{\star}{=} \min_{\mathbf{v}_i\in\mathbb{R}^p}\{\sum_{i=1}^M d_i\|\mathbf{v}_i\|_\infty : \mathbf{x} = \sum_{i=1}^M \mathbf{v}_i, \forall \mathrm{supp}(\mathbf{v}_i) \subseteq \mathcal{G}_i\}$,

Group cover sparsity: Minimal group cover [5, 25, 19]



$\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

Structure: *We seek the signal covered by a minimal number of groups.*

Objective: $\mathbb{1}^T \boldsymbol{s} \rightarrow \boldsymbol{d}^T \boldsymbol{\omega}$

Linear description: At least one group containing a sparse coefficient is selected

$$\boxed{\boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

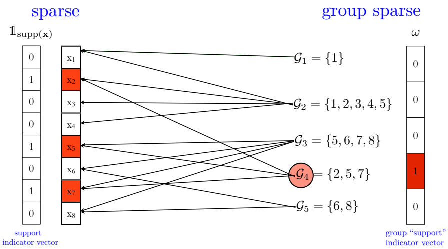
When $\boldsymbol{B}$ is an interval matrix, or $\mathfrak{G}$ has a *loopless* group intersection graph it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\boldsymbol{d}^T\boldsymbol{\omega} : \boldsymbol{B}\boldsymbol{\omega} \geq |\mathbf{x}|\}$ for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise

$$\stackrel{\star}{=} \min_{\mathbf{v}_i \in \mathbb{R}^p}\{\sum_{i=1}^M d_i \|\mathbf{v}_i\|_\infty : \mathbf{x} = \sum_{i=1}^M \mathbf{v}_i, \forall \text{supp}(\mathbf{v}_i) \subseteq \mathcal{G}_i\},$$

Remark: Weights $\boldsymbol{d}$ can depend on the sparsity within each groups (not TU) [14].

Budgeted group cover sparsity



Structure: *We seek the sparsest signal covered by G groups.*

Objective: $\boldsymbol{d}^T\boldsymbol{\omega} \rightarrow \mathbb{1}^T\boldsymbol{s}$

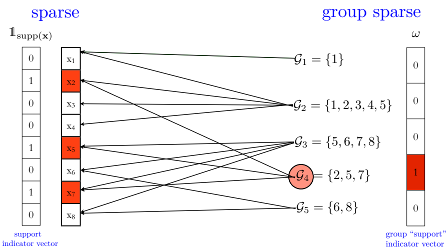
Linear description: At least one of the G selected groups cover each sparse coefficient.

$$\boxed{\boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}, \mathbb{1}^T\boldsymbol{\omega} \leq G}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

When $\begin{bmatrix} \boldsymbol{B} \\ \mathbb{1} \end{bmatrix}$ is an interval matrix, it is TU.

Budgeted group cover sparsity



Structure: *We seek the sparsest signal covered by G groups.*

Objective: $\boldsymbol{d}^T\boldsymbol{\omega} \to \mathbb{1}^T\boldsymbol{s}$

Linear description: At least one of the G selected groups cover each sparse coefficient.

$$\boxed{\boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}, \mathbb{1}^T\boldsymbol{\omega} \leq G}$$

where $\boldsymbol{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\boldsymbol{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

When $\begin{bmatrix} \boldsymbol{B} \\ \mathbb{1} \end{bmatrix}$ is an interval matrix, it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\|\mathbf{x}\|_1 : \boldsymbol{B}\boldsymbol{\omega} \geq |\mathbf{x}|, \mathbb{1}^T\boldsymbol{\omega} \leq G\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Budgeted group cover example: Interval overlapping groups

- Basis pursuit (BP): $\|\mathbf{x}\|_1$
- Sparse group Lasso (SGL$_q$):

$$(1-\alpha)\sum_{\mathcal{G}\in\mathfrak{G}} \sqrt{|\mathcal{G}|}\|\mathbf{x}_{\mathcal{G}}\|_q + \alpha\|\mathbf{x}_{\mathcal{G}}\|_1$$

- TU-relax (TU):

$$\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M} \{\|\mathbf{x}\|_1 : \boldsymbol{B\omega} \geq |\mathbf{x}|, \mathbb{1}^T\boldsymbol{\omega} \leq G\}$$

for $\mathbf{x}\in[-1,1]^p$, ∞ otherwise.

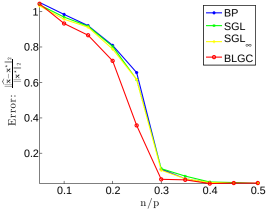
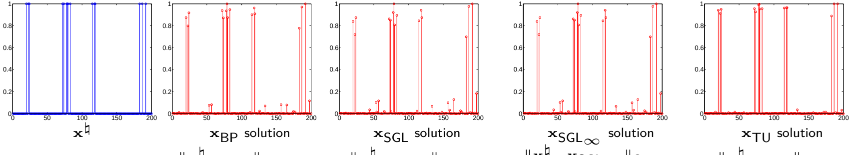
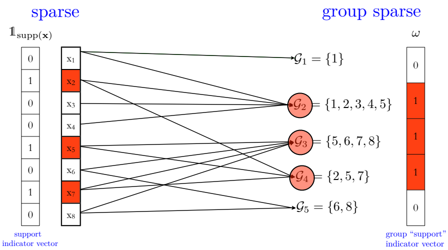


Figure: *Recovery for $n = 0.25p$, $s = 15$, $p = 200$, $G = 5$ out of $M = 29$ groups.*



relative errors: $\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{BP}}\|_2}{\|\mathbf{x}^\natural\|_2} = .128$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{SGL}}\|_2}{\|\mathbf{x}^\natural\|_2} = .181$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{SGL}_\infty}\|_2}{\|\mathbf{x}^\natural\|_2} = .085$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}_{\text{TU}}\|_2}{\|\mathbf{x}^\natural\|_2} = .058$

Group intersection sparsity [20, 34, 3]



Structure: *We seek the signal intersecting with minimal number of groups.*

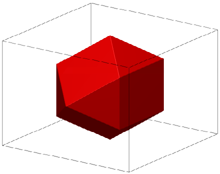
$$\text{Objective: } \mathbb{1}^T \boldsymbol{s} \rightarrow \boldsymbol{d}^T \boldsymbol{\omega}$$

Linear description: All groups containing a sparse coefficient are selected

$$\boxed{\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}}$$

where $\boldsymbol{H}_k(i, j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Group intersection sparsity [20, 34, 3]



$\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$
(left) intersection (right) cover.

Structure: *We seek the signal intersecting with minimal number of groups.*

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

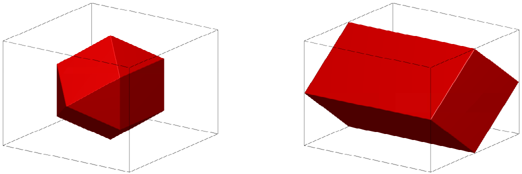
Linear description: All groups containing a sparse coefficient are selected

$$\boxed{\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}_k |\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Group intersection sparsity [20, 34, 3]



$\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$
(left) intersection (right) cover.

Structure: *We seek the signal intersecting with minimal number of groups.*

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$ (*submodular*)

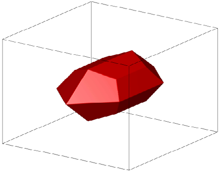
Linear description: All groups containing a sparse coefficient are selected

$$\boxed{\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}_k |\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\} \overset{\star}{=} \sum_{\mathcal{G} \in \mathfrak{G}} \|x_{\mathcal{G}}\|_\infty$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Group intersection sparsity [20, 34, 3]



$\mathfrak{G} = \{\{1,2,3\},\{2\},\{3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

Structure: *We seek the signal intersecting with minimal number of groups.*

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$ (*submodular*)

Linear description: All groups containing a sparse coefficient are selected

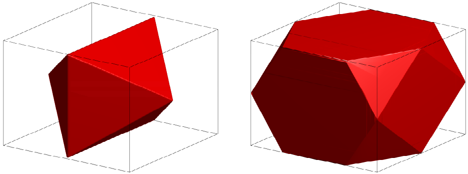
$$\boxed{\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\boldsymbol{d}^T\boldsymbol{\omega} : \boldsymbol{H}_k|\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\} \overset{\star}{=} \sum_{\mathcal{G}\in\mathfrak{G}} \|x_{\mathcal{G}}\|_\infty$ for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Remark: For hierarchical $\mathfrak{G}_H$, group intersection and tree sparsity models coincide.

Beyond linear costs: Graph dispersiveness



(left) $\|\mathbf{x}\|_s^{**} = 0$ (right) $\|\mathbf{x}\|_s^{**} \leq 1$ for $\mathcal{E} = \{\{1,2\},\{2,3\}\}$ (chain graph)

Structure: *We seek a signal dispersive over a given graph* $\mathcal{G}(\mathfrak{P}, \mathcal{E})$

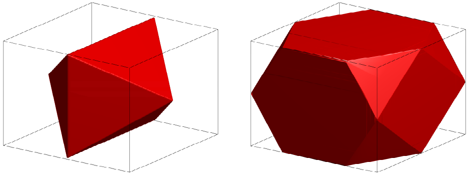
Objective: $\mathbb{1}^T \boldsymbol{s} \rightarrow \sum_{(i,j)\in\mathcal{E}} s_i s_j$ (non-linear, supermodular function)

Linearization:

$$\boxed{\|\mathbf{x}\|_s = \min_{\mathbf{z}\in\{0,1\}^{|\mathcal{E}|}} \{\textstyle\sum_{(i,j)\in\mathcal{E}} z_{ij} : z_{ij} \geq s_i + s_j - 1\}}$$

When edge-node incidence matrix of $\mathcal{G}(\mathfrak{P}, \mathcal{E})$ is TU (e.g., bipartite graphs), it is TU.

Beyond linear costs: Graph dispersiveness



(left) $\|\mathbf{x}\|_s^{**} = 0$ (right) $\|\mathbf{x}\|_s^{**} \leq 1$ for $\mathcal{E} = \{\{1,2\},\{2,3\}\}$ (chain graph)

Structure: *We seek a signal dispersive over a given graph $\mathcal{G}(\mathfrak{P}, \mathcal{E})$*

Objective: $\mathbb{1}^T \boldsymbol{s} \to \sum_{(i,j)\in\mathcal{E}} s_i s_j$ (non-linear, supermodular function)

Linearization:

$$\boxed{\|\mathbf{x}\|_s = \min_{\mathbf{z}\in\{0,1\}^{|\mathcal{E}|}} \{\textstyle\sum_{(i,j)\in\mathcal{E}} z_{ij} : z_{ij} \geq s_i + s_j - 1\}}$$

When edge-node incidence matrix of $\mathcal{G}(\mathfrak{P}, \mathcal{E})$ is TU (e.g., bipartite graphs), it is TU.

Biconjugate: $\|\mathbf{x}\|_s^{**} = \sum_{(i,j)\in\mathcal{E}} (|x_i| + |x_j| - 1)_+$ for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise.

Outline

Total unimodularity in discrete optimization

From sparsity to *structured sparsity*

Convex relaxations for structured sparse recovery

Enter nonconvexity

Conclusions

An important alternative

Problem (Projection)

Define $\mathcal{M}_{s,G} := \{\mathbf{x} : \boldsymbol{e}^T \boldsymbol{s} \leq s, \boldsymbol{d}^T \boldsymbol{\omega} \leq G, \; \boldsymbol{M} \begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{s} \end{bmatrix} \leq \boldsymbol{c}, \; \boldsymbol{s} = \mathbb{1}_{\mathrm{supp}(\mathbf{x})}\}$.

The projection of $\mathbf{x}$ *onto* $\mathcal{M}_{s,G}$ *in* ℓ_q*-norm is defined as* $\mathcal{P}_{q,\mathcal{M}_{s,G}}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$*,*

$$\mathcal{P}_{q,\mathcal{M}_{s,G}}(\mathbf{x}) \in \arg\min_{\mathbf{u} \in \mathbb{R}^p} \{\|\mathbf{x} - \mathbf{u}\|_q^q : \mathbf{u} \in \mathcal{M}_{s,G}\}$$

- $\hat{\mathbf{x}} = \mathcal{P}_{M_{k,G}}(\mathbf{x})$ *is the best model-based approximation of* $\mathbf{x}$*.*
- *The interesting cases are* $q = 1, 2$*.*

Observation: Model-based approximation corresponds to an ILP

- NP-Hard in general (weighted max cover formulation [5])
- TU structures play a major role
- Pseudo-polynomial time solutions via dynamic programming

An important alternative

Problem (Projection)

Define $\mathcal{M}_{s,G} := \{\mathbf{x} : \boldsymbol{e}^T\boldsymbol{s} \leq s, \boldsymbol{d}^T\boldsymbol{\omega} \leq G,\ \boldsymbol{M}\begin{bmatrix}\boldsymbol{\omega}\\ \boldsymbol{s}\end{bmatrix} \leq \boldsymbol{c},\ \boldsymbol{s} = \mathbb{1}_{\mathrm{supp}(\mathbf{x})}\}$.

The projection of $\mathbf{x}$ *onto* $\mathcal{M}_{s,G}$ *in* ℓ_q*-norm is defined as* $\mathcal{P}_{q,\mathcal{M}_{s,G}}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$,

$$\mathcal{P}_{q,\mathcal{M}_{s,G}}(\mathbf{x}) \in \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\{\|\mathbf{x}-\mathbf{u}\|_q^q : \mathbf{u} \in \mathcal{M}_{s,G}\}$$

- $\hat{\mathbf{x}} = \mathcal{P}_{\boldsymbol{M}_{k,G}}(\mathbf{x})$ *is the best model-based approximation of* $\mathbf{x}$.
- *The interesting cases are* $q = 1, 2$.

Observation: Model-based approximation corresponds to an ILP

- NP-Hard in general (weighted max cover formulation [5])
- TU structures play a major role
- Pseudo-polynomial time solutions via dynamic programming

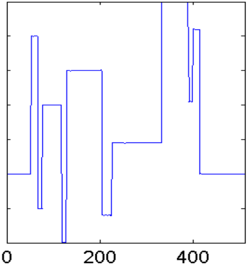
Model-based CS [6, 22]: $n = \mathcal{O}(\log|\mathcal{M}_{s,G}|)$ with *iid* Gaussian (dense)

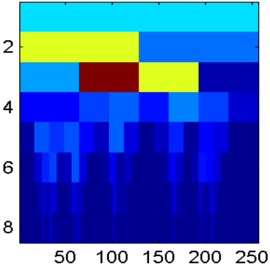
- $n = \mathcal{O}(s)$ for tree structure
- iterative projected gradient descent $\mathbf{x}^{k+1} \in \mathcal{P}_{2,\mathcal{M}_{s,G}}\left(\mathbf{x}^k + \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^k)\right)$

Model-based sketching [4]: $n = o\left(s\log(p/s)\right)$ with expanders (sparse)

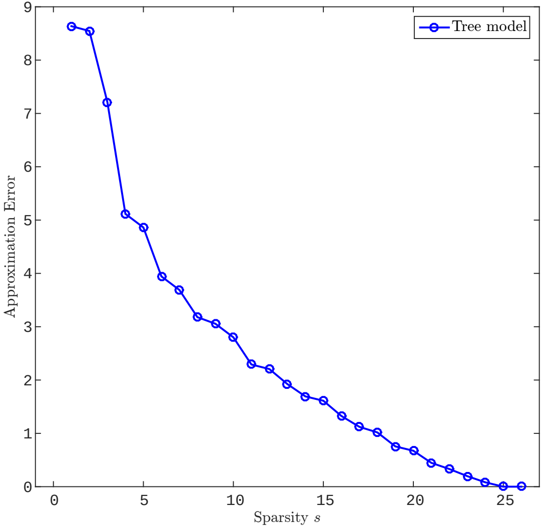
- $n = \mathcal{O}(s\log(p/s)/\log\log(p))$ for tree structure (empirical: $n = \mathcal{O}(s)$)
- iterative projected median descent $\mathbf{x}^{k+1} \in \mathcal{P}_{1,\mathcal{M}_{s,G}}(\mathbf{x}^k + \mathfrak{M}(\mathbf{b} - \mathbf{A}\mathbf{x}^k))$

Tree sparsity example: Pareto frontier [8, 5]

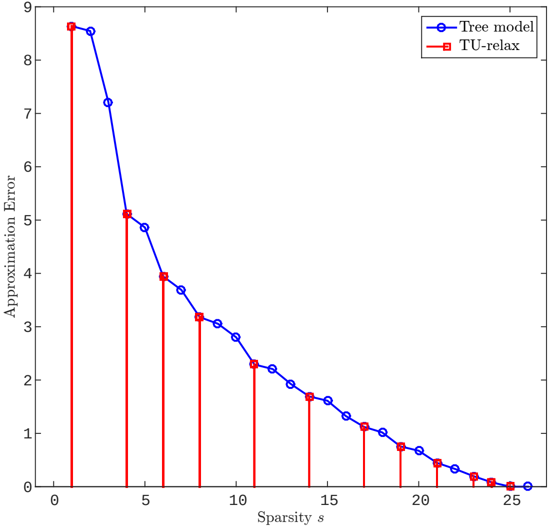




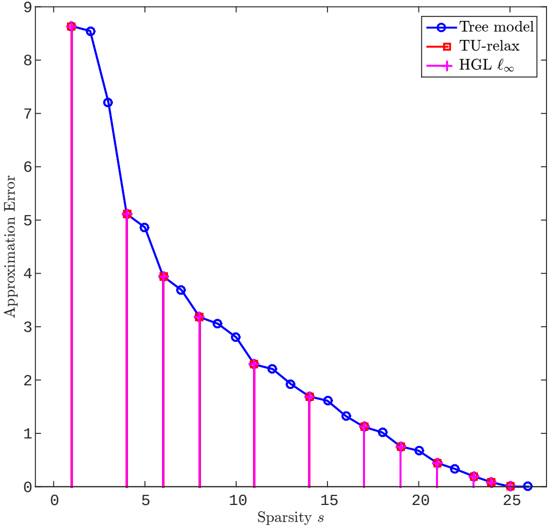
Tree sparsity example: Pareto frontier [8, 5]



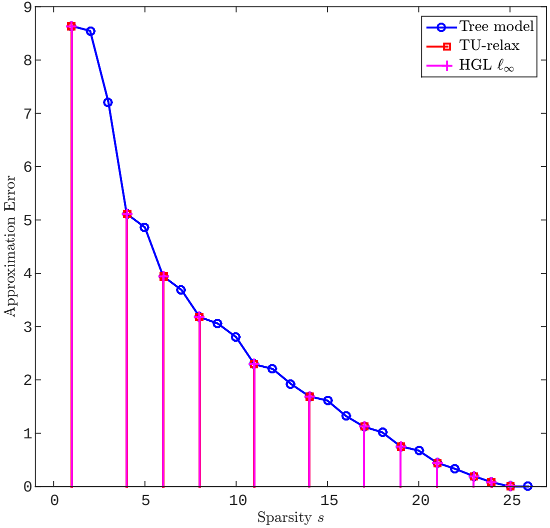
Tree sparsity example: Pareto frontier [8, 5]



Tree sparsity example: Pareto frontier [8, 5]

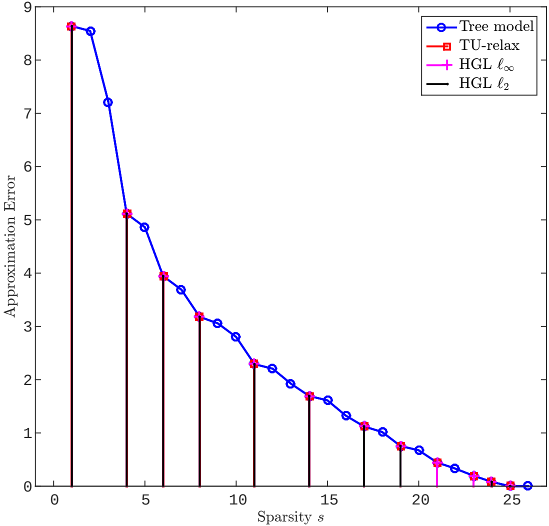


Tree sparsity example: Pareto frontier [8, 5]

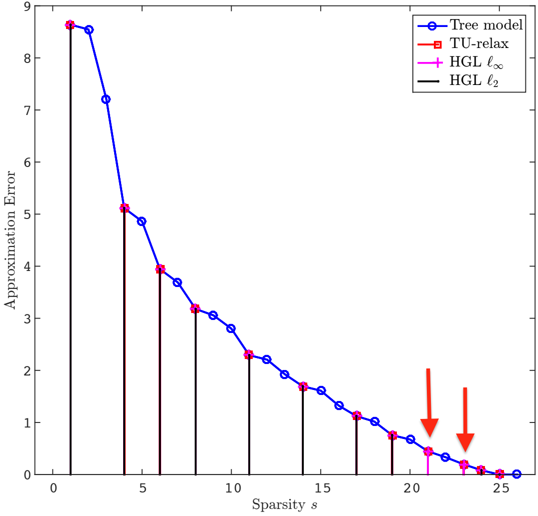


Just kidding, they are the same.

Tree sparsity example: Pareto frontier [8, 5]

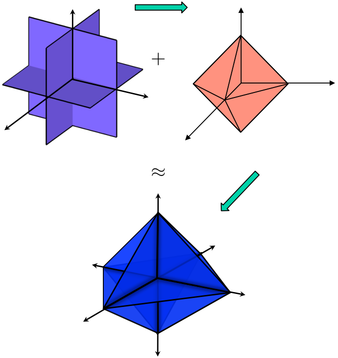


Tree sparsity example: Pareto frontier [8, 5]

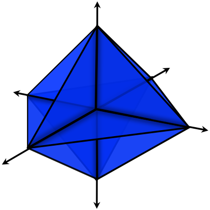


CLASH [22]

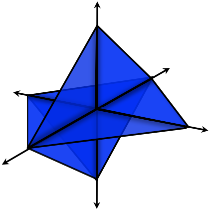
Combinatorial Selection
$+$
Least Absolute Shrinkage



CLASH set:



Model-CLASH set:



Retains the same theoretical guarantees

combinatorial origami

ILP and matroid structured models...

Outline

Total unimodularity in discrete optimization

From sparsity to *structured sparsity*

Convex relaxations for structured sparse recovery

Enter nonconvexity

Conclusions

Conclusions

Our work: TU modeling framework & convex template & non-convex algorithms

- Many more convex programs (not necessarily norms)
- TU models: tight convexifications, non-submodular examples
- Easy to design and "usually" efficient via an LP
- London calling...

Alternatives:

1. Atomic norms [11, 10]
   - Given a set $\mathcal{A}$, use the biconjugation of $g(\mathbf{x}) = \inf_{0\leq t\leq c} t + \iota_{t\mathcal{A}}(\mathbf{x})$, for $c > 0$
   - Reverse engineer the set to obtain structured sparsity
   - "Usually" tractable since the norm is reverse engineered
2. Monotone submodular penalties and extensions [3]
   - Tight convexification via Lovász extension
   - Reverse engineer the submodular set function (not always possible)
3. ℓ_q-regularized combinatorial functions [24]
   - Tight convexification (also explains latent group lasso like norms)
   - Not always efficiently computable
   - Reverse engineered and may loose structure, e.g., group knapsack model

References I

[1] Ben Adcock, Anders C. Hansen, Clarice Poon, and Bogdan Roman.
Breaking the coherence barrier: A new theory for compressed sensing.
http://arxiv.org/abs/1302.0561, Feb. 2013.

[2] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp.
Living on the edge: Phase transitions in convex programs with random data.
Information and Inference, 3:224–294, 2014.
arXiv:1303.6672v2 [cs.IT].

[3] Francis Bach.
Structured sparsity-inducing norms through submodular functions.
Adv. Neur. Inf. Proc. Sys. (NIPS), pages 118–126, 2010.

[4] B. Bah, L. Baldassarre, and V. Cevher.
Model-based sketching and recovery with expanders.
In *Proc. ACM-SIAM Symp. Disc. Alg.*, number EPFL-CONF-187484, 2014.

[5] L. Baldassarre, N. Bhan, V. Cevher, and A. Kyrillidis.
Group-sparse model selection: Hardness and relaxations.
arXiv preprint arXiv:1303.3207, 2013.

[6] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
IEEE Trans. Inf. Theory, 56(4):1982–2001, April 2010.

References II

[7] Peter L. Barlett and Shahar Mendelson.
Rademacher and Gaussian complexities: Risk bounds and structural results.
J. Mach. Learn. Res., 3, 2002.

[8] N. Bhan, L. Baldassarre, and V. Cevher.
Tractability of interpretability via selection of group-sparse models.
In *IEEE Int. Symp. Inf. Theory*, 2013.

[9] Venkat Chandrasekaran and Michael I. Jordan.
Computational and statistical tradeoffs via convex relaxation.
Proc. Nat. Acad. Sci., 110(13):E1181–E1190, 2013.

[10] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky.
The convex geometry of linear inverse problems.
Found. Comp. Math., 12:805–849, 2012.

[11] S. Chen, D. Donoho, and M. Saunders.
Atomic decomposition by basis pursuit.
SIAM J. Sci. Comp., 20(1):33–61, 1998.

[12] William Cook, László Lovász, and Alexander Schrijver.
A polynomial-time test for total dual integrality in fixed dimension.
In *Mathematical programming at Oberwolfach II*, pages 64–69. Springer, 1984.

References III

[13] Marco F. Duarte, Dharmpal Davenport, Mark A. adn Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk.
Single-pixel imaging via compressive sampling.
IEEE Sig. Proc. Mag., 25(2):83–91, March 2008.

[14] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
preprint, 2014.
arXiv:1411.1990v1 [cs.LG].

[15] S. Fujishige.
Submodular functions and optimization, volume 58.
Elsevier Science, 2005.

[16] W Gerstner and W. Kistler.
Spiking neuron models: Single neurons, populations, plasticity.
Cambridge university press, 2002.

[17] FR Giles and William R Pulleyblank.
Total dual integrality and integer polyhedra.
Linear algebra and its applications, 25:191–196, 1979.

References IV

[18] C. Hegde, M. Duarte, and V. Cevher.
Compressive sensing recovery of spike trains using a structured sparsity model.
In *Sig. Proc. with Adapative Sparse Struct. Rep. (SPARS)*, 2009.

[19] J. Huang, T. Zhang, and D. Metaxas.
Learning with structured sparsity.
J. Mach. Learn. Res., 12:3371–3412, 2011.

[20] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion.
Multi-scale mining of fmri data with hierarchical structured sparsity.
In *Pattern Recognition in NeuroImaging (PRNI)*, 2011.

[21] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
J. Mach. Learn. Res., 12:2297–2334, 2011.

[22] A. Kyrillidis and V. Cevher.
Combinatorial selection and least absolute shrinkage via the CLASH algorithm.
In *IEEE Int. Symp. Inf. Theory*, pages 2216–2220. Ieee, 2012.

[23] George L Nemhauser and Laurence A Wolsey.
Integer and combinatorial optimization, volume 18.
Wiley New York, 1999.

References V

[24] G. Obozinski and F. Bach.
Convex relaxation for combinatorial penalties.
arXiv preprint arXiv:1205.1240, 2012.

[25] G. Obozinski, L. Jacob, and J.P. Vert.
Group lasso with overlaps: The latent group lasso approach.
arXiv preprint arXiv:1110.0413, 2011.

[26] G. Obozinski, B. Taskar, and M.I. Jordan.
Joint covariate selection and joint subspace selection for multiple classification problems.
Statistics and Computing, 20(2):231–252, 2010.

[27] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi.
Simple bounds for noisy linear inverse problems with exact side information.
2013.
arXiv:1312.0641v2 [cs.IT].

[28] C Seshadhri and Jan Vondrák.
Is submodularity testable?
Algorithmica, 69(1):1–25, 2014.

References VI

[29] Paul D Seymour.
Decomposition of regular matroids.
Journal of combinatorial theory, Series B, 28(3):305–359, 1980.

[30] Quoc Tran-Dinh and Volkan Cevher.
Constrained convex minimization via model-based excessive gap.
In *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2014.

[31] Klaus Truemper.
Alpha-balanced graphs and matrices and GF(3)-representability of matroids.
J. Comb. Theory Ser. B, 32(2):112–139, 1982.

[32] Roman Vershynin.
Estimation in high dimensions: a geometric perspective.
http://arxiv.org/abs/1405.5103, May 2014.

[33] Peng Zhao, Guilherme Rocha, and Bin Yu.
Grouped and hierarchical model selection through composite absolute penalties.
Department of Statistics, UC Berkeley, Tech. Rep, 703, 2006.

[34] Peng Zhao and Bin Yu.
On model selection consistency of Lasso.
J. Mach. Learn. Res., 7:2541–2563, 2006.

References VII

[35] H. Zhou, M.E. Sehl, J.S. Sinsheimer, and K. Lange.
Association screening of common and rare genetic variants by penalized regression.

Bioinformatics, 26(19):2375, 2010.