Sparse Prediction with the k-Support Norm

Andreas Argyriou, Rina Foygel, and Nathan Srebro

Kuo-Shih Tseng

2014.04.02

Sparse regression

$$Y_{(m,1)} = X_{(m,n)} W_{(n,1)}$$

Y : mesurement $\Rightarrow$ How many measurements?

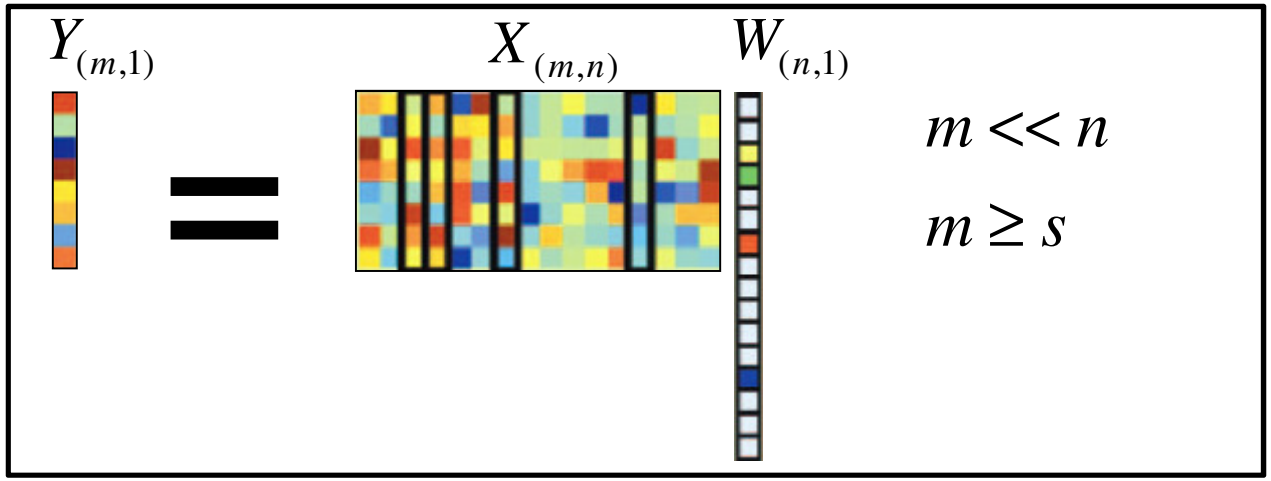
SP5

X : design matrix $\Rightarrow$ Which condition?

SP6, SP7

W : sparse vector $\Rightarrow W$ is what we want?

Opt9

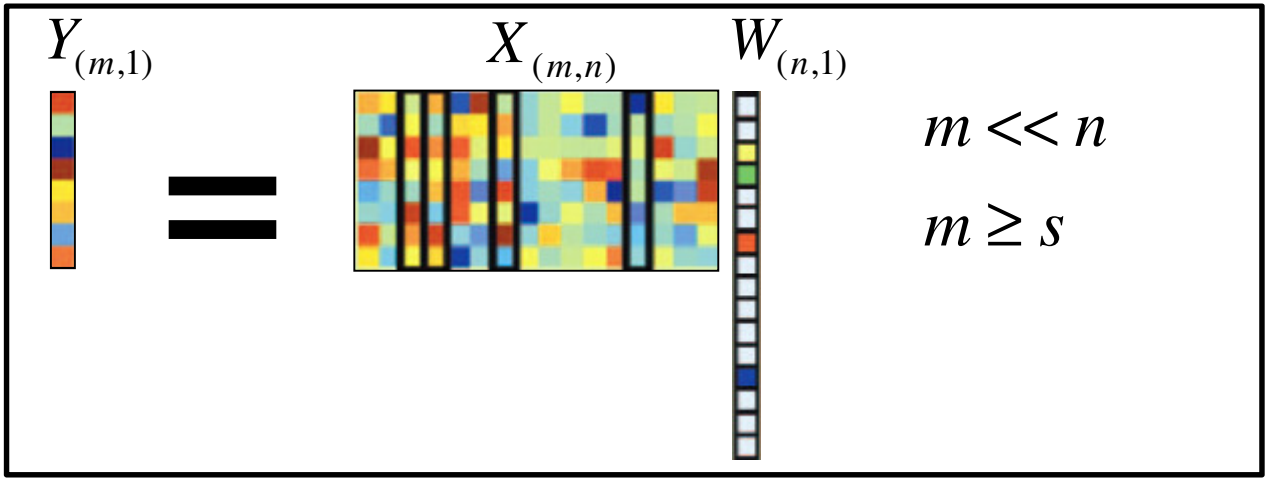


$$\hat{W} = \min_{W} \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 \right\}$$

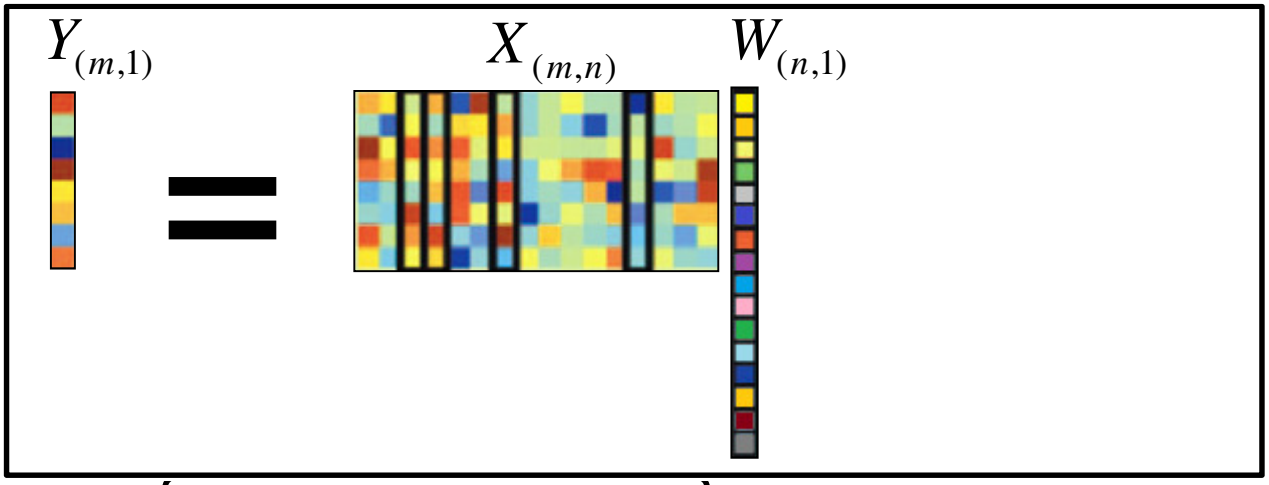
Outline

- Sparsity v.s. Accuracy
- K-support norm
- Compute K-support norm
- Optimization algorithm
- K-support norm v.s. Elastic net
- Experiments

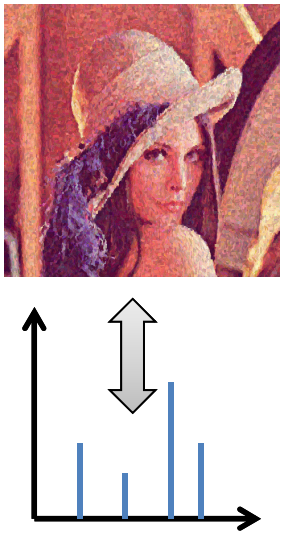
Sparsity v.s. Accuracy

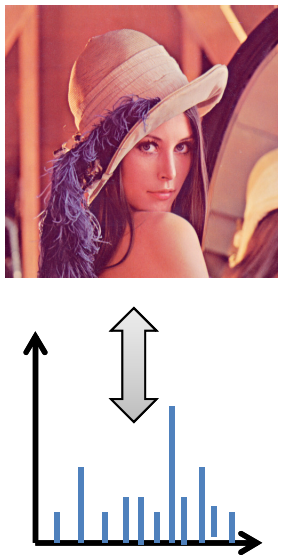


$$\min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 \right\}$$



$$\min_W \left\{ \|Y - XW\|_2^2 + \lambda_2 \|W\|_2^2 \right\}$$





Sparsity v.s. Accuracy

- Trade off between sparsity and accuracy

Lasso : **(sparse solutions)**

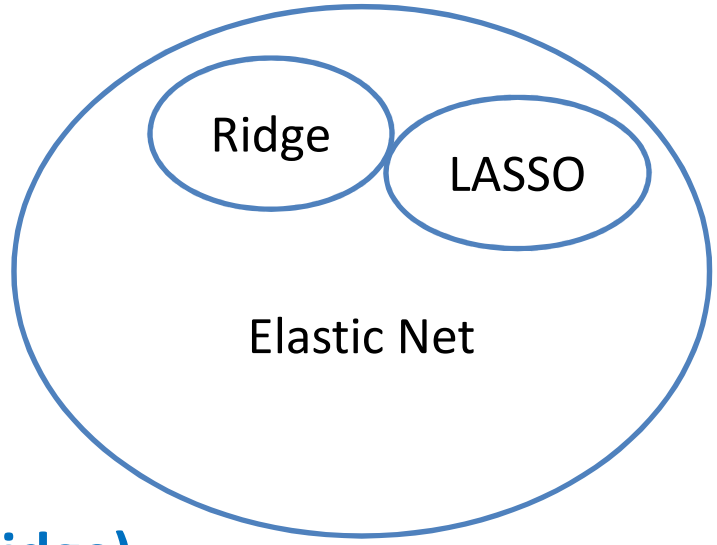
$$\min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 \right\}$$

Ridge : **(group solutions)**

$$\min_W \left\{ \|Y - XW\|_2^2 + \lambda_2 \|W\|_2^2 \right\}$$

Elastic Net : **(consisted of LASSO & Ridge)**

$$\min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2^2 \right\}$$



Sparsity v.s. Accuracy

$$\text{Elastic net}: \quad \min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2^2 \right\}$$

$$\text{Motivation}: \quad \min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_0 + \lambda_2 \|W\|_2^2 \right\}$$

Non-convex?

relaxation

$$\text{K-support norm}:$$

$$\min_W \left\{ \|Y - XW\|_2^2 + \frac{\lambda}{2} \left(\|W\|_k^{sp} \right)^2 \right\}$$

This paper

- What's k-support norm
- How to compute the k-support norm?
- How to learn the k-support norm problem?
- How "loose" a relaxation is the Elastic Net?

$$\text{K-support norm}: \min_W \left\{ \|Y - XW\|_2^2 + \frac{\lambda}{2} \left(\|W\|_k^{sp} \right)^2 \right\}$$

$$\text{Elastic net}: \quad \min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2^2 \right\}$$

$$\text{Lasso}: \quad \min_W \left\{ \|Y - XW\|_2^2 + \lambda_1 \|W\|_1 \right\}$$

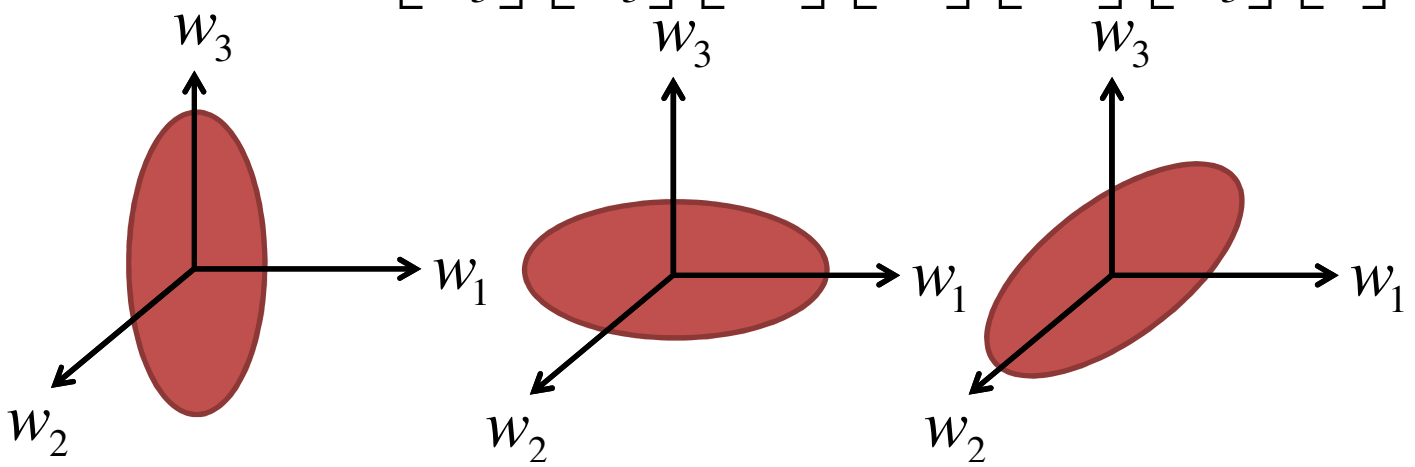
K-support norm

- Sparse signal set

$$S_k^{(2)} = \left\{ w \in R^d \mid \|w\|_0 \le k, \|w\|_2 \le 1 \right\}$$

Suppose $k = 2,\ d = 3$ $\quad (2-\text{support norm in } R^3)$

$$\|w\|_0 \le 2 \Rightarrow \begin{bmatrix} 0 \\ w_2 \\ w_3 \end{bmatrix}, \begin{bmatrix} w_1 \\ 0 \\ w_3 \end{bmatrix}, \begin{bmatrix} w_1 \\ w_2 \\ 0 \end{bmatrix}, \begin{bmatrix} w_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ w_2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ w_3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Non-convex?

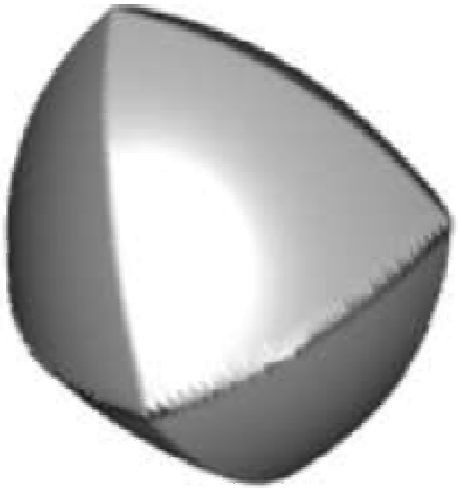
K-support norm

- Unit ball of K-support norm

  $C_k = conv(S_k^{(2)})$

- Accurate / Sparse?

K-support norm



Elastic net

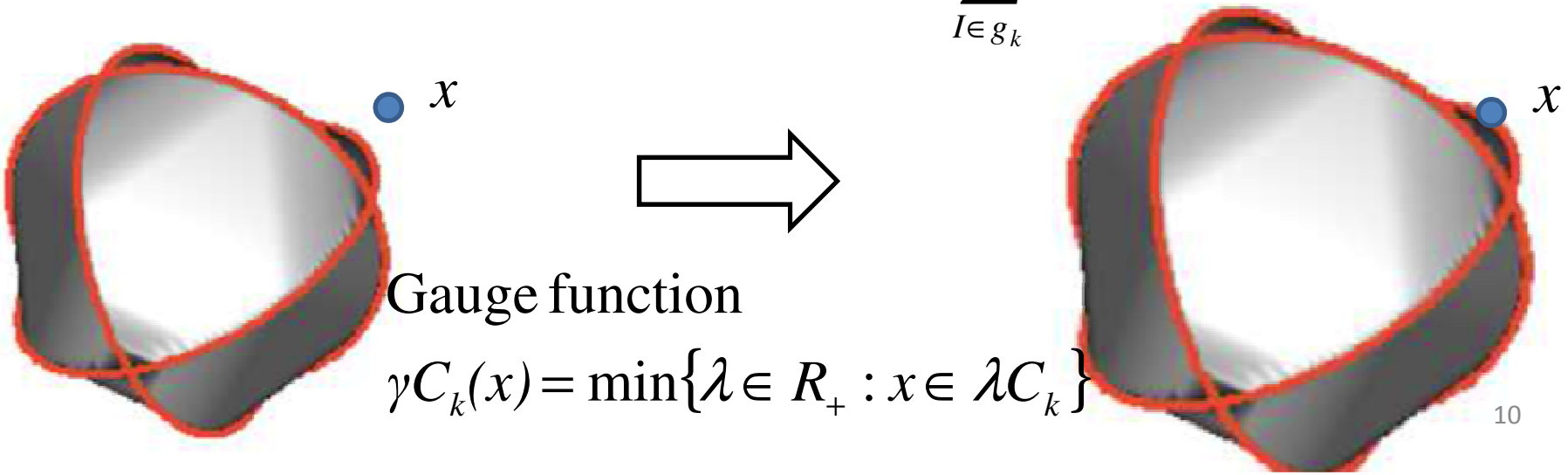
K-support norm

Defintion 2.1 Let $k \in \{1,...,d\}$. The k-support norm $\|\bullet\|_k^{sp}$ is defined, for every $w \in R^d$, *as*

$$\|w\|_k^{sp} := \min\left\{ \sum_{I \in g_k} \|v_I\|_2 : \mathrm{supp}(v_I) \subseteq I, \sum_{I \in g_k} v_I = w \right\}$$

g_k is the set of all subset of $\{1,...,d\}$ of cardinality at most k.

where $v_I = \mu_I z_I, \mu_I \geq 0, z_I \in C_k, \forall I \in g_k, \sum_{I \in g_k} \mu_I = 1$



Gauge function

$$\gamma C_k(x) = \min\{\lambda \in R_+ : x \in \lambda C_k\}$$

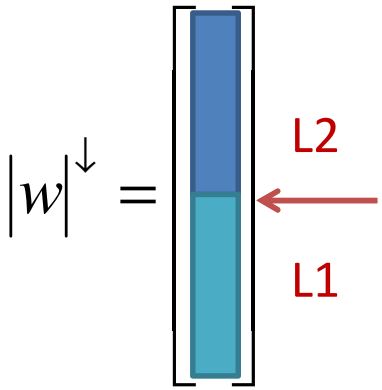
Compute K-support norm

Proposition 2.1

1) Find r $\quad$ s.t. $|w|^{\downarrow}_{k-r-1} > \frac{1}{r+1}\sum_{i=k-r}^{d}|w|^{\downarrow}_{i} \geq |w|^{\downarrow}_{k-r}$

$r \in \{0,1,\ldots\ldots,k-1\}$, $|w|^{\downarrow}_{i}$ is the i-th largest element of w

2) Compute $\|w\|^{sp}_{k} = \sqrt{\sum_{i=1}^{k-r-1}\left(|w|^{\downarrow}_{i}\right)^2 + \frac{1}{r+1}\left(\sum_{i=k-r}^{d}|w|^{\downarrow}_{i}\right)^2}$



$$\|w\|^{sp}_{k} = \sqrt{\underbrace{\sum_{i=1}^{k-r-1}\left(|w|^{\downarrow}_{i}\right)^2}_{l2-norm \text{ on larger terms}} + \frac{1}{r+1}\underbrace{\left(\sum_{i=k-r}^{d}|w|^{\downarrow}_{i}\right)^2}_{l1-norm \text{ on smaller terms}}}$$

Compute K-support norm

Proposition 2.1

1) Find r $\quad$ s.t. $|w|^{\downarrow}_{k-r-1} > \frac{1}{r+1}\sum_{i=k-r}^{d}|w|^{\downarrow}_{i} \geq |w|^{\downarrow}_{k-r}$

$r \in \{0,1,\ldots\ldots,k-1\}$, $|w|^{\downarrow}_{i}$ is the i-th largest element of w

2) Compute $\|w\|^{sp}_{k} = \sqrt{\sum_{i=1}^{k-r-1}\left(|w|^{\downarrow}_{i}\right)^2 + \frac{1}{r+1}\left(\sum_{i=k-r}^{d}|w|^{\downarrow}_{i}\right)^2}$

EX:

$$|w|^{\downarrow} = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 0 \end{bmatrix} \begin{matrix} \\ \leftarrow \text{L2} \\ \\ \text{L1} \end{matrix}$$

$d = 4, k = 3$

1) find $r \in \{0,1,2\}$

$r = 0$

$|w|^{\downarrow}_{2} > \sum_{i=3}^{4}|w|^{\downarrow}_{i} \geq |w|^{\downarrow}_{3}$

$2 > 1 \geq 1$ (satisfy!)

2) compute $\|w\|^{sp}_{k}$

$$= \sqrt{\sum_{i=1}^{2}\left(|w|^{\downarrow}_{i}\right)^2 + \left(\sum_{i=3}^{4}|w|^{\downarrow}_{i}\right)^2}$$

$= \sqrt{14}$

Compute K-support norm

Proof of Proposition 2.1

Conjugate function

$$f^*(w) = \sup_{u \in dom\ f} \left(\langle u, w \rangle - f(u) \right)$$

Dual Norm

$$\|u\|_k^{sp*} = \max \left\{ \langle x, y \rangle : \|w\|_k^{sp} \le 1 \right\}$$

$$= \max \left\{ \left(\sum_{i \in I} u_i^2 \right)^{\frac{1}{2}} : I \in g_k \right\} = \left(\sum_{i=1}^{k} \left(|u|_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}} =: \|u\|_k^2$$

g_k is the set of all subset of $\{1,...,d\}$ of cardinality at most k

$$\therefore \frac{1}{2} \left(\|w\|_k^{sp} \right)^2 = \max \left\{ \langle u, w \rangle - \frac{1}{2} \left(\|u\|_k^2 \right)^2 : u \in R^d \right\}$$

Compute K-support norm

Proof of Proposition 2.1

$$\frac{1}{2}\left(\|w\|_k^{sp}\right)^2 = \max\left\{\langle u, w\rangle - \frac{1}{2}\left(\|u\|_k^2\right)^2 : u \in R^d\right\}$$

$$= \max\left\{\sum_{i=1}^{d} \alpha_i |w|_i^{\downarrow} - \frac{1}{2}\sum_{i=1}^{k} \alpha_i^2 : \alpha_1 \geq \ldots . \alpha_d \geq 0\right\}$$

$$= \max\left\{\sum_{i=1}^{k-1} \alpha_i |w|_i^{\downarrow} + \alpha_k \sum_{i=k}^{d} |w|_i^{\downarrow} - \frac{1}{2}\sum_{i=1}^{k} \alpha_i^2 : \alpha_1 \geq \ldots . \alpha_k \geq 0\right\}$$

$$A_r = \sum_{i=k-r}^{d} |w|_i^{\downarrow}$$

$$|w|^{\downarrow} = \left[\quad\right] \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{k-r} \\ \\ \\ \\ A_r \end{matrix}$$

1) $A_0 < |w|_{k-1}^{\downarrow}$ (r=0)

$$\begin{cases} \alpha_i = |w|_i^{\downarrow}, i = 1, \ldots, (k-1) \\ \alpha_i = A_0, i = k, \ldots, d \end{cases}$$

2) $A_0 \geq |w|_{k-1}^{\downarrow}$

$|w|_{k-1}^{\downarrow} \leq optimal\ \alpha_k, \alpha_{k-1} \leq A_0$

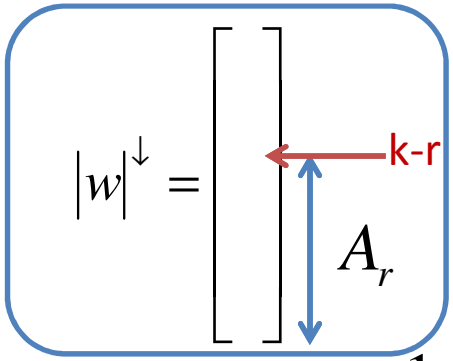
Compute K-support norm

$$= \max\left\{ \sum_{i=1}^{k-1} \alpha_i |w|_i^{\downarrow} + \alpha_k \sum_{i=k}^{d} |w|_i^{\downarrow} - \frac{1}{2} \sum_{i=1}^{k} \alpha_i^2 : \alpha_1 \geq .... \alpha_k \geq 0 \right\}$$

$$= \max\left\{ \sum_{i=1}^{k-2} \alpha_i |w|_i^{\downarrow} - \frac{1}{2} \sum_{i=1}^{k-2} \alpha_i^2 + A_1 \alpha_{k-1} - \alpha_{k-1}^2 : \alpha_1 \geq .... \alpha_{k-1} \geq 0 \right\}$$

$$A_0 \geq |w|_{k-1}^{\downarrow}, |w|_{k-2}^{\downarrow} > \frac{A_1}{2}$$

$$\vdots \quad (induction)$$



$$\frac{A_{r-1}}{r} \geq |w|_{k-r}^{\downarrow}, |w|_{k-r-1}^{\downarrow} > \frac{A_r}{r+1} \Rightarrow |w|_{k-r-1}^{\downarrow} > \frac{1}{r+1} \sum_{i=k-r}^{d} |w|_i^{\downarrow} \geq |w|_{k-r}^{\downarrow}$$

$$\begin{cases} \alpha_i = |w|_i^{\downarrow}, i = 1,..., k-r-1 \\ \alpha_i = \dfrac{A_r}{r+1}, i = k-r,....., d \end{cases} \Rightarrow \|w\|_k^{sp} = \sqrt{\sum_{i=1}^{k-r-1} \left(|w|_i^{\downarrow}\right)^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{d} |w|_i^{\downarrow} \right)^2}$$

Optimization algorithm

$$\min_W \left\{ \|y - XW\|_2^2 + \frac{\lambda}{2} \left(\|W\|_k^{sp} \right)^2 \right\}$$

Algorithm 2 Accelerated K-support regularization

$w_1 = \alpha_1 \in R^d, \theta_1 \leftarrow 1, L - Lipschitz\ gradient$

for $t = 1,2,.....$ do

$$\theta_{t+1} \leftarrow \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}$$

$$w_{t+1} \leftarrow \text{prox}_{\frac{\lambda}{2L}\left(\|\bullet\|_k^{sp}\right)^2} \left(\alpha_t - \frac{1}{L} X^T \left(X\alpha_t - y \right) \right) \text{using algorithm1}$$

$$\alpha_{t+1} \leftarrow w_{t+1} + \frac{\theta_t - 1}{\theta_{t+1}} \left(w_{t+1} - w_t \right)$$

end for

Optimization algorithm

Algorithm 1 *Computation of the proximity operator*

Input : $v \in R^d$, **Output** : $q = prox_{\frac{\lambda}{2L}\left(\|\bullet\|_2^{sp}\right)^2}(v)$

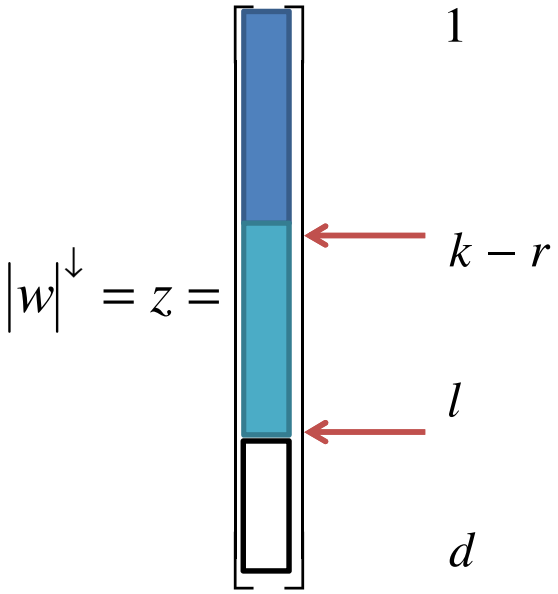
Find $r \in \{0,\ldots,k-1\}, l \in \{k,\ldots,d\}$ s.t.

$$\begin{cases} \dfrac{1}{L+1} z_{k-r-1} > \dfrac{T_{r,l}}{l-k+(L+1)r+L+1} \geq \dfrac{1}{L+1} z_{k-r} \\ z_l > \underbrace{\dfrac{T_{r,l}}{l-k+(L+1)r+L+1}}_{A_W} \geq z_{l+1} \end{cases}$$

where $z := |v|^{\downarrow}, T_{r,l} := \sum_{i=k-r}^{l} z_i$ →

$$q_i \leftarrow \begin{cases} \dfrac{L}{L+1} z_i, i = 1,\ldots,k\text{-}r\text{-}1 \\ z_i - A_W, \quad i = k\text{-}r,\ldots,l \\ 0, \qquad i = l+1,\ldots,d \end{cases}$$

reorder and change signs of q



Complexity :

$$O(\underbrace{dk}_{find\ r,l} + \underbrace{d \log d}_{sorting})$$

(relative shrinkage)

(soft thresholding)

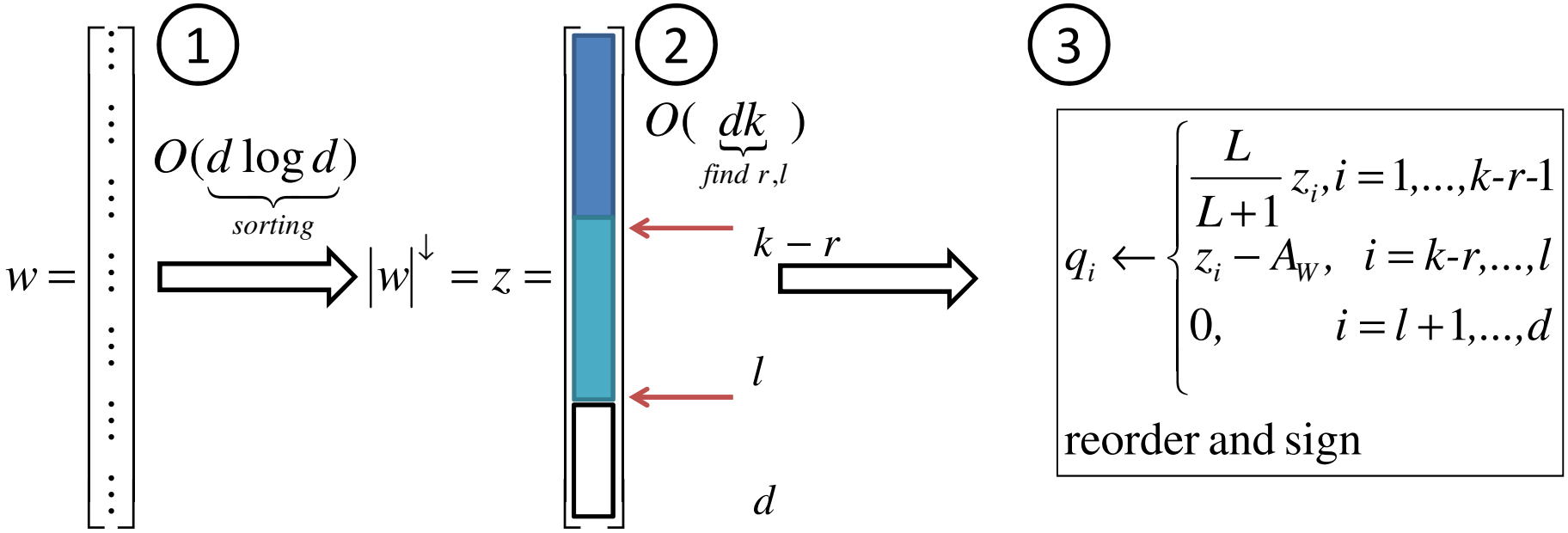
$$sign(w_i)\left(|w_i| - A_W\right)_+$$

Optimization algorithm

- Parameters (λ, k)

k doesn't mean $k-sparse \Rightarrow |w|_0 \le k$

$r \in \{0,\ldots,k-1\}, l \in \{k,\ldots,d\}$



Optimization algorithm

- Proof of algorithm 1

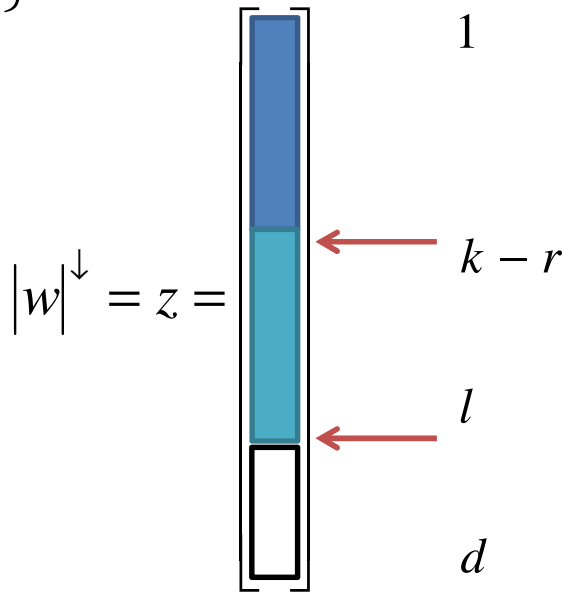
$$prox_w(x) = \arg\min\left\{\frac{1}{2}\|u-x\|^2 + w(u) : u \in R^d\right\}$$

$$Lz - Lq = Lv - Lq \in \partial\frac{1}{2}\left(\|\bullet\|_k^{sp}\right)^2(q)$$

$$\sum_{i=k-r}^{d} q_i = \sum_{i=k-r}^{l}\left(z_i - \frac{T_{r,l}}{l-k+(L+1)r+L+1}\right)$$

$$= T_{r,l} - \frac{(l-k+r+1)T_{r,l}}{l-k+(L+1)r+L+1}$$

$$= L(r+1)\frac{T_{r,l}}{l-k+(L+1)r+L+1} := A_r$$



Optimization algorithm

$$q_i \leftarrow \begin{cases} \dfrac{L}{L+1} z_i, i = 1,\ldots,k\text{-}r\text{-}1 \\ z_i - A_W, \quad i = k\text{-}r,\ldots,l \\ 0, \qquad\quad i = l+1,\ldots,d \end{cases}$$

(relative shrinkage)

(soft thresholding)

1) $i \le k - r - 1$

$$Lz_i - Lq_i = q_i \Rightarrow q_i = \frac{L}{L+1} z_i$$

$$A_r = L(r+1)\frac{T_{r,l}}{l-k+(L+1)r+L+1}$$

2) $k - r \le i \le l$

$$Lz_i - Lq_i = \frac{1}{r+1} A_r \Rightarrow q_i = z_i - \frac{1}{L(r+1)} A_r = z_i - A_W$$

3) $i \ge l$

$q_i = 0$

$$sign(w_i)\left(|w_i| - A_W\right)_+$$

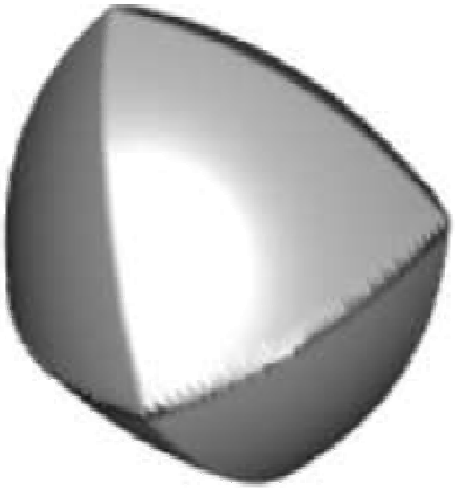
K-support norm v.s. Elastic net

Proposition 3.1 $\|w\|_k^{el} \le \|w\|_k^{sp} < \sqrt{2}\|w\|_k^{el}$

$$\|w\|_k^{el} = \max\left\{\|w\|_2, \frac{\|w\|_1}{\sqrt{k}}\right\}, \|w\|_k^{sp} > \langle w, u\rangle$$

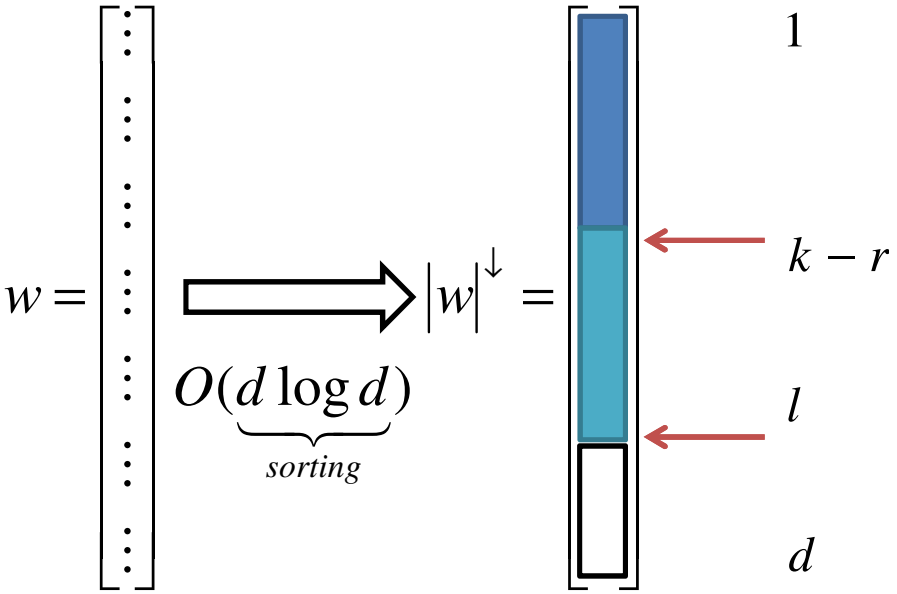
$$w = \left(k^{1.5}, \underbrace{1,1,\ldots 1}_{k^2}\right)^T, u = \left(\frac{1}{\sqrt{2}}, \underbrace{\frac{1}{\sqrt{2k}}, \ldots, \frac{1}{\sqrt{2k}}}_{k^2}\right)^T$$

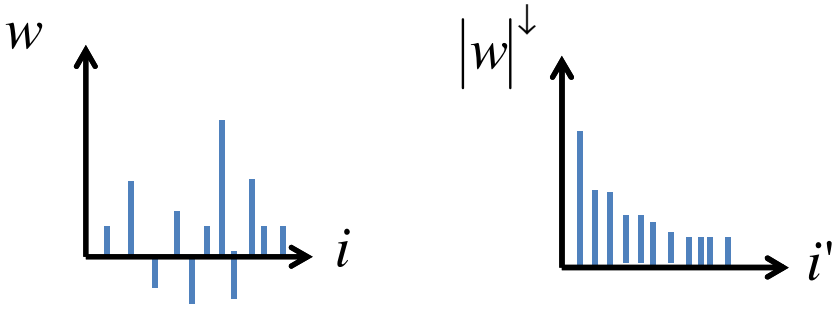
$$\begin{cases} \|w\|_k^{el} = k^{1.5}\left(1+\frac{1}{\sqrt{k}}\right) \Rightarrow \|w\|_k^{sp} < \sqrt{2}\|w\|_k^{el} \\ \|w\|_k^{sp} > \sqrt{2}k^{1.5} \end{cases}$$



K-support norm v.s. Group LASSO

- Group based on the **magnitude** of weighting





Unknown structure

↓

K – support norm

Known structure

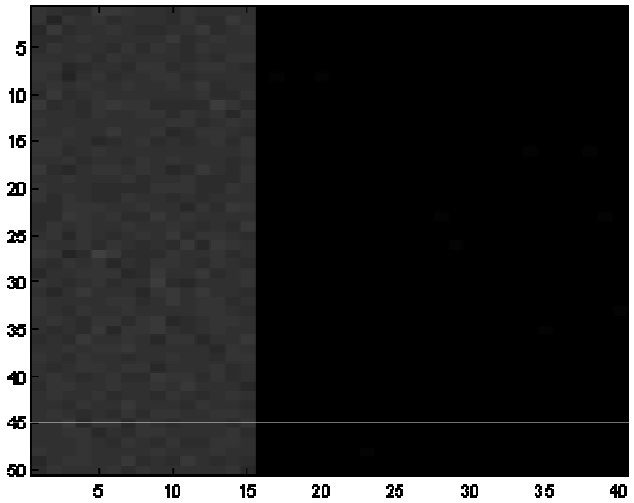
↓

Group LASSO

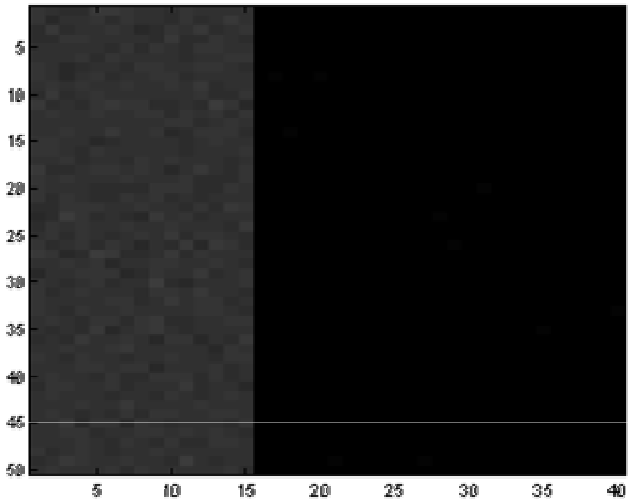
Experiments

- Synthetic data
  - d=40 (sparse regression), strong correlations
  - Sample size 450= 50 training + 50 validation + 350 test
- South African heart data
  - d=9 (classification)
  - Sample size 462= 400 training + 30 validation + 32 test
- 20 Newsgroup
  - (classification)
  - Sample size 20k = 14k training + 1k validation + 5k test

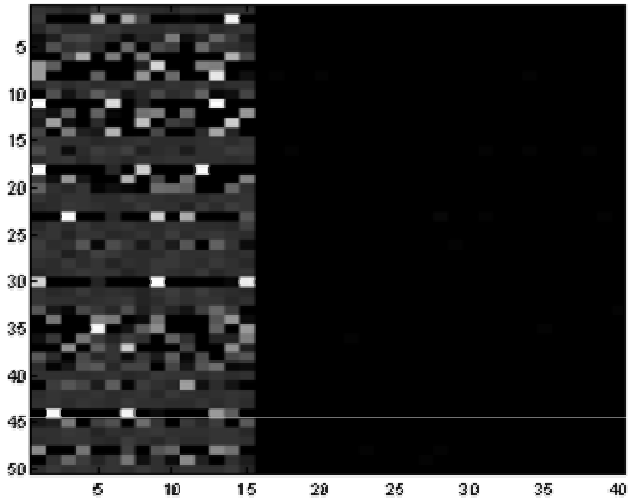
Experiments



K-support



LASSO



Elastic net

	Synthetic	Heart		Newsgroups	
Method	MSE (SE)	MSE (SE)	Accuracy (SE)	MSE	Accuracy
Lasso	0.2746 (0.02)	0.18 (0.005)	66.41 (0.53)	0.70	73.02
Elastic net	0.3119 (0.03)	0.18 (0.005)	66.41 (0.53)	0.71	72.53
k-support	**0.2342** (0.02)	0.18 (0.005)	66.41 (0.53)	**0.69**	**73.40**

???

???

Accuracy: K-support > Elastic net > Lasso

Experiments

- Paper

Accuracy: K-support > Elastic net > Lasso
Sparsity: K-support < Elastic net < Lasso

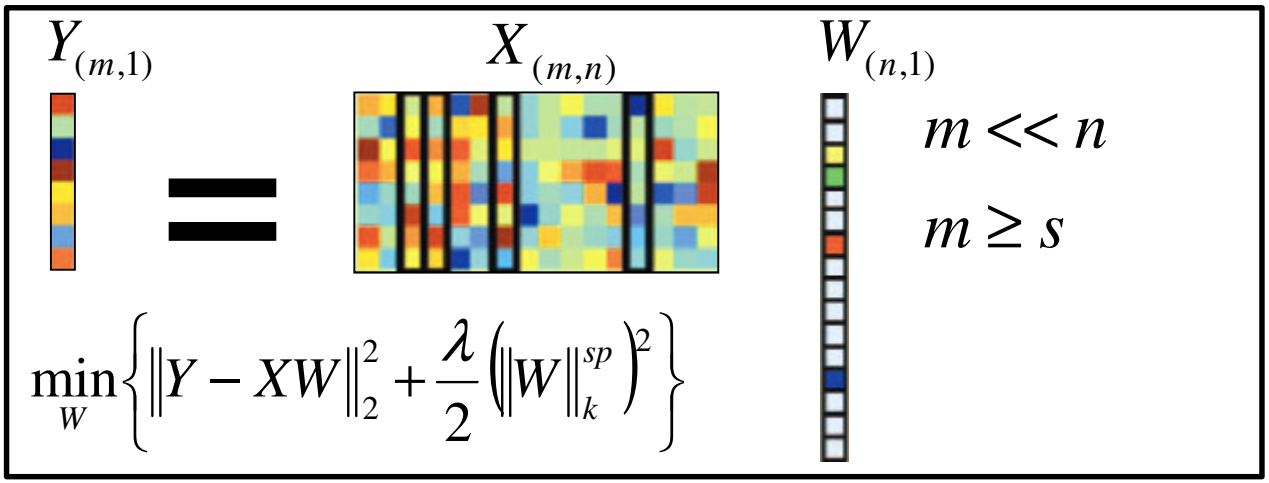
	Synthetic	Heart		Newsgroups	
Method	MSE (SE)	MSE (SE)	Accuracy (SE)	MSE	Accuracy
Lasso	0.2746 (0.02)	0.18 (0.005)	66.41 (0.53)	0.70	73.02
Elastic net	0.3119 (0.03)	0.18 (0.005)	66.41 (0.53)	0.71	72.53
k-support	**0.2342** (0.02)	0.18 (0.005)	66.41 (0.53)	**0.69**	**73.40**

- Poster

	synthetic data	heart data	20 newsgroups
	MSE (SE) over 50 trials	MSE (SE) over 50 trials	MSE
Lasso	0.2685 (0.02)	0.18 (0.005)	0.70
Elastic net	0.2274 (0.02)	0.18 (0.005)	0.70
k-support norm	**0.2143** (0.02)	0.18 (0.005)	**0.69**

Conclusions

- K-support norm
  - More accurate (less sparse) norm than elastic net
  - An efficient algorithm to compute it
  - An accelerated algorithm to learn it
- Apply to compressed sensing?



Reference

- Andreas Argyriou, Rina Foygel, and Nathan Srebro, ”Sparse Prediction with the k-Support Norm,” NIPS, 2012. (paper & poster)