Convex relaxation for Combinatorial Penalties

Guillaume Obozinski

Equipe Imagine
Laboratoire d'Informatique Gaspard Monge
Ecole des Ponts - ParisTech

École des Ponts
ParisTech

Joint work with Francis Bach

Fête Parisienne in Computation, Inference and Optimization
IHES - March 20th 2013

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \left|\mathrm{Supp}(w)\right|$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

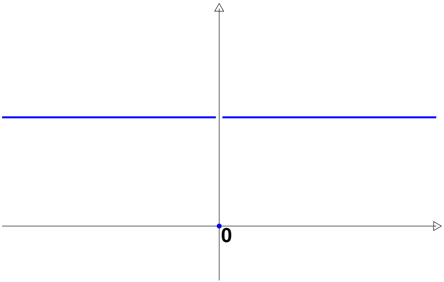
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

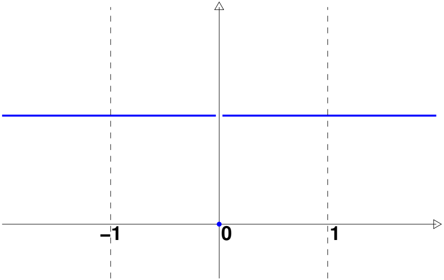
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

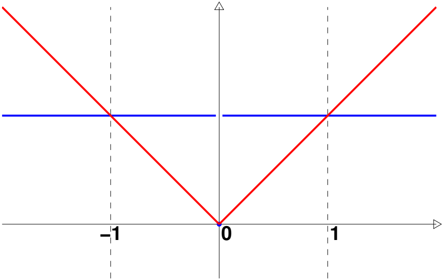
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

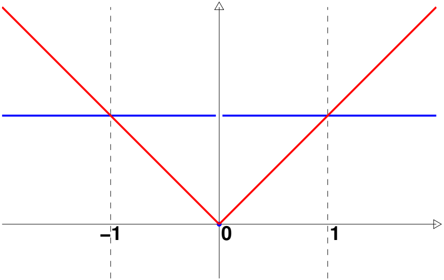
- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$



Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

Lasso

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \|w\|_1$$

… to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

… to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

Examples

- The variables should be selected in groups.

… to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

Examples

- The variables should be selected in groups.
- The variables lie in a hierarchy.

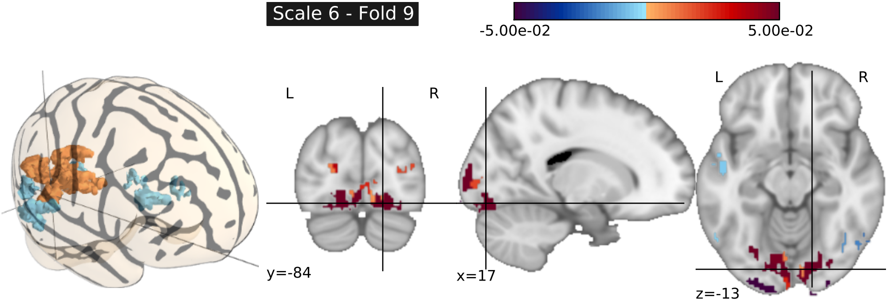
... to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

Examples

- The variables should be selected in groups.
- The variables lie in a hierarchy.
- The variables lie on a graph or network and the support should be localized or densely connected on the graph.

Difficult inverse problem in Brain Imaging



Jenatton et al. (2012)

Hierarchical Dictionary Learning

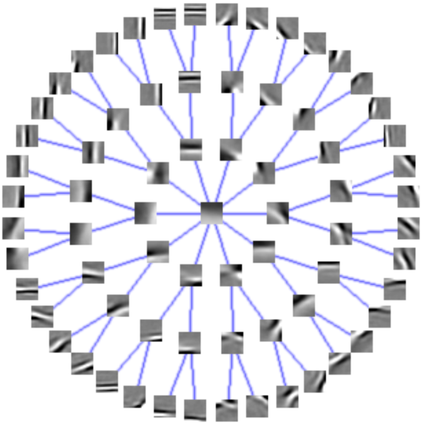


Figure: Hierarchical dictionary of image patches

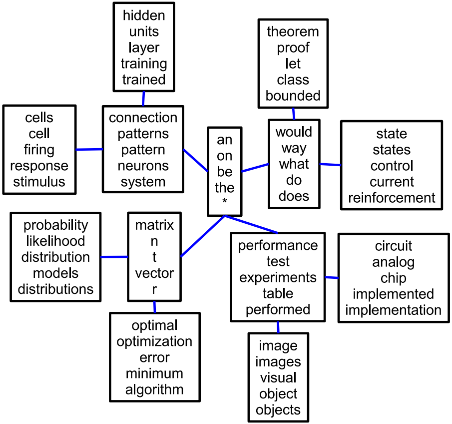


Figure: Hierarchical Topic model

Mairal, Jenatton, Obozinski and Bach (2010)

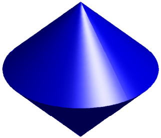
Ideas in structured sparsity

Group Lasso and ℓ_1/ℓ_p norm (Yuan and Lin, 2006)

Group Lasso

Given $\mathcal{G} = \{A_1, \ldots, A_m\}$ a partition of $V := \{1, \ldots, d\}$ consider

$$\|w\|_{\ell_1/\ell_p} = \sum_{A \in \mathcal{G}} \delta^A \|w_A\|_p$$

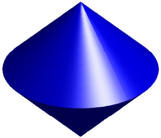


Group Lasso and ℓ_1/ℓ_p norm (Yuan and Lin, 2006)

Group Lasso

Given $\mathcal{G} = \{A_1, \ldots, A_m\}$ a partition of $V := \{1, \ldots, d\}$ consider

$$\|w\|_{\ell_1/\ell_p} = \sum_{A \in \mathcal{G}} \delta^A \|w_A\|_p$$

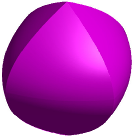


Overlapping groups: direct extension of Jenatton et al. (2011).

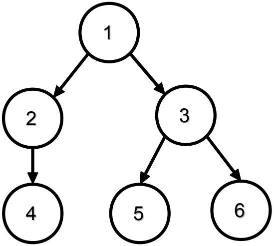
Interesting induced structures

→ Induce patterns of rooted subtree

→ Induce "convex" patterns on a grid



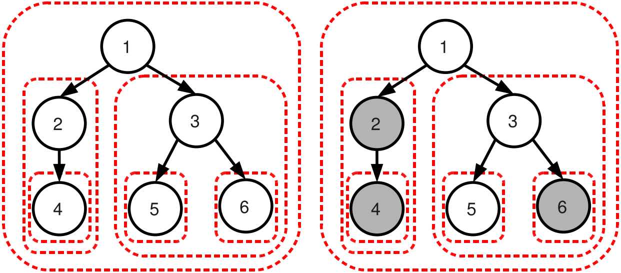
Hierarchical Norms (Zhao et al., 2009; Bach, 2008)



(Jenatton, Mairal, Obozinski and Bach, 2010a)

- A covariate can only be selected after its ancestors
- Structure on parameters w

Hierarchical Norms (Zhao et al., 2009; Bach, 2008)



(Jenatton, Mairal, Obozinski and Bach, 2010a)

- A covariate can only be selected after its ancestors
- Structure on parameters w
- Hierarchical penalization: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|_2$ where groups g in $\mathcal{G}$ are equal to the set of descendants of some nodes in a tree.

A new approach based on combinatorial functions

General framework

Let $V = \{1, \dots, d\}$.
Given a set function $F : 2^V \mapsto \mathbb{R}_+$

General framework

Let $V = \{1, \ldots, d\}$.
Given a set function $F : 2^V \mapsto \mathbb{R}_+$ consider

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

General framework

Let $V = \{1, \ldots, d\}$.
Given a set function $F : 2^V \mapsto \mathbb{R}_+$ consider

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

Examples of combinatorial functions

- Use **recursivity** or **counts** of structures (e.g. tree) with DP
- **Block-coding** (Huang et al., 2011)

$$\tilde{G}(A) = \min_{B_i} F(B_1) + \ldots + F(B_k) \quad \text{s.t.} \quad B_1 \cup \ldots \cup B_k \supset A$$

- **Submodular functions** (Work on convex relaxations by Bach (2010))

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

$\rightarrow$ Greedy algorithms

$\rightarrow$ Non-convex methods

$\rightarrow$ Relaxation

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

$\rightarrow$ Greedy algorithms

$\rightarrow$ Non-convex methods

$\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$	

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$ $\downarrow$ $L(w) + \lambda \lVert w \rVert_1$	

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\|A\|$	$F(A)$
$L(w) + \lambda \|\mathrm{Supp}(w)\|$ $\downarrow$ $L(w) + \lambda \\|w\\|_1$	$L(w) + \lambda F(\mathrm{Supp}(w))$

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

$\rightarrow$ Greedy algorithms

$\rightarrow$ Non-convex methods

$\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$	$L(w) + \lambda F(\mathrm{Supp}(w))$
$\downarrow$	$\downarrow$?
$L(w) + \lambda \lVert w \rVert_1$	$L(w) + \lambda$ ...?...

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?

Previous relaxation result

Bach (2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization
- Required for F allowing large supports such as $A \mapsto 1_{\{A \neq \varnothing\}}$

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization
- Required for F allowing large supports such as $A \mapsto 1_{\{A \neq \varnothing\}}$
- Leads to a penalty which is *coercive* so that a convex relaxation on $\mathbb{R}^d$ will not be trivial.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\to$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda>0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

Leads us to consider:

$$\mathrm{pen}_h(w) \quad = \quad \inf_{\lambda>0} \frac{1}{\lambda}\big(\mu\, F(\mathrm{Supp}(\lambda w)) + \nu\, \|\lambda w\|_p^p\,\big)$$

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\to$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda>0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

Leads us to consider:

$$\begin{aligned} \mathrm{pen}_h(w) &= \inf_{\lambda>0} \frac{1}{\lambda}\big(\mu\, F(\mathrm{Supp}(\lambda w)) + \nu\, \|\lambda w\|_p^p\big) \\ &\propto \Theta(w) := \|w\|_p\, F(\mathrm{Supp}(w))^{1/q} \qquad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\quad \mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

$$\text{with unit ball:} \quad \mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

$$\text{with unit ball:} \quad \mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p \, F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p \, F(\mathrm{Supp}(w))^{1/q}$:

$$\Theta^*(s) \quad = \quad \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p \, F(\mathrm{Supp}(w))^{1/q}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

$$\text{with unit ball:} \quad \mathcal{B}_{\Omega_p^*} := \left\{ s\in\mathbb{R}^d \mid \forall A\subset V,\ \|s_A\|_q^q \leq F(A) \right\}$$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$:

$$\begin{aligned}
\Theta^*(s) &= \max_{w\in\mathbb{R}^d} w^\top s - \|w\|_p\, F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A\subset V} \max_{w_A\in\mathbb{R}^A} w_A^\top s_A - \|w_A\|_p\, F(A)^{1/q}
\end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

$$\text{with unit ball:} \quad \mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

$$\begin{aligned}
\Theta^*(s) &= \max_{w\in\mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A\subset V} \max_{w_A\in\mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q} \\
&= \max_{A\subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}}
\end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

$$\text{with unit ball:} \quad \mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$$

Proposition

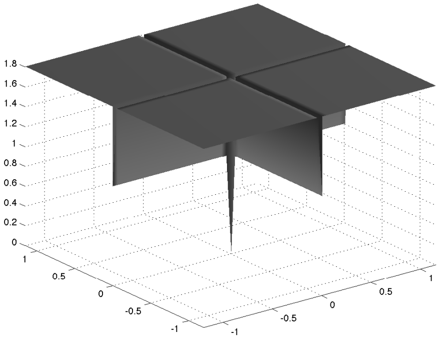
The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

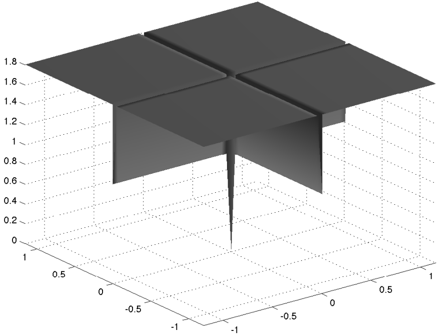
$$\begin{aligned}
\Theta^*(s) &= \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A \subset V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q} \\
&= \max_{A \subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}} = \iota_{\{\Omega_p^*(s) \leqslant 1\}}
\end{aligned}$$

Graphs of the different penalties for $w \in \mathbb{R}^2$

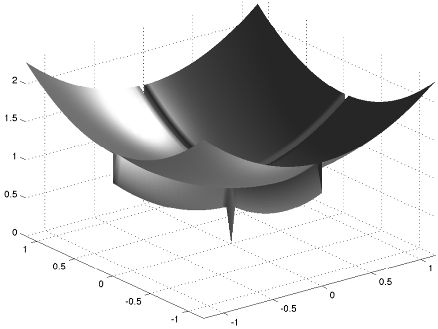


$F(\mathrm{Supp}(w))$

Graphs of the different penalties for $w \in \mathbb{R}^2$

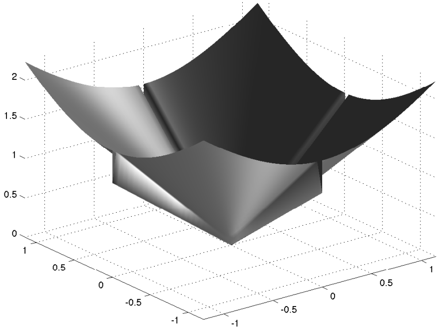


$F(\mathrm{Supp}(w))$



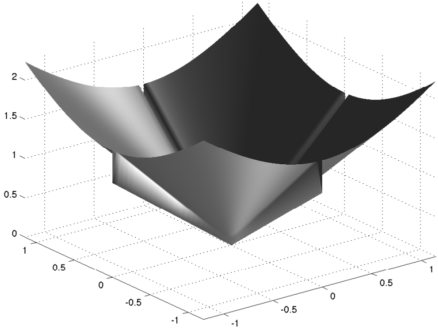
$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_2^2$

Graphs of the different penalties for $w \in \mathbb{R}^2$

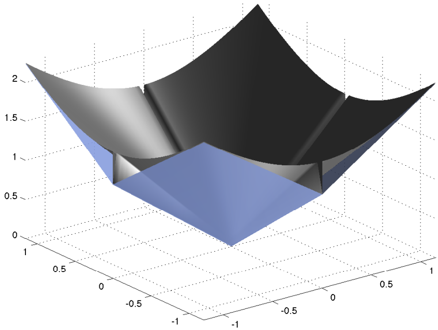


$$\Theta(w) = \sqrt{F(\mathrm{Supp}(w))}\|w\|_2$$

Graphs of the different penalties for $w \in \mathbb{R}^2$



$\Theta(w) = \sqrt{F(\mathrm{Supp}(w))}\|w\|_2$

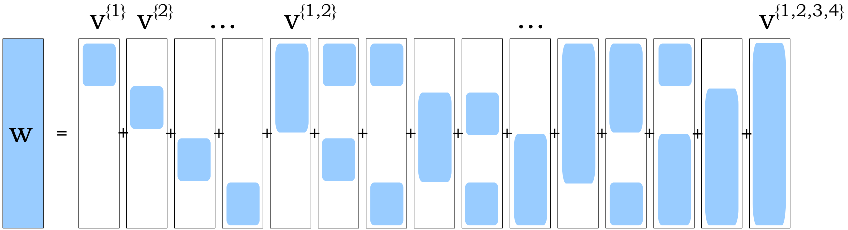


$\Omega^F(w)$

A large latent group Lasso (Jacob et al., 2009)

$$\mathcal{V} = \{v = (v^A)_{A \subset V} \in \left(\mathbb{R}^V\right)^{2^V} \text{ s.t. } \mathrm{Supp}(v^A) \subset A\}$$

$$\Omega_p(w) \quad = \quad \min_{v \in \mathcal{V}} \sum_{A \subset V} F(A)^{\frac{1}{q}} \|v^A\|_p \quad \text{s.t.} \quad w = \sum_{A \subset V} v^A,$$



Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\dots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\ldots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

- When $p=\infty$ and F is submodular, our relaxation coincides with that of Bach (2010).

Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\dots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

- When $p=\infty$ and F is submodular, our relaxation coincides with that of Bach (2010).
- However, when $\mathcal{G}$ is not a partition and $p<\infty$, Ω_p is not in general an ℓ_1/ℓ_p-norms !

$\rightarrow$ New norms... e.g. the k-support norm of Argyriou et al. (2012).

Example

Consider V=$\{1,2,3\}$.

$\mathcal{G} = \{\{1,2\},\{1,3\},\{2,3\}\}$

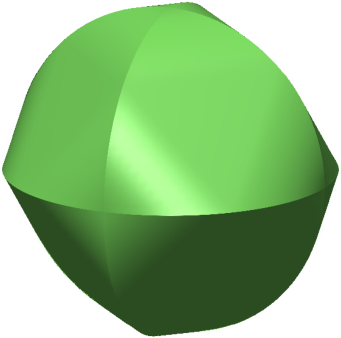
- $F(\{1,2\}) = 1$,
- $F(\{1,3\}) = 1$,
- $F(\{2,3\}) = 1$,
- $F(A) = \infty$ or defined by block-coding.

Example

Consider V=$\{1,2,3\}$.

$\mathcal{G} = \{\{1,2\},\{1,3\},\{2,3\}\}$

- $F(\{1,2\}) = 1$,
- $F(\{1,3\}) = 1$,
- $F(\{2,3\}) = 1$,
- $F(A) = \infty$ or defined by block-coding.



How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

The relaxation fails

How tight is the relaxation? Example: the range function

Consider $V = \{1, \dots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow \Omega_p^F(w) = \|w\|_1$

The relaxation fails

- New concept of Lower Combinatorial envelope provides a tool to analyze the tightness of the relaxation.

Submodular penalties

A function $F : 2^V \mapsto \mathbb{R}$ is *submodular* if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{1}$$

For these functions $\Omega_\infty^F(w) = f(|w|)$ for f the Lovász extension of F.

Properties of submodular function

- f is computed efficiently (via the so-called "greedy" algorithm)
- decomposition ("weak" separability) properties
- F and f can be minimized in polynomial time.

Submodular penalties

A function $F : 2^V \mapsto \mathbb{R}$ is *submodular* if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{1}$$

For these functions $\Omega_\infty^F(w) = f(|w|)$ for f the Lovász extension of F.

Properties of submodular function

- f is computed efficiently (via the so-called "greedy" algorithm)
- decomposition ("weak" separability) properties
- F and f can be minimized in polynomial time.

... leads to properties of the corresponding submodular norms

- Regularized empirical risk minimization problems solved efficiently
- Statistical guarantees in terms of consistency and support recovery.

Consistency for the Lasso (Bickel et al., 2009)

- Assume that $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, Id_n)$
- Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$.
- Denote $J = \mathrm{Supp}(w^*)$.
- Assume the ℓ_1-Restricted Eigenvalue condition:

$$\forall \Delta \ \text{ s.t. } \ \|\Delta_{J^c}\|_1 \leqslant 3\,\|\Delta_J\|_1, \quad \Delta^\top Q \Delta \geqslant \kappa\,\|\Delta_J\|_1^2.$$

Then we have

$$\frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \frac{72|J|\sigma^2}{\kappa}\Big(\frac{2\log p + t^2}{n}\Big),$$

with probability larger than $1 - \exp(-t^2)$.

Support Recovery for the Lasso (Wainwright, 2009)

- Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, Id_n)$
- Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d \times d}$.
- Denote by $J = \mathrm{Supp}(w^*)$.
- Define $\nu = \min_{j, w_j^* \neq 0} |w_j^*| > 0$
- Assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$
- Assume the Irrepresentability Condition, i.e., that for $\eta > 0$,

$$\|\!| Q_{JJ}^{-1} Q_{JJ^c} \|\!|_{\infty,\infty} \leqslant 1 - \eta.$$

Then, if $\frac{2}{\eta}\sqrt{\frac{2\sigma^2 \log(p)}{n}} < \lambda < \frac{\kappa\nu}{|J|}$, the minimizer $\hat{w}$ is unique and has support equal to J, with probability larger than $1 - 4\exp(-c_1 n \lambda^2)$.

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:

$F(A) = \mathrm{range}(A) = i_{\max}(A) - i_{\min}(A) + 1.$

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:
$F(A) = \mathrm{range}(A) = i_{\mathsf{max}}(A) - i_{\mathsf{min}}(A) + 1.$

$\rightarrow$ F is not submodular...

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:

$F(A) = \mathrm{range}(A) = i_{\max}(A) - i_{\min}(A) + 1.$

$\rightarrow$ F is not submodular...

$\rightarrow$ $G(A) = |A|$

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:

$F(A) = \mathrm{range}(A) = i_{\max}(A) - i_{\min}(A) + 1.$

$\rightarrow$ F is not submodular...

$\rightarrow$ $G(A) = |A|$

But $F(A) := d - 1 + range(A)$ is submodular !

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:

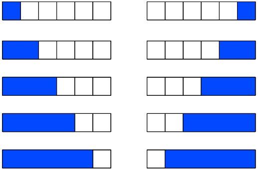
$F(A) = \mathrm{range}(A) = i_{\mathrm{max}}(A) - i_{\mathrm{min}}(A) + 1.$

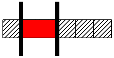
$\rightarrow$ F is not submodular...

$\rightarrow$ $G(A) = |A|$

But $F(A) := d - 1 + range(A)$ is submodular !

In fact $F(A) = \sum_{B \in \mathcal{G}} 1_{\{A \cap B \neq \varnothing\}}$ for B of the form:





Jenatton et al. (2011) considered $\Omega(w) = \sum_{B \in \mathcal{B}} \|w_B \circ d_B\|_2$.

Experiments

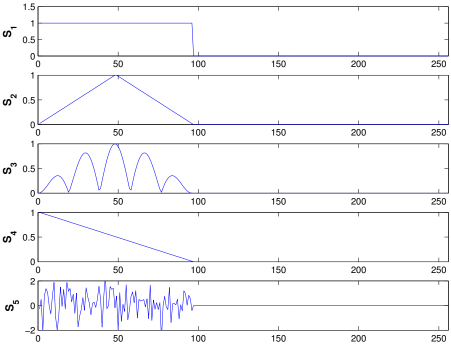


Figure: Signals

S_1 constant

S_2 triangular shape

S_3 $x \mapsto |\sin(x)\sin(5x)|$

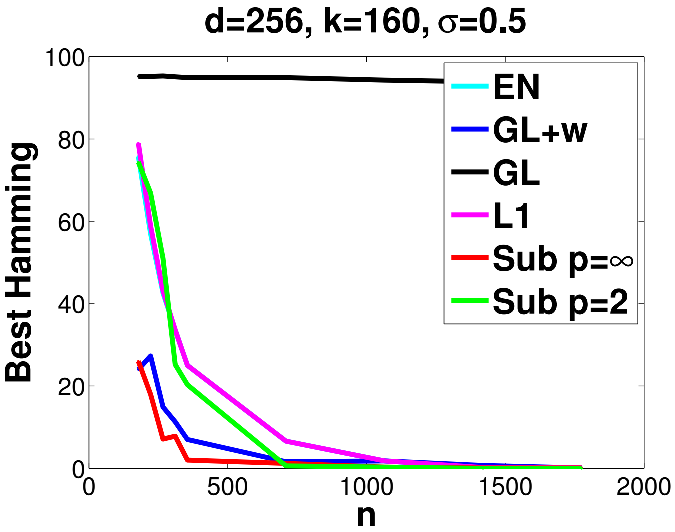
S_4 a slope pattern

S_5 i.i.d. Gaussian pattern

Compare:

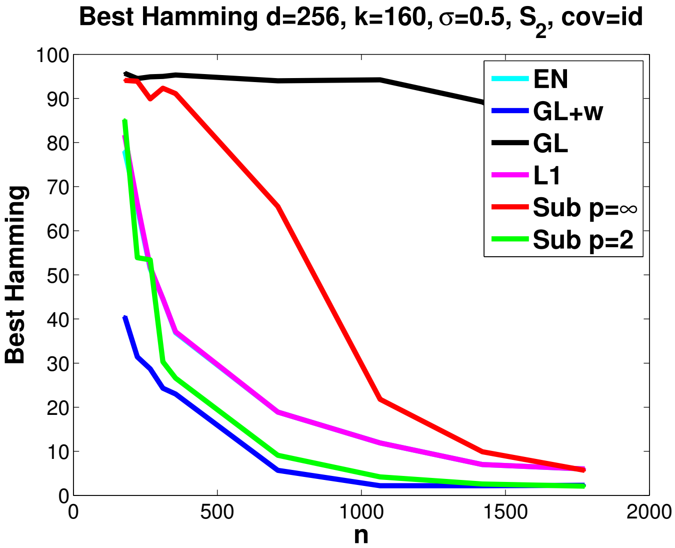
- Lasso
- Elastic Net
- Naive ℓ_2 group-Lasso
- Ω_2 for $F(A) = d - 1 + \mathrm{range}(A)$
- Ω_∞ for $F(A) = d - 1 + \mathrm{range}(A)$
- The weighted ℓ_2 group-Lasso of (Jenatton et al., 2011).

Constant signal



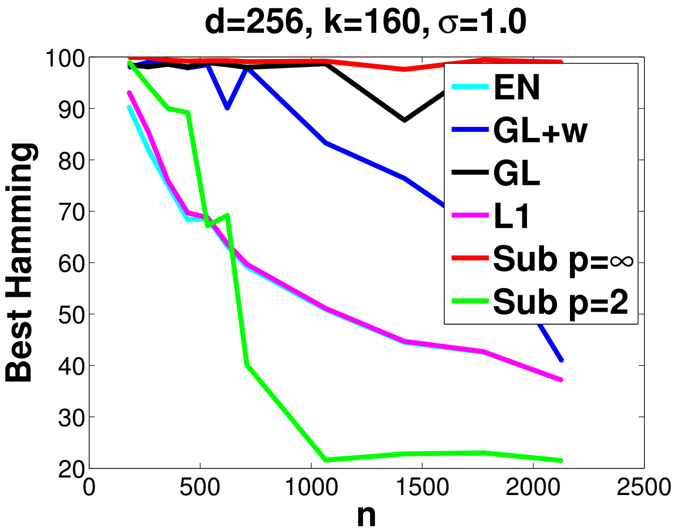
- $d = 256$
- $k = 160$
- $\sigma = .5$

Triangular signal



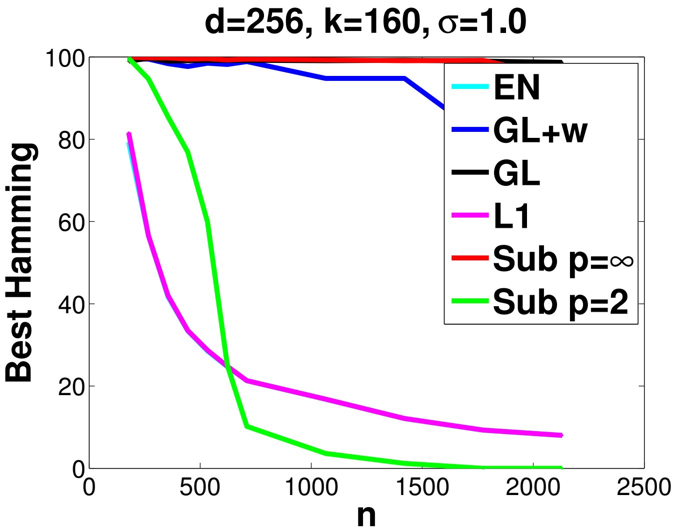
- $d = 256$
- $k = 160$
- $\sigma = .5$

$(x_1, x_2) \mapsto |\sin(x_1)\sin(5x_1)\sin(x_2)\sin(5x_2)|$ signal in 2D



- $d = 256$
- $k = 160$
- $\sigma = .5$

i.i.d Random signal in 2D



- $d = 256$
- $k = 160$
- $\sigma = .5$

Summary

- A convex relaxation for functions penalizing
  (a) the support via a general set function
  (b) the ℓ_p norm of the parameter vector w.
- Principled construction of:
  - known norms like the group Lasso or ℓ_1/ℓ_p-norm
  - many new sparsity inducing norms
- Caveat: the relaxation can fail to capture the structure (e.g. range function)
- For submodular functions we can obtain efficient algorithms, and theoretical results such as consistency and support recovery guarantees.

References I

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474.

Bach, F. (2008). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*.

Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. In *Adv. NIPS*.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *J. Mach. Learn. Res.*, 12:3371–3412.

Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In *ICML*.

Jenatton, R., Audibert, J., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824.

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., and Thirion, B. (2012). Multiscale mining of fMRI data with hierarchical structured sparsity. *SIAM Journal on Imaging Sciences*, 5(3):835–856.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2011). Convex and network flow optimization for structured sparsity. *JMLR*, 12:2681–2720.

References II

Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1- constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, 68:49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.