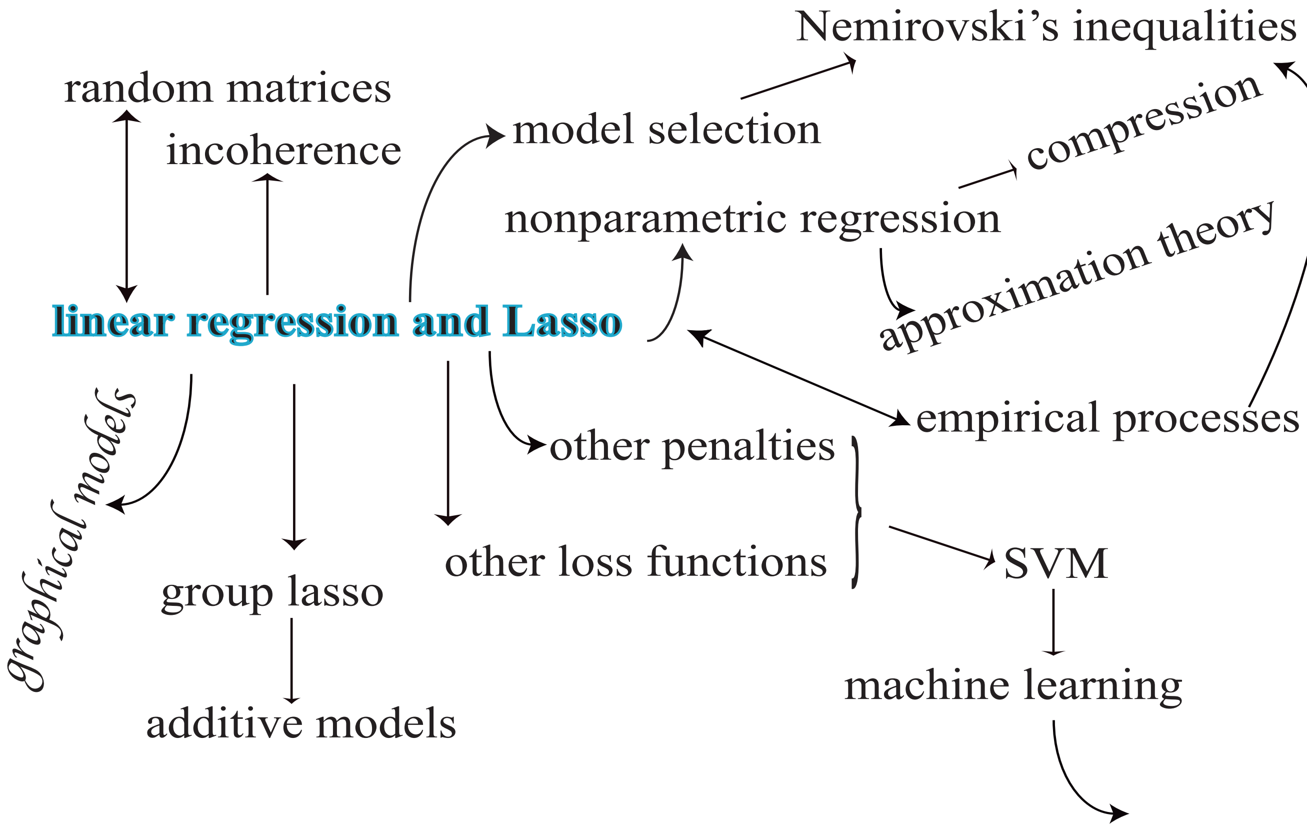
M-estimation

and Complexity Regularization

Sara van de Geer
Seminar für Statistik, ETH Zürich

Toulouse, June 2008

High-dimensional road map



Linear regression and the Lasso

Observations $\{X_i, Y_i\}_{i=1}^n$:
co-variables $X_i \in \mathbb{R}^p$, response variables $Y_i \in \mathbb{R}$.
Linear model:

$$Y_i = \beta_1 X_{i,1} + \ldots + \beta_{i,p} X_{i,p} + \epsilon_i, \ \imath = 1, \ldots, n,$$

with $\beta_1, \ldots, \beta_p$ unknown parameters.
High-dimensional data: $p \gg n$!
Least squares with Lasso penalty:

$$\hat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Oracle result

Let $f^0 := \arg\min_{\text{all } f} \mathbb{E}\|Y - f\|_2^2$.

For appropriate choice of λ, of order $\sqrt{\log p / n}$:

$$\mathbb{E}\|X\hat{\beta} - f^0\|_2^2$$

$$\leq (1+\delta)\{\min_{\beta} \|X\beta - f^0\|^2 + \lambda^2 \#\{\beta_j \neq 0\}\}.$$

(see Bühlmann and Meinshausen (2006), Candes and Tao (2007), vdG (2007), ...)

Extensions

To other loss functions (vdG (2008)),
e.g., support vector machine loss
(Tarigan and vdG (2006))

Technical tools

Contraction and concentration inequalities,
the behavior of suprema of stochastic processes indexed by functions.

Cross road to model selection

$$\hat{f} := \arg \min_{j=1,\ldots,p} \|Y - f_j\|_2^2.$$

Aim is to show that $\|\hat{f} - f_0\|_2^2$ is close to

$$\|f^* - f^0\|_2^2 := \min_{j=1,\ldots,p} \|f_j - f^0\|_2^2$$

(recall $f^0 := \arg\min_{\text{all } f} \mathbb{E}\|Y - f\|_2^2$).
Recent work concerns the case where the errors $\epsilon := Y - f_0$ have only lower order moments, e.g., oracle results of the form

$$\sqrt{\mathbb{E}\frac{\|\hat{f} - f^0\|_2^2}{\|f^* - f^0\|_2^2}} \le 1 + \mathrm{rest},$$

with

$$\mathrm{rest} = C\sqrt{\frac{\lambda}{\|f_* - f^0\|_2^2}} + c\left(\frac{K}{\|f^* - f^0\|_2^2}\right)^{\frac{s}{s+1}} \lambda^{\frac{s-1}{s+1}},$$

where

$$\lambda := \frac{2\log(2p)}{n}, K := E|\epsilon|^s$$

(Mitchell and vdG (2008)).

Junction to Nemirovski inequalities

Lemma (Dümbgen, vdG, Wellner (2008)) *Let $X_1, \ldots, X_n$ be independent centered random variables in $\mathbf{R}^p$ and set $S_n = \sum_{i=1}^n X_i$. Then*

$$\sqrt{\mathbb{E}\|S_n\|_\infty} \le (1 + 3.46)\sqrt{\log(2p)}\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|_\infty^2}.$$

Cross road to additive models with many components

$$Y_i = f_1(X_{i,1}) + \ldots + f_p(X_{i,p}) + \epsilon_i, \ i = 1, \ldots, n,$$

with f_j unknown functions satisfying a smoothness assumption, e.g.,

$$I^2(f_j) := \int |f_j^{(s)}(x)|^2 dx < \infty.$$

Estimator of group Lasso type:

$$\hat{f} = \arg\min \left\{ \|Y - \sum_{j=1}^{p} f_j\|_2^2 + \mathrm{pen}(f) \right\},$$

with

$$\mathrm{pen}(f) := \sum_{j=1}^{p} \mathrm{pen}(f_j)$$

$$\mathrm{pen}(f_j) := \lambda \sqrt{\|f_j\|_2^2 + \lambda^2 I^2(f_j)} + \lambda^2 I^2(f_j).$$

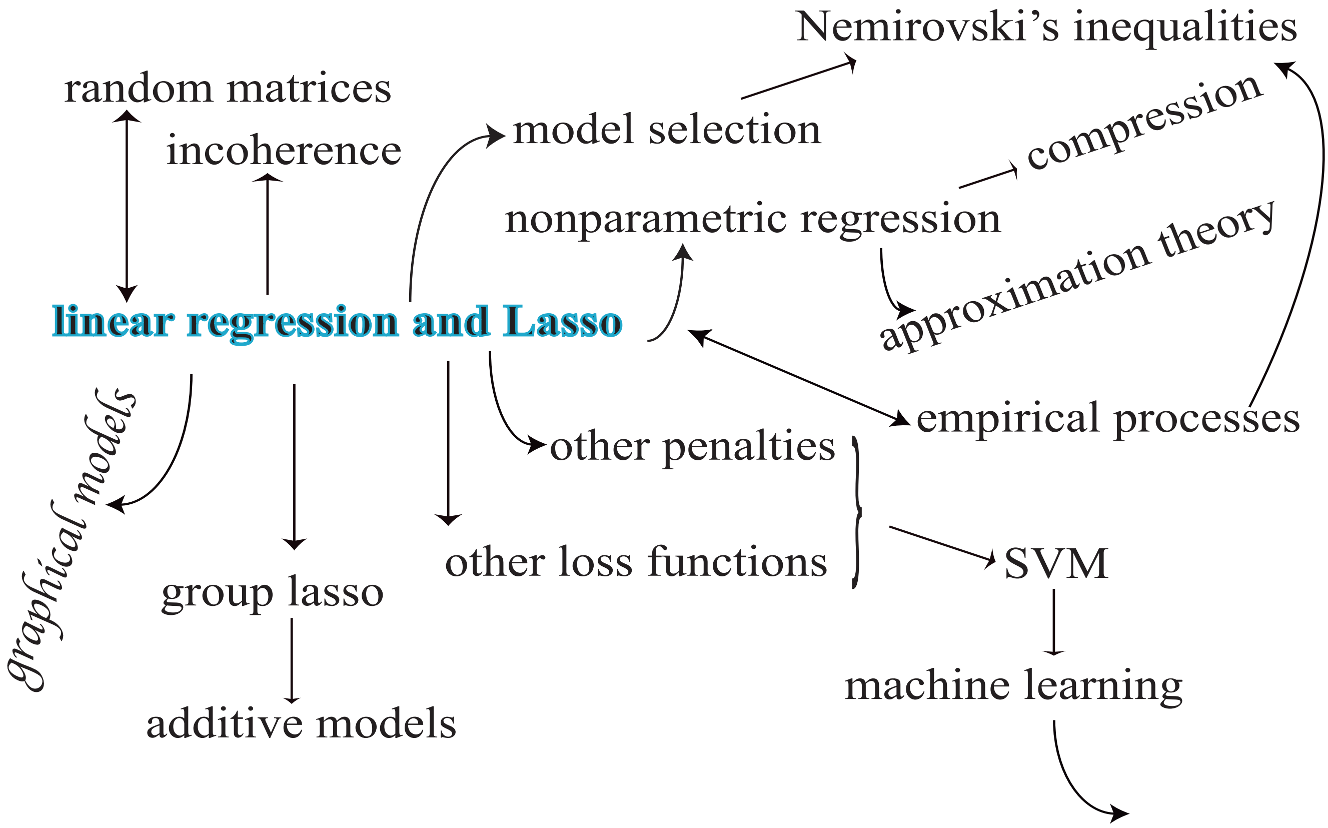
Oracle result

$$\mathbb{E}\|\hat{f} - f^0\|_2^2$$

$$\leq (1+\delta) \min_f \left\{ \|f - f^0\|_2^2 + \lambda^{2-\gamma} \sum_{f_j \neq 0} I^2(f_j) \vee 1 \right\}.$$

(Bühlmann, Meier and vdG (2008)).

High-dimensional road map



Regression model

$$Y_i = f^0(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

with

$Y_i \in \mathbb{R}$,

$x_i \in \mathcal{X}$ (fixed design),

$f^0 : \mathcal{X} \to \mathbb{R}$ an unknown function.

Penalized least squares

We study the estimator

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \mathrm{pen}(f) \right\},$$

where $\mathrm{pen}(f)$ is a penalty,
depending on some measure of complexity $I(f)$,
and on a smoothing parameter λ_n.

Notation

Let

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

be the empirical distribution of the co-variables.

Define

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} f^2(x_i).$$

Let

$$\mathcal{F} \subset L_2(Q_n),$$

be some linear space of functions.

Complexity measure

Let $I : \mathcal{F} \longrightarrow [0, \infty)$ be some map. Think of $I(f)$ measuring the *complexity* of the function f.

Example: smooth functions.

$$\mathcal{X} := [0,1],$$

$$I(f) := \left(\int |f^{(s)}(x)|^q dx \right)^{\frac{1}{q}}.$$

Here, $1 \leq q \leq \infty$.

Example: linear functions.

$$\mathcal{F} := \{ f_\beta(\cdot) := \sum_{j=1}^p \beta_j \psi_j(\cdot) : \ \beta \in \mathbf{R}^p \},$$

with $\{\psi_j\}$ a given dictionary of functions on $\mathcal{X}$.
Moreover, possibly $p \gg n$.
Take the ℓ_γ complexity measure

$$I^\gamma(f) := \sum_{j=1}^p |\beta_j|^\gamma := \|\beta\|_\gamma^\gamma.$$

Special cases:

- $\gamma = 1$: $I(f_\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, the LASSO.
- $\gamma = 0$: BIC

  $I^0(f_\beta) = \|\beta\|_0^0 = \mathrm{card}\{j : \ \beta_j \neq 0\} := N_\beta$.

Remark

- •• $\gamma \to 0$:

  $I^{0+}(f_\beta) := \sum_{j=1}^p \log\left(1 + \frac{|\beta_j|}{\lambda_n}\right)$?

Let β^* be arbitrary (later it will be the oracle), but satisfying the compatibility condition below. Let $f^* := f_{\beta^*}$ and let

$$\mathcal{A}_* := \{j : \beta_j^* \neq 0\}$$

be the *active* set, with cardinality

$$N_* := \mathrm{card}(\mathcal{A}_*).$$

Define $\beta_{\text{in}} = \beta \text{l}\{j \in \mathcal{A}_*\}$ and $\beta_{\text{out}} = \beta \text{l}\{j \notin \mathcal{A}^*\}$. Let

$$f = f_\beta := f_{\text{in}} + f_{\text{out}}$$

with

$$f_{\text{in}} := f_{\beta_{\text{in}}} = \sum_{j \in \mathcal{A}_*} \beta_j \psi_j,$$

and

$$f_{\text{out}} := f_{\beta_{\text{out}}} = \sum_{j \notin \mathcal{A}_*} \beta_j \psi_j.$$

Compatibility assumption: For all β, we have the eigenvalue assumption

$$\|\beta_{\text{in}}\|_2 \leq \|f_{\text{in}}\|_n / \psi_*,$$

and the canonical correlation assumption

$$\frac{|(f_{\text{in}}, f_{\text{out}})_n|}{\|f_{\text{in}}\|_n \|f_{\text{out}}\|_n} \leq \rho_* < 1.$$

Relaxed compatibility assumption:
Opposition does not pay off.

For all β with $I(f_{\rm out}) \le 3I(f_{\rm in})$, we have the eigen-value assumption

$$\|\beta_{\rm in}\|_2 \le \|f_{\rm in}\|_n / \psi_*^2,$$

and the opposition assumption

$$\frac{(f_{\rm in}, f_{\rm out})_n}{\|f_{\rm in}\|_n \|f_{\rm out}\|_n} \ge \rho_* > -1.$$

Remark

In the case of ℓ_1 penalization:

The *Relaxed compatibility condition* is related to the *Restricted Eigenvalue* (RE) Property in Bickel, Ritov and Tsybakov (2007).

The *Restricted Isometry Property* (RIP) (Candes and Tao (2007)) is sufficient.

Related: *Mutual Incoherence*, *Uniform Uncertainty Principle* (UUP), *Irrepresentability Condition*.

(vdG (2007) calls it the *Compatibility Condition*, or simply *Condition C*.)

Conjugate

Let $0 \leq \gamma \leq 1$. The conjugate of γ is defined as

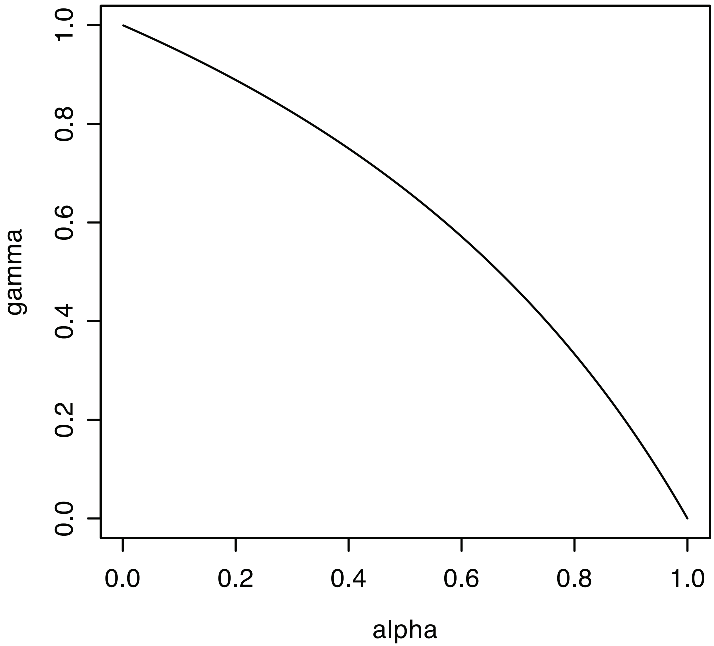
$$\alpha := g(\gamma),$$

where

$$g(\gamma) = \frac{2(1-\gamma)}{2-\gamma}.$$

Note that

$$g = g^{-1}.$$



1.0
0.8
0.6
0.4
0.2
0.0
gamma
0.0
0.2
0.4
0.6
0.8
1.0
alpha

The empirical process

Define

$$(\epsilon, f)_n := \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i).$$

Let $\alpha = g(\gamma)$ be the conjugate of γ.

We will assume the *Empirical Process Condition*: with large probability

$$\sup_{f \in \mathcal{F}} \frac{|(\epsilon, f)_n|}{\|f\|_n^{\alpha} I^{1-\alpha}(f)} \leq \lambda_n.$$

Generally

$$\lambda_n \sim \frac{1}{\sqrt{n}} \times \text{ possible log factors.}$$

Example: smooth functions.

$$I(f) := \left(\int |f^{(s)}(x)|^q dx \right)^{\frac{1}{q}} .$$

Then

$$\alpha = 1 - \frac{1}{2s},\ \gamma = \frac{2}{2s+1},$$

and

$$\lambda_n \sim \frac{1}{\sqrt{n}}.$$

Example: linear functions.

$$I^{\gamma}(f) := \sum_{j=1}^{p} |\beta_j|^{\gamma}.$$

Then

$$\alpha = \frac{2(1-\gamma)}{2-\gamma} = g(\gamma)$$

and

$$\lambda_n \sim \sqrt{\frac{\log(p)}{n}}.$$

Special cases:

- $\gamma = 1 \Rightarrow \alpha = 0$:

$$|(\epsilon, f_\beta)_n| = |\sum_{j=1}^{p} \beta_j (\epsilon, \psi_j)_n| \le \|\beta\|_1 \max_{1 \le j \le p} |(\epsilon, \psi_j)_n|$$

$$\le \lambda_n \|\beta\|_1.$$

Note: with correlated ψ_j, this can be improved to some $\alpha > 0$ (entropy conditions).

- $\gamma = 0 \Rightarrow \alpha = 1$:

$$|(\epsilon, f_\beta)_n| \le \lambda_n \|f_\beta\|_n \sqrt{\|\beta\|_0^0} = \lambda_n \|f_\beta\|_n \sqrt{N_\beta}.$$

Entropy conditions

Let $H(\cdot, \{f \in \mathcal{F} : I(f) \le 1\}, Q_n)$ be the entropy of $\{f \in \mathcal{F} : I(f) \le 1\}$. Assume that I is scalable and

$$H(\delta, \{f \in \mathcal{F} : I(f) \le 1\}, Q_n) \le A_n \delta^{-2(1-\alpha)}.$$

Then the *Empirical Process Condition* holds: with large probability

$$\sup_{f \in \mathcal{F}} \frac{|(\epsilon, f)_n|}{\|f\|_n^\alpha I^{1-\alpha}(f)} \le \lambda_n,$$

with

$$\lambda_n \sim \sqrt{\frac{A_n}{n}}.$$

Basic inequality

Lemma 1 *We have the basic inequality*

$$\|\hat{f} - f^0\|_n^2 + \text{pen}(\hat{f})$$

$$\leq 2|(\epsilon, \hat{f} - f^*)_n| + \text{pen}(f^*) + \|f^* - f^0\|_n^2.$$

Proof. This is rewriting,

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{f}(x_i)|^2 + \text{pen}(\hat{f})$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} |Y_i - f^*(x_i)|^2 + \text{pen}(f^*).$$

$\square$

Recall the *Empirical Process Condition*: with large probability

$$2|(\epsilon, f)_n| \leq \ 2\lambda_n \|f\|_n^{\alpha} I^{1-\alpha}(f).$$

The penalty should be such that it kills the empirical process.
Now, use that for positive a and b,

$$a^{\alpha} b^{1-\alpha} \leq a^2 + b^{\gamma}.$$

This implies

$$2\lambda_n a^\alpha b^{1-\alpha} \le a^2 + (2\lambda_n)^{2-\gamma} b^\gamma .$$

Theorem *Take*

$$\mathrm{pen}(f) := 2 \times (2\lambda_n)^{2-\gamma} I^\gamma(f),$$

where $\gamma = g(\alpha)$ *is the conjugate of* α*. Then on the set*

$$\mathcal{S} := \left\{ \sup_{f \in \mathcal{F}} \frac{|(\epsilon, f)_n|}{\|f\|_n^\alpha I^{1-\alpha}(f)} \le \lambda_n \right\},$$

we have

$$\|\hat{f} - f^0\|_n^2 + \mathrm{pen}(\hat{f}) \le 3\mathrm{pen}(f^*) + \|f^* - f^0\|_n^2 .$$

If pen *is the* ℓ_γ *penalty, then it is sparsity decomposable:*

$$\text{pen}(f) = \text{pen}(f_{\text{in}}) + \text{pen}(f_{\text{out}}),\ \text{pen}(f^*_{\text{out}}) = 0,$$

and sub-linear:

$$\text{pen}(f + \tilde{f}) \leq \text{pen}(f) + \text{pen}(\tilde{f}).$$

If the the relaxed compatibility condition holds, then on $\mathcal{S}$*, for* $\phi_*^2 := \psi_*^2(1-\rho_*^2)$*,*

$$\|\hat{f} - f^0\|_n^2 + \text{pen}(\hat{f} - f^*) \leq 16\frac{N_*\lambda_n^2}{\phi_*^2} + 3\|f^* - f^0\|_n^2.$$

Example: smooth functions.

$$I^2(f) := \int_0^1 |f^{(s)}(x)|^2 dx.$$

Then

$$\gamma = \frac{2}{2s+1}, \; \alpha = 1 - \frac{1}{2s}.$$

and

$$\lambda_n^{2-\gamma} \sim n^{-\frac{2-\gamma}{2}} = n^{-\frac{2s}{2s+1}}.$$

So we take

$$\mathrm{pen}(f) \sim n^{-\frac{2s}{2s+1}} \left(\int |f^{(s)}(x)|^2 dx \right)^{\frac{1}{2s+1}} .$$

We find

$$\|\hat{f} - f^*\|_n^2 + \mathrm{pen}(\hat{f}) \leq 3\mathrm{pen}(f^*).$$

Standard penalty:

$$\mathrm{standardpen}(f) := \lambda^2 \int |f^{(s)}(x)|^2 dx.$$

By data depend choice of λ

$$\mathrm{standardpen}(f) = \mathrm{pen}(f).$$

Example: linear functions

$$I^{\gamma}(f_{\beta}) := \|\beta\|_{\gamma}^{\gamma}.$$

Then $\alpha = g(\gamma)$, and

$$\lambda_n \sim \sqrt{\log(p)/n}.$$

So we take

$$\mathrm{pen}(f_{\beta}) \sim \left(\frac{\log(p)}{n}\right)^{\frac{2-\gamma}{2}} \|\beta\|_{\gamma}^{\gamma}.$$

Special cases:

- $\gamma = 1$: $\text{pen}(f_\beta) = \|\beta\|_1 \sqrt{\log(p)/n}$.
- $\gamma = 0$: $\text{pen}(f_\beta) = \|\beta\|_0^0 \log(p)/n$.

For general γ, under the compatibility condition, we get, taking $f^* = f^0$ (or the projection of f^0 onto the space of linear functions $\{f_\beta : \ \beta \in \mathbf{R}^p\}$),

$$\|\hat{f} - f^0\|_n^2 \sim \frac{\log(p)}{n} \frac{N_*}{\phi_*^2}$$

$$\|\hat{\beta} - \beta^0\|_\gamma^\gamma \sim \left(\frac{\log(p)}{n} \right)^{\frac{\gamma}{2}} \frac{N_*}{\phi_*^2}.$$

- $\gamma = 1$: $\|\hat{\beta} - \beta^0\|_1 \sim \sqrt{\frac{\log(p)}{n}} \frac{N_*}{\phi_*^2}$.
- $\gamma = 0$: $\|\hat{\beta} - \beta^0\|_0^0 \sim N_*$.

Numerics. With $\gamma = 0$, the estimator is computationally infeasible. For $0 < \gamma < 1$, one has

$$\frac{d}{d\beta}|\beta|^{\gamma} = \gamma \frac{1}{|\beta|^{1-\gamma}}.$$

Adaptive Lasso:

$$\min \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\beta}(x_i)|^2 + \tilde{\lambda}_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\beta_j^{\text{init}}|^{1-\gamma}} \right\}.$$

Additive model

Let $\mathcal{X} = [0,1]^p$ with p large, and

$$\mathcal{F} := \{f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j) : \; f_j \in \mathcal{F}\}.$$

Let

$$I^q(f_j) := \int |f_j^{(s)}(z)|^q dz, \; f_j \in \mathcal{F}_0$$

Define the *active* set

$$\mathcal{A}_* = \{j : \|f_j^*\|_n \neq 0\},$$

and let $N_* = \mathrm{card}(\mathcal{A}_*)$.

Empirical process

Let

$$\mathcal{S} := \{ |(\epsilon, f)_n| \le \sum_{j=1}^{p} |(\epsilon, f_j)_n| \le \lambda_n \sum_{j=1}^{p} \|f_j\|_n^{\alpha} I^{1-\alpha}(f$$

where $\alpha = g(\gamma)$, and $\gamma := 2/(2s+1)$. Moreover, let

$$\lambda_n \sim \sqrt{\log(p)/n}.$$

The set $\mathcal{S}$ has large probability (under certain conditions).

To come up with an appropriate penalty, we again use that

$$a^{\alpha} b^{1-\alpha} \leq a^2 + b^{\gamma}.$$

This leads to the penalty

$$\mathrm{pen}(f) \sim \lambda_n^{2-\gamma} \sum_{j=1}^{p} I(f_j)^{\gamma}.$$

Lemma 2 *Suppose the strong compatibility condition*

$$\sum_{j=1}^{p} \|f_j\|_n^2 \le \|\sum_{j=1}^{p} f_j\|_n^2 / \phi_*^2.$$

Then on $\mathcal{S}$,

$$\|\hat{f} - f^{)}\|_n^2 + \mathrm{pen}(\hat{f}) \le 3\{\mathrm{pen}(f^*) + \|f^* - f^0\|_n^2\}.$$

We note that

$$\text{pen}(f^*) \sim \lambda_n^{2-\gamma} \sum_{j \in \mathcal{A}_*} I^\gamma(f_j^*)$$

$$\sim n^{-\frac{2s}{2s+1}} \sum_{j \in \mathcal{A}_*} I^\gamma(f_j^*) \leq n^{-\frac{2s}{2s+1}} N_* / \phi_*^{2-\gamma},$$

assuming that $I(f_j^*) \leq 1$ for all j.
So we have an oracle inequality.

Numerics.

The estimator is computationally intractable. For example, with $q = 2$, we have

$$\mathrm{pen}(f) = \lambda_n^{\frac{4s}{2s+1}} \sum_{j=1}^{p} \left(\int |f_j^{(s)}(z)|^2 dz \right)^{\frac{1}{2s+1}}.$$

Alternatively, we may use that

$$a^{\alpha}b^{1-\alpha} \le a + b.$$

Hence

$$\lambda a^{\alpha}b^{1-\alpha} \le \lambda^{\frac{2-\gamma}{2}} a + \lambda^{2-\gamma} b.$$

This leads to the penalty

$$\mathrm{pen}(f) \sim \lambda_n^{\frac{2-\gamma}{2}} \sum_{j=1}^{p} \|f_j\|_n + \lambda_n^{2-\gamma} \sum_{j=1}^{p} I(f_j).$$

Compatibility condition: For all $f = \sum_{j=1}^{p} f_j$,

$$\sum_{j \in \mathcal{A}_*} \|f_j\|_n^2 \le \|\sum_{j=1}^{p} f_j\|_n^2 / \phi_*^2.$$

Lemma 3 *Assume the compatibility condition. Then on $\mathcal{S}$,*

$$\|\hat{f} - f^0\|_n^2 \sim \lambda_n^{2-\gamma} \left(\frac{N_*}{\phi_*} + \sum_{j \in \mathcal{A}_*} I(f_j^*) \right) + \|f^* - f^0\|_n^2,$$

and

$$\sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n \sim \lambda_n^{\frac{2-\gamma}{2}} \left(\frac{N_*}{\phi_*} + \sum_{j \in \mathcal{A}_*} I(f_j^*) \right)$$

$$+\lambda^{-\frac{2-\gamma}{2}} \|f^* - f_0\|_n^2.$$

Numerics. With $q = 2$, the penalty is

$$\mathrm{pen}(f) \sim \lambda_n^{\frac{2-\gamma}{2}} \left(\sum_{j=1}^{p} \|f_j\|_n + \sqrt{\lambda_n^{2-\gamma} \int_0^1 |f_j^{(s)}(z)|^2 dz} \right).$$

This is computationally similar to the group lasso penalty, but the two terms are intertwined.

It would be computationally easier to use

$$\mathrm{pen}(f) \sim$$

$$\lambda_n^{\frac{2-\gamma}{2}} \sum_{j=1}^{p} \sqrt{\|f_j\|_n^2 + \lambda_n^{2-\gamma} \int_0^1 |f_j^{(s)}(z)|^2 dz}.$$

However, so far the theory does not work for that penalty.

Uniting computational feasibility and oracle behavior

Let

$$\mathrm{pen}(f) := \sum_{j=1}^{p} \mathrm{pen}(f_j),$$

with

$$\mathrm{pen}(f_j) := \lambda^{\frac{2-\gamma}{2}} \sqrt{\|f_j\|_n^2 + \lambda^{2-\gamma} I^2(f_j)} + \lambda^{2-\gamma} I^2(f_j).$$

Theorem *Take* $2\sqrt{2}\lambda_n \leq \lambda \leq 1$. *Suppose the compatibility condition is met. Then on the set* $\mathcal{S}$, *it holds that*

$$\|\hat{f} - f^0_{add}\|_n^2 + \lambda^{\frac{2-\gamma}{2}} \sum_{j=1}^{p} \|\hat{f}_j - f_j^*\|_n$$

$$\leq 3\|f^* - f^0_{add}\|_n^2 + 4\lambda^{2-\gamma} \sum_{j \in \mathcal{A}_*} [I^2(f_j^*) + \frac{3}{\phi_*^2}],$$

where f^0_{add} *is the projection of* f^0 *on the space of additive functions.*

Remark One may take $\lambda \sim \sqrt{\log p/n}$. When $I^2(f_j) := \int \left(f_j''(x)dx\right)^2$, this gives $\lambda^{2-\gamma}$ of order $(\log p/n)^{4/5}$, which is up to the log-term the usual rate for estimating a twice differentiable function. If the oracle f^* has bounded smoothness $I(f_j^*)$ for all j, the rate is thus $N_*(\log p/n)^{4/5}$, with being the number of active variables the oracle needs. This is, again up to the log-term, the same rate one would obtain if it was known beforehand which of the p functions are relevant.

Remark Let $\mathcal{A}_0 = \{j : \|f^0_{add,j}\|_n \neq 0\}$ be the active set of f^0_{add}. Assume the compatibility condition holds for $\mathcal{A}_0$, with constant ϕ_0. Suppose also that for $j \in \mathcal{A}_0$, $I(f^0_{add,j}) \leq 1$ (say) . The theorem tells us that on $\mathcal{S}$,

$$\sum_{j=1}^{p} \|\hat{f}_j - f^0_{add,j}\|_n \leq 16\lambda^{\frac{2-\gamma}{2}}|\mathcal{A}_0|/\phi_0^2.$$

Hence, if

$$\|f^0_{add,j}\|_n > 16\lambda^{\frac{2-\gamma}{2}}|\mathcal{A}_0|/\phi_0^2,\ j \in \mathcal{A}_0,$$

we have (on $\mathcal{S}$), that the estimated active set $\{j : \|\hat{f}_j\|_n \neq 0\} \supset \mathcal{A}_0$.

Simulation

We use the penalty

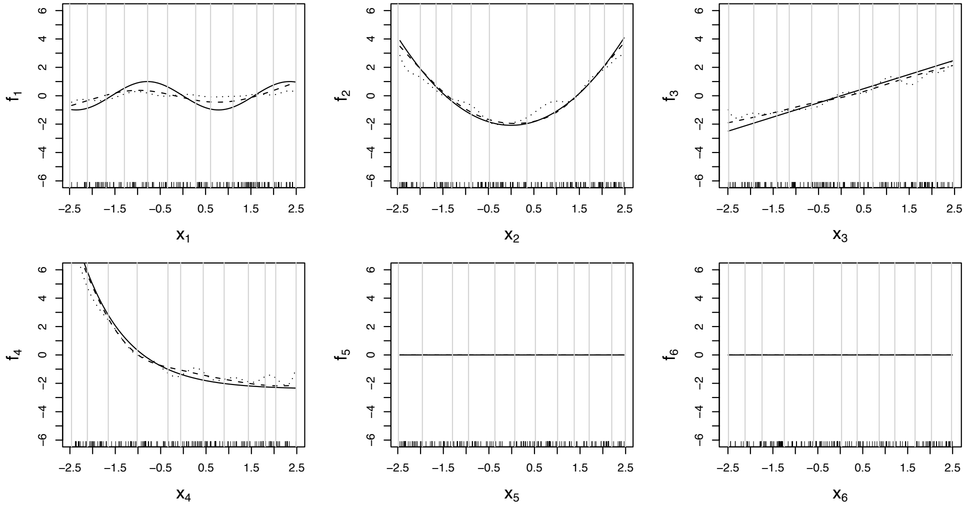
$$\mathrm{pen}(f_j) = \lambda_1 \sqrt{\|f_j\|_n^2 + \lambda_2 I^2(f_j)} + \lambda_3 I^2(f_j).$$

The parameters λ_1 and λ_2 are selected by cross-validation, and either $\lambda_3 := \lambda_2$ or $\lambda_3 := 0$.

For each function f_j we use a cubic B-spline parametrization with a reasonable amount of knots or basis functions. A typical choice would be to use $K - 4 \asymp \sqrt{n}$ interior knots that are placed at the empirical quantiles of x_j. Hence, we parametrize

$$f_j(x) = \sum_{k=1}^{K} \beta_{j,k} b_{j,k}(x),$$

where $b_{j,k} : \mathbb{R} \to \mathbb{R}$ are the B-spline basis functions and $\beta_j \in \mathbb{R}^K$ are the corresponding parameter vectors.



True functions f_j (solid) and estimated functions $\hat{f}_j$ (dashed) for the first 6 components of a simulation run of Example 1. Small vertical bars indicate original data and grey vertical lines knot positions. The dotted lines are the function estimates when no smoothness penalty is used, i.e. when setting $\lambda_2 = 0$.

Example 1 ($n = 150,\ p = 200,\ N_0 = 4,\ \mathrm{SNR} \approx 15$)
This example is similar to example 1 in Wasserman et al.(2008). The model is

$$Y_i = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + \varepsilon_i,$$

with

$$f_1(x) = -\sin(2x),\ f_2(x) = x_2^2 - 25/12,\ f_3(x) = x,$$
$$f_4(x) = e^{-x} - 2/5 \cdot \sinh(5/2).$$

The covariates are simulated from independent Uniform(-2.5, 2.5) distributions.

Example 2 ($n = 100$, $p = 1000$, $N_0 = 4$, $\mathrm{SNR} \approx 6.7$)
As above but high-dimensional and correlated. The covariates are simulated according to a multivariate normal distribution with covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$; $i, j = 1, \ldots, p$.

Example 3 ($n = 100$, $p = 80$, $N_0 = 4$, $\mathrm{SNR} \approx 7.9$
This is similar to example 1 in Zhang (2006) but with more predictors. The model is

$$Y_i = 5f_1(x^{(1)})+3f_2(x^{(2)})+4f_3(x^{(3)})+6f_4(x^{(4)})+\varepsilon_i,$$

$$\varepsilon_i \; N(0, 1.74),$$

with

$$f_1(x) = x,\ f_2(x) = (2x-1)^2,\ f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}$$

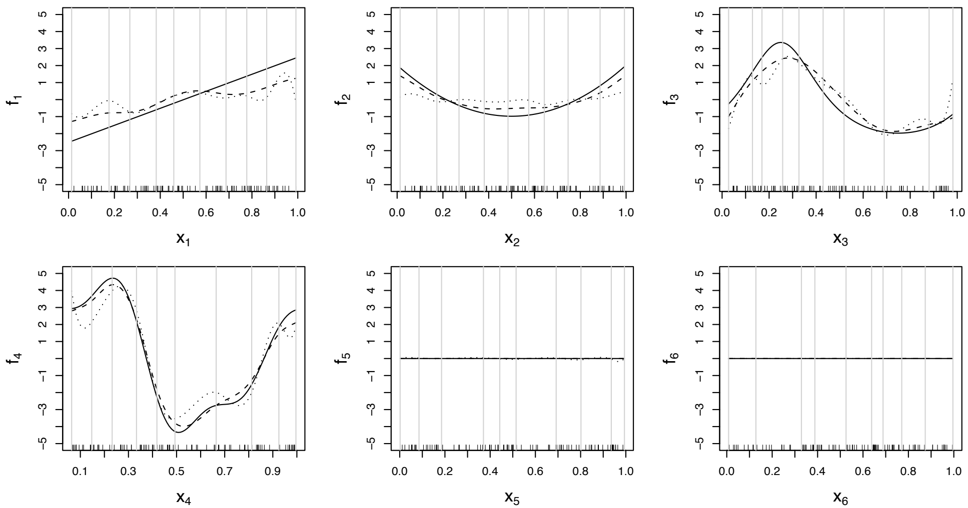
and

$$f_4(x) = 0.1 \sin(2\pi x) + 0.2 \cos(2\pi x) + 0.3 \sin^2(2\pi x) + 0.4 \cos^3(2\pi x) + 0.5 \sin^3(2\pi x).$$

The covariates are simulated according to

$$x^{(j)} = \frac{W^{(j)} + tU}{1 + t},$$

where $W^{(1)}, \ldots, W^{(p)}$ and U are i.i.d. Uniform(0, 1). The case $t = 1$ results in a design with correlation 0.5 between all covariates.



True functions f_j (solid) and estimated functions $\hat{f}_j$ (dashed) for the first 6 components of a simulation run of Example 3. The dotted lines are the function estimates when no smoothness penalty is used, i.e. when setting $\lambda_2 = 0$.

Moreover, we also consider a “high-frequency” situation where we use $f_3(8x)$ and $f_4(4x)$ instead of $f_3(x)$ and $f_4(x)$. The corresponding signal-to-noise ratios for these models are $\mathrm{SNR} \approx 8.1$.

Example 4 ($n = 100$, $p = 60$, $N_0 = 12$, $\mathrm{SNR} \approx 9\,(t = 0)$, ≈ 11.25 **)**
This is similar to example 2 in Zhang (2006) but with fewer observations. We use the same functions as in example 3. The model is

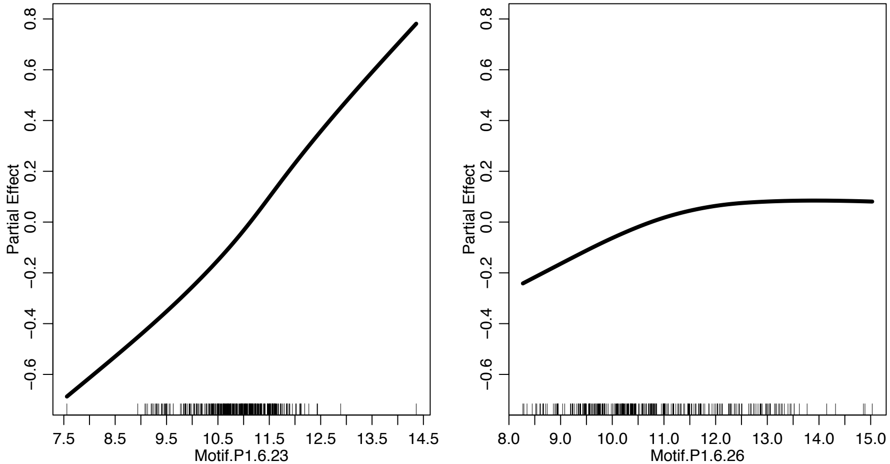
$$\begin{aligned} Y_i = \; & f_1(x^{(1)}) + f_2(x^{(2)}) + f_3(x^{(3)}) + f_4(x^{(4)}) + \\ & 1.5f_1(x^{(5)}) + 1.5f_2(x^{(6)}) + 1.5f_3(x^{(7)}) + 1.5f_4 \\ & 2f_1(x^{(9)}) + 2f_2(x^{(10)}) + 2f_3(x^{(11)}) + 2f_4(x^{(12} \end{aligned}$$

with ε_i i.i.d. $N(0, 0.5184)$. The covariates are simulated as in Example 3.

Model	TP_{SSP}	FP_{SSP}	TP_{boost}	FP_{boost}
Example 1	4.00 (0.00)	24.24 (14.23)	4.00 (0.00)	22.54 (12.91)
Example 2	3.48 (0.61)	34.66 (17.10)	3.60 (0.63)	28.76 (20.15)
Example 3	3.93 (0.29)	19.25 (9.55)	3.92 (0.27)	18.69 (8.38)
Example 3 “high-freq”	2.80 (0.78)	12.26 (7.61)	2.16 (0.94)	9.23 (9.74)
Example 4	10.63 (1.15)	19.49 (7.27)	10.67 (1.25)	23.76 (9.89)

Motif Regression In motif regression problems, the aim is to predict gene expression levels or binding intensities based on information on the DNA sequence. For our specific dataset, from the Ricci lab at ETH Zurich, we have binding intensities Y_i of a certain transcription factor (TF) at 287 regions on the DNA. Moreover, for each region i, motif scores $x_i^{(1)}, \ldots, x_i^{(p)}, p = 196$ are available. We used 5 fold cross-validation to determine the prediction optimal tuning parameters, yielding 28 active functions. To assess the stability of the estimated model, we performed a nonparametric bootstrap analysis. The two functions which appear most often in the bootstrapped model estimates are depicted in the next figure.

Indeed, `Motif.P1.6.26` is the true (known) binding site. A follow-up experiment showed that the TF does not direclty bind to `Motif.P1.6.23`. Hence, this motif is a candidate for a binding site of a co-factor (another TF) and needs further experimental validation.



Estimated functions $\hat{f}_j$ of the two most stable motifs. Small vertical bar indicate original data.

On the compatibility condition

Well-conditioned active set condition *We say that the active set* $\mathcal{A}_*$ *is well conditioned if for some constant* $0 < \psi_* \leq 1$, *and for all* $\{f_j\}_{j \in \mathcal{A}^*}$,

$$\sum_{j \in \mathcal{A}_*} \|f_j\|_n^2 \leq \|\sum_{j \in \mathcal{A}_*} f_j\|_n^2 / \psi_*^2.$$

Writing f_j as linear function of base functions, with coefficients β_j,

$$f_j = B_j \beta_j,$$

with B_j the B-spline matrix of the jth predictor, one sees that ψ_*^2 can be taken as the smallest eigenvalue of the matrix

$$\left((B_j^T B_j)^{-1/2} (B_j^T B_k) (B_k^T B_k)^{-1/2} \right)_{j,k \in \mathcal{A}_*}.$$

The inner product between functions f and $\tilde{f}$ is denoted by $(f, \tilde{f})_n := \sum_{i=1}^n f(x_i)\tilde{f}(x_i)/n$.

No perfect canonical dependence condition *We say that the active and non-active variables have no perfect canonical dependence, if for a constant* $0 \leq \rho_* < 1$, *and all* $\{f_j\}_{j=1}^p$, *we have for* $f_{\rm in} := \sum_{j \in \mathcal{A}_*} f_j$ *and* $f_{\rm out} := \sum_{j \notin \mathcal{A}_*} f_j$, *that*

$$\frac{|(f_{\rm in}, f_{\rm out})_n|}{\|f_{\rm in}\|_n \|f_{\rm out}\|_n} \leq \rho_*.$$

Again, writing $f_j = B_j\beta_j$, one sees that ρ_* can be taken as the canonical correlation between the linear space spanned by $\{B_j\}_{j\in\mathcal{A}_*}$ and the linear space spanned by $\{B_j\}_{j\notin\mathcal{A}_*}$. Note that the condition $\rho_* < 1$ allows for perfect linear dependencies between non-active B_j.

The next Lemma makes the link between the compatibility condition and the above two conditions.

Lemma *Let* $f = f_{\text{in}} + f_{\text{out}}$ *satisfy*

$$\frac{|(f_{\text{in}}, f_{\text{out}})_n|}{\|f_{\text{in}}\|_n \|f_{\text{out}}\|_n} \le \rho_* < 1.$$

Then

$$\|f_{\text{in}}\|_n^2 \le \|f\|_n^2 / (1 - \rho_*^2).$$

Corollary *A well-conditioned active set in combination with no perfect canonical dependence, implies the compatibility condition with* $\phi_*^2 = \psi_*^2(1 - \rho_*^2)$.

References

Bickel, P. Ritov, Ya. and Tsybakov A. (2007). Simultaneous analysis of Lasso and Dantzig selector.

Bunea, F. and Tsybakov, A.B. and Wegkamp, M.H. (2007), Sparsity oracle inequalities for the Lasso, *Electr. Journal of Statist.* **1**

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n, *Ann.Statist.*

B. Tarigan and S.A. van de Geer (2006). Classifiers of support vector machine type, with ℓ_1 penalty. *Bernoulli* **12**, 1045–1076.

van de Geer, S. (2007). High-dimensional generalized linear models and the Lasso.To appear in *Ann. Statist.*

van de Geer, S. (2007). The deterministic Lasso. JSM proceedings, paper nr. 489.