The additive model revisited

Sara van de Geer

January 8, 2013

but first something else

The additive model revisited

Sara van de Geer

January 8, 2013

but first something else

Contents

Sharp oracle inequalities

Let $S \in \mathcal{S}$ be some index set and $\{\mathcal{F}_S\}_{S\in\mathcal{S}}$ be a collection of models. Moreover let $L(X, f)$ be a loss function and $R(f) := \mathbb{E}L(X, f)$. We say that the estimator $\hat{f}$ satisfies a *sharp oracle inequality* if with large probability

$$R(\hat{f}) \leq \min_{S\in\mathcal{S}}\left\{\min_{f\in\mathcal{F}_S} R(f) + \mathrm{Remainder}(S)\right\}.$$

Non-sharp oracle inequalities are of the form: with large probability

$$R(\hat{f}) - R(f^0) \leq (1+\delta)\min_{S\in\mathcal{S}}\left\{\min_{f\in\mathcal{F}_S}(R(f) - R(f^0)) + \mathrm{Remainder}_\delta(S)\right\},$$

where $\delta > 0$ and

$$f^0 := \min_{f\in\cup_{S\in\mathcal{S}}\mathcal{F}_S} R(f).$$

Sharp oracle inequalities with structured sparsity penalities

High-dimensional linear model:

$$Y = X\beta^0 + \epsilon,$$

with $Y \in \mathbb{R}^n$, X and $n \times p$ matrix and $\beta^0 \in \mathbb{R}^p$.
We believe that β^0 can be well approximated by a “structured sparse” β.
Let Ω be some given norm on $\mathbb{R}^p$.

Norm-penalized estimator:

$$\hat{\beta} := \hat{\beta}_\Omega := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 / n + 2\lambda\Omega(\beta) \right\}.$$

Aim:
(Sharp) sparsity oracle inequalities for $\hat{\beta}$.

Notation: for $\beta \in \mathbb{R}^p$ and $S \subset \{1, \ldots, p\}$

$$\beta_{j,S} := \beta_j \mathrm{l}\{j \in S\}.$$

Example

ℓ_1-norm

$$\Omega(\beta) := \|\beta\|_1 := \sum_{j=1}^{p} |\beta_j| \rightsquigarrow \text{ Lasso}$$

The ℓ_1-norm is *decomposable*:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^c}\|_1 \forall \beta \forall S.$$

Definition

We say that the norm Ω is weakly decomposable for S if there exists a norm Ω_{S^c} on $\mathbb{R}^{p-|S|}$ such that for all $\beta \in \mathbb{R}^p$,

$$\Omega(\beta) \geq \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}).$$

Definition

We say that S is an allowed set (for Ω) if Ω is weakly decomposable for S.

Example

The group Lasso norm:

$$\Omega(\beta) := \|\beta\|_{2,1} := \sum_{t=1}^{T} \sqrt{|G_t|}\|\beta_{G_t}\|_2, \ \beta \in \mathbb{R}^p,$$

where $G_1, \ldots, G_T$ is a partition of $\{1, \ldots, p\}$ into disjoint groups.
It is (weakly) decomposable for $S = \cup_{t \in \mathcal{T}} G_t$ with $\Omega_{S^c} = \Omega$.
Thus, for any β, $S := \cup\{G_t : \ \|\beta_{G_t}\|_2 \neq 0\}$ is an allowed set.

Example

From Micchelli et al. (2010)
Let $\mathcal{A} \subset [0, \infty)^p$ be some convex cone. Define

$$\Omega(\beta) := \Omega(\beta; \mathcal{A}) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^{p} \left(\frac{\beta_j^2}{a_j} + a_j \right).$$

Let $\mathcal{A}_S := \{a_S : \ a \in \mathcal{A}\}$.

Definition

We call $\mathcal{A}_S$ an allowed set, if $\mathcal{A}_S \subset \mathcal{A}$.

Lemma

Suppose $\mathcal{A}_S$ is an allowed set. Then S is allowed, i.e. S is weakly decomposable for Ω.

We use the notation

$$\|v\|_n^2 := v^T v/n, \ v \in \mathbb{R}^n.$$

Definition

Suppose S is an allowed set. Let $L > 0$ be some constant. The Ω-eigenvalue (for S) is

$$\delta_\Omega(L, S) := \min\left\{ \|X\beta_S - X\beta_{S^c}\|_n : \ \Omega(\beta_S) = 1, \ \Omega^{S^c}(\beta_{S^c}) \le L \right\}.$$

The Ω-effective sparsity is

$$\Gamma_\Omega^2(L, S) := \frac{1}{\delta_\Omega^2(L, S)}.$$

The dual norm of Ω is denoted by Ω_*, that is

$$\Omega_*(w) := \sup_{\Omega(\beta) \leq 1} |w^T \beta|, \ w \in \mathbb{R}^p.$$

We moreover let $\Omega_*^{S^c}$ be the dual norm of Ω^{S^c}.

A sharp oracle inequality

Theorem

Let $\beta \in \mathbb{R}^p$ *be arbitrary and let Let* $S \supset \{j : \beta_j \neq 0\}$ *be an allowed set. Define*

$$\lambda^S := \Omega_* \Big((\epsilon^T X)_S / n \Big), \ \lambda^{S^c} := \Omega_*^{S^c} \Big((\epsilon^T X)_{S^c} / n \Big).$$

Suppose $\lambda > \lambda^{S^c}$. *Define*

$$L_S := \left(\frac{\lambda + \lambda^S}{\lambda - \lambda^{S^c}} \right).$$

Then

$$\| X(\hat{\beta} - \beta^0) \|_n^2 \leq \| X(\beta - \beta^0) \|_n^2 + \Big[(\lambda + \lambda^S) \Big]^2 \Gamma_\Omega^2 (L_S, S).$$

Related results: Bach (2010).

What about convergence of the Ω-estimation error?

Theorem

Let $\beta \in \mathbb{R}^p$ *be arbitrary and let Let* $S \supset \{j : \beta_j \neq 0\}$ *be an allowed set. Define*

$$\lambda^S := \Omega_*\Big((\epsilon^T X)_S/n\Big),\ \lambda^{S^c} := \Omega_*^{S^c}\Big((\epsilon^T X)_{S^c}/n\Big).$$

Suppose

$$\lambda > \lambda^{S^c}.$$

Define for some $0 \leq \delta < 1$

$$L_S := \left(\frac{\lambda + \lambda^S}{\lambda - \lambda^{S^c}}\right)\left(\frac{1+\delta}{1-\delta}\right).$$

Then

$$\|X(\hat{\beta} - \beta^0)\|_n^2 + \delta(\lambda - \lambda^{S^c})\Omega^{S^c}(\hat{\beta}_{S^c}) + \delta(\lambda + \lambda^S)\Omega(\hat{\beta}_S - \beta)$$

$$\leq \|X(\beta - \beta^0)\|_n^2 + \Big[(1+\delta)(\lambda + \lambda^S)\Big]^2 \Gamma_\Omega^2(L_S, S).$$

Special case where $\Omega = \|\cdot\|_1$

Theorem

(Koltchinskii et al. (2011)) Let for $S \subset \{1, \ldots, p\}$

$$\lambda_0 := \|(\epsilon^T X)\|_\infty / n.$$

Define for $\lambda > \lambda_0$

$$L := \frac{\lambda + \lambda_0}{\lambda - \lambda_0}.$$

Then

$$\|X(\hat{\beta} - \beta^0)\|_n^2 \leq \min_{\beta \in \mathbb{R}^p} \left\{ \|X(\beta - \beta^0)\|_n^2 + (\lambda + \lambda_0)^2 \Gamma^2(L, \|\beta\|_0) \right\}.$$

Compatibility (restricted eigenvalue condition)

Recall that for the ℓ_1-norm

$$\Gamma^2(L,S) = \frac{1}{\delta^2(L,S)},$$

with

$$\delta(L,S) := \min\Big\{ \|X\beta_S - X\beta_{S^c}\|_n : \ \|\beta_S\|_1 = 1, \ \|\beta_{S^c}\|_1 \leq L \Big\}.$$

We have

$$\Gamma^2(L,S) \leq \frac{|S|}{\kappa^2(L,S)},$$

where $\kappa^2(L,S)$ is the restricted eigenvalue (Bickel et al. (2009)).

Consider the case $S = \{1\}$, and write $X_1 := X_S$, $X_2 := X_{S^c}$. Let $X_1 \hat{P} X_2$ be the projection (in $\mathbb{R}^n$) of X_1 on X_2 and $X_1 \hat{A} X_2 := X_1 - X_1 \hat{P} X_2$ be the antiprojection. Define

$$\hat{\gamma}^0 := \arg\min\{\|\gamma\|_1 : \ X_1 \hat{P} X_2 = X_2 \gamma\}.$$

Then clearly

$$\delta(L, \{1\}) = \|X_1 \hat{A} X_2\|_n \ \forall \ L \geq \|\hat{\gamma}^0\|_1.$$

When $n < p$ one readily sees that

$$\delta(L, \{1\}) = 0 \ \forall \ L \geq \|\hat{\gamma}^0\|_1.$$

Suppose now that the rows of X are i.i.d. with sub-Gaussian distribution Q. Let $X_1 P X_2$ be the projection of X_1 on X_2 in $L_2(Q)$ and $X_1 A X_2 := X_1 - X_1 P X_2$. Let $\| \cdot \|$ be the $L_2(Q)$-norm. Define

$$\gamma^0 := \arg\min\{\|\gamma\|_1 : \ X_1 P X_2 = X_2 \gamma\}.$$

Then with large probability, for $L\sqrt{\log p/n}$ small

$$\delta(L, S) \geq (1 - \epsilon)\|X_1 A X_2\| \ \forall \ L \geq \|\gamma^0\|_1.$$

and moreover,

$$(X_1 A X_1)^T (X_1 P X_2)/n \asymp \sqrt{\frac{\log p}{n}}.$$

Oracle inequalities for parameters of interest

High-dimensional linear model:

$$Y = X_1\beta_1^0 + X_2\beta_2^0 + \epsilon,$$

$$\beta_1^0 \in \mathbb{R}^q,\ \beta_2^0 \in \mathbb{R}^{p-q},$$

and the entries of ϵ i.i.d. sub-Gaussian. Suppose the rows of X are i.i.d with sub-Gaussian distribution Q.
We are interested in estimating β_1^0.
Lasso estimator:

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_1) := \arg\min_{\beta_1,\ \beta_2} \left\{ \|Y - X_1\beta_1 - X_2\beta_2\|_2^2/n + \lambda\|\beta_1\|_1 + \lambda\|\beta_2\|_1 \right\}.$$

Notation

Let $X_1 P X_2$ be the projection of X_1 on X_2 in $L_2(Q)$, and define

$$\tilde{X}_1 := X_1 - X_1 P X_2 = X_1 A X_2 .$$

Let

$$\Sigma_1 := \mathbb{E} \tilde{X}_1^T \tilde{X}_1 / n,$$

and let $\tilde{\Lambda}_1^2$ be its smallest eigenvalue.
Define

$$C^0 := \arg\min \left\{ \|C\|_{1,\infty} : \; X_1 P X_2 = X_2 C \right\},$$

where

$$\|C\|_{1,\infty} := \max_{1 \le k \le q} \|\gamma_k\|_1 , \; C := (\gamma_1, \ldots, \gamma_{p-q}).$$

Condition 1 $1/\tilde{\Lambda}_1 = \mathcal{O}(1)$

Condition 2 $\|\beta^0\|_1 = \mathcal{O}(1)$ and $s_1 := \|\beta_1^0\|_0 \vee 1 = o\left(\sqrt{\frac{n}{\log p}}\right)$.

Theorem

Take $\lambda \asymp \sqrt{\log p / n}$. *Then*

$$\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(1).$$

If moreover

$$\|C^0\|_{1,\infty} = \mathcal{O}(1) \text{ (\textit{i.e.} } \ell_1 - \textit{smoothness of the projection}),$$

then

$$\|\hat{\beta}_1 - \beta_1^0\|_1 = \mathcal{O}_{\mathbb{P}}\left(s_1 \sqrt{\frac{\log p}{n}}\right) = o_{\mathbb{P}}(1).$$

Special case: $q = 1$ (recall $q = \dim(\beta_1)$). Then $s_1 = 1$ and hence

$$|\hat{\beta}_1 - \beta_1^0| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right).$$

The high-dimensional partial linear model

Joint work with *Patric Müller*.

Additive model:

$$Y = X\beta^0 + g^0(Z) + \epsilon, \text{ with } \epsilon \perp (X, Z).$$

We assume that the entries of $(X, Z) \in \mathbb{R}^p \times \mathcal{Z}$ are i.i.d. with distribution Q and that the entries of ϵ are i.i.d. sub-Gaussian.

We will assume that g^0 has a given "smoothness" $m > 1/2$ and that β^0 is sparse, with $X\beta^0$ is "smoother" than g^0.

Estimator:

$$(\hat{\beta}, \hat{g}) := \arg\min_{\beta,\, g} \left\{ \|Y - X\beta - g(Z)\|_2^2/n + \lambda\|\beta\|_1 + \mu^2 J^2(g) \right\},$$

where J is some (semi-)norm on the space of functions on $\mathcal{Z}$.

Notation

We write $\tilde{X} := XAZ := X - XPZ$ where $XPZ := E(X|Z)$.

The smallest eigenvalue of $\mathbb{E}\tilde{X}^T\tilde{X}/n$ is denoted by $\tilde{\Lambda}^2$.

The largest eigenvalue of $\mathbb{E}(XPZ)^T(XPZ)/n$ is denoted by Λ_P^2.

$\|\cdot\|$ is the $L_2(Q)$-norm.

Condition 1 $\max_{i,j} |X_{i,j}| = \mathcal{O}(1)$.
Condition 2 $1/\tilde{\Lambda} = \mathcal{O}(1)$ *and* $\Lambda_P = \mathcal{O}(1)$.
Condition 3 *For some fixed constant A it holds that*

$$\mathcal{H}\big(u, \{g: \ \|g\| \le 1, \ J(g) \le 1\}, \|\cdot\|_\infty\big) \le Au^{-1/m}, \ u > 0.$$

Condition 4

$$\sup_{\|g\| \le 1, \ J(g) \le 1} \|g\|_\infty = \mathcal{O}(1).$$

Condition 5 $s := \|\beta^0\|_0 = o(n^{\frac{1}{2m+1}} / \log p)$ *and* $J(g^0) = \mathcal{O}(1)$.

Theorem

Take $\lambda \asymp \sqrt{\log p/n}$ *and* $\mu \asymp n^{-\frac{m}{2m+1}}$. *Then*

$$\|X(\hat{\beta} - \beta^0) + (\hat{g} - g^0)\|^2 + \lambda\|\hat{\beta} - \beta^0\|_1 + \mu^2 J^2(\hat{g}) = \mathcal{O}_{\mathbb{P}}(n^{-\frac{2m}{2m+1}}).$$

If moreover

$$J(h) = \mathcal{O}(1),$$

where $h(Z) = E(X|Z)$ *(i.e. J-smoothness of the projection) then*

$$\|\tilde{X}(\hat{\beta} - \beta^0)\|^2 + \lambda\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}\left(\frac{s \log p}{n}\right) = o_{\mathbb{P}}(n^{-\frac{2m}{2m+1}}).$$

The additive model with different smoothness per component

Joint work with *Enno Mammen*
Additive model:

$$Y = f^0(X) + g^0(Z) + \epsilon \text{ with } \epsilon \perp (X, Z)$$

We assume that the entries of $(X, Z) \in \mathcal{X} \times \mathcal{Z}$ are i.i.d. with distribution $Q_{X,Z}$ and that the entries of ϵ are i.i.d. sub-Gaussian.

The density of $Q_{X,Z}$ with respect to some product measure is denoted by $q_{X,Z}$, with marginal densities q_X and q_Z.

We will assume that f^0 has given "smoothness" $k > 1/2$ and g^0 has given "smoothness" $m > 1/2$, with $k > m$ (i.e., f^0 is "smoother" than g^0).

Notation:

We define

$$r(x,z) := \frac{q_{X,Z}(x,z)}{q_X(x)q_Z(z)},$$

and

$$\gamma_\infty^2 := \|r(\cdot,\cdot)\|_\infty.$$

Moreover, we let

$$\gamma^2 := \int (r-1)^2 q_X q_Z.$$

We define

$$f_P = E(f(X)|Z=\cdot),\ f_A := f - f_P.$$

Condition 1 *For some fixed constants A_I and A_J it holds that*

$$\mathcal{H}_B(u, \{f : \ \|f\| \le 1, \ I(f) \le 1\}, \| \cdot \|) \le A_I u^{-1/k}, \ u > 0,$$

and

$$\mathcal{H}_B(u, \{g : \ \|g\| \le 1, \ J(g) \le 1\}, \| \cdot \|) \le A_J u^{-1/m}, \ u > 0.$$

Condition 2 *For all $R \le 1$ and for some fixed constants B_I and B_J it holds that*

$$\sup_{\|f\| \le R, \ I(f) \le 1} \|f\|_\infty \le B_I R^{1-\frac{1}{2k}},$$

and

$$\sup_{\|g\| \le R, \ J(g) \le 1} \|g\|_\infty \le B_J R^{1-\frac{1}{2m}}.$$

Condition 3 *It holds that $\gamma < 1$.*
Condition 4 $I(f^0) = \mathcal{O}(1)$ *and* $J(g^0) = \mathcal{O}(1)$.

Theorem

Take $\lambda \asymp n^{-\frac{k}{2k+1}}$ *and* $\mu \asymp n^{-\frac{m}{2m+1}}$. *Then*

$$\|\hat{f} - f^0 + \hat{g} - g^0\|^2 + \lambda^2 I^2(\hat{f}) + \mu^2 J^2(\hat{g}) = \mathcal{O}_{\mathbb{P}}(n^{-\frac{2m}{2m+1}}).$$

If moreover for some constant Γ *and for all* f, $J(f_P) \leq \Gamma\|f\|$
(i.e. J-smoothness of the projection), then

$$\|\hat{f} - f^0\|^2 + \lambda^2 I^2(\hat{f}) = \mathcal{O}_{\mathbb{P}}(n^{-\frac{2k}{2k+1}}) = o_{\mathbb{P}}(n^{-\frac{2m}{2m+1}}).$$

Conclusion

- The theory for the ℓ_1-penalty goes through for any weakly decomposable norms

- Sparsity oracle inequalities however require small ”effective sparsity” (i.e., on restricted eigenvalues or compatibility conditions)

- If one is only interested in specific components, one can relax the compatibility conditions

- But then one ”needs” to assume sparse projections on the nuisance part, or ...

- Or replace sparsity assumptions by smoothness assumptions...