Referee's report manuscript ... Faithful variable screening for high-dimensional convex regression

Authors : Min Xu, Minhua Chen and John Lafferty

SUMMARY

The paper deals with the problem of variable selection in nonparametric convex regression in dimension p. The model is the following :

$$Y^{(i)} = f(X^{(i)}) + W^{(i)} \quad \text{for } i = 1, \dots, n$$

where the $X^{(i)}$'s and the $W^{(i)}$'s are i.i.d.

Assuming f is sparse, the goal is to identify the set S_0 of indices of relevant variables.

Notation :

for a vector $\mathbf{x} \in \mathbb{R}^p$, and $j = 1, \dots, p$, $\mathbf{x}_{-j}$ denotes the vector with the j-th coordinate removed and $\mathbf{x}_S$ is the subvector of x restricted to the coordinates in S.

For a vector $x_j \in \mathbb{R}^n$, x_{ij} designates its i^{th} component.

The authors show that, for convex functions, it is sufficient to estimate a sum of one-dimensional convex functions, leading to significant computational advantages, and allowing the intrinsic dimension s to scale polynomially in n, rather than exponentially as in the general case (cf [1]).

More precisely, an additive approximation of f is given by

$$\{f_k^*\}, \mu^* := \arg \min_{\{f_k\}, \mu} \big\{ \mathbf{E}\big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k)\big)^2 \quad : \mathbf{E} f_k(X_k) = 0 \big\}. \qquad (1)$$

f is said "additively faithful" in case $f_k^* = 0$ implies that f does not depend on x_k. A density p supported on $[0,1]^p$ satisfies the "boundary flatness condition" if for all j and for all x_j,

$$\frac{\partial p(\mathbf{x_{-j}} | x_j |)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x_{-j}} | x_j |)}{\partial x_j^2} \quad \text{at } x_j = 0 \text{ and } x_j = 1.$$

The authors show that f is additively faithful if the density p of the design satisfies the boundary flatness condition and if f is convex and twice differentiable. The boundary flatness condition is shown to be a rather weak condition.

The optimal additive functions in (1) could be estimated by smoothing but the authors prefer an approach that is free of smoothing parameters : the additive model is approximated by a convex additive model, that is to say, they look

for the solutions of (1) with f_k^* convex. Faithfulness is not assured anymore. But it can be restored by coupling the f_k^*'s with a set of univariate concave fits on the residual $f - f^*$:

$$\{g_k^*\} := \arg \min_{g_k \in -\mathcal{C}_0^1} \{\mathbf{E}\big(f(X) - \mu^* - \sum_{k \neq k'} f_{k'}^*(X_{k'}) - g_k(X_k)\big)^2\}.$$

where $\mathcal{C}_0^1$ (resp. $-\mathcal{C}_0^1$) denotes the set of one-dimensional convex (resp. concave) functions with population mean zero.

The authors show that if p is a positive density that satisfies the boundary flatness condition and if $\frac{\partial p(\mathbf{x_{-j}}|x_j|)}{\partial x_j}$ and $\frac{\partial^2 p(\mathbf{x_{-j}}|x_j|)}{\partial x_j^2}$ are all continuous as functions of x_j, and if f is convex and twice differentiable, then , $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k.

Notation : $S = \{k : f_k^* = 0 \text{ and } g_k^* = 0\}^c$.

These are the main results for the population setting.

Next the authors describe the optimization procedure in the finite sample case :.

- *Input* : $(x_1, y_1), \ldots, (x_n, y_n)$, regularization parameter λ.
- *First stage* : Estimate a sparse addtive convex model :

$$\hat{f}_1, \ldots, \hat{f}_p, \hat{\mu} = \arg \min_{f_1, \ldots, f_p \in \mathcal{C}_0^1} \frac{1}{n} \sum_{i=1}^n \big(y_i - \mu - \sum_{k=1}^p f_k(x_{ik})\big)^2 + \lambda \sum_{i=1}^p \|f_k\|_\infty$$

- *second stage* : estimate concave functions for each k such that $\left\|\hat{f}_k\right\|_\infty = 0$

$$\hat{g}_k = \arg \min_{g_k \in -\mathcal{C}_0^1} \frac{1}{n} \sum_{i=1}^n \big(y_i - \hat{\mu} - \sum_{k=1}^p \hat{f}_k(x_{ik}) - g_k(x_{ik})\big)^2 + \lambda \sum_{i=1}^p \|g_k\|_\infty$$

- *output* : component functions $\{\hat{f}_k\}$ and relevant variables $\hat{S}$ where

$$\hat{S}^c = \{k : \hat{f}_k = 0 \text{ and } \hat{g}_k = 0\}$$

The optimization appears to be infinite dimensional but the authors show it is equivalent to a finite dimensional quadratic program, using the fact that the subgradients of f_k are non-decreasing for f_k convex. This gives $O(np)$ constraints :

$$f_{i+1k} = f_{ik} + \beta_{ik}(x_{i+1k} - x_{ik}) \text{ for all } i, i' = 1, \ldots, n,$$

with

$$\beta_{i+1k} \geq \beta_{ik}$$

if $x_{1k} \leq \ldots \leq x_{nk}$.

The authors then suggest to use a block coordinate descent method.

An alternative formulation is suggested, replacing the order constraints with positivity constraints, which simplifies the analysis.

Finally, a section is devoted to the conditions under which $P(\hat{S} = S)$ is small. The assumptions made on X, W and f are the following ones : (called assumptions A)

- X_S and X_{S^c} are independent.
- f is convex, twice differentiable (and $\mathbf{E}f(X) = 0$).
- $\|f\|_\infty \leq B$ and $\|f_k^*\|_\infty \leq B$ for all k and for a certain constant B.
- W is mean zero sub-Gaussian, independent of X, with scale σ, i.e., for all $t \in \mathbb{R}$, $\mathbf{E}e^{tW} \leq e^{\sigma^2 t^2/2}$.
- The density p is bounded away from 0 and satisfies the boundary flatness condition.
- $p = O\big(\exp(cn)\big)$ for $0 < c < 1$.

(the assumption of B-boundedness is only used to apply the bracketing number results of [2])

- Results for the control of false positives :

  If assumptions A hold and if $\lambda \geq 512s \max(\sigma, B)\sqrt{\frac{\log^2 np}{n}}$, then the solution to the optimization procedure (with the extra constraint that $\hat{f}_k$ and $\hat{g}_k$ be bounded by B) is such that $\hat{f}_k = \hat{g}_k = 0$ for all $k \in S^c$.

- Results for the control of false positives :

  Notation :

  $$\alpha_+ = \inf_{h \in \mathcal{C}_p :\ \mathrm{supp}(h) \subsetneq \mathrm{supp}(f^*)} \big\{ \mathbf{E}\big(f(X) - h(X)\big)^2 - \mathbf{E}\big(f(X) - f^*(X)\big)^2 \big\}$$

  $$\alpha_- = \min_{k \in S : g_k^* \neq 0} \big\{ \mathbf{E}\big(f(X) - h(X)\big)^2 - \mathbf{E}\big(f(X) - f^*(X) - g_k^*(X_k)\big)^2 \big\}$$

  where $\mathcal{C}_p$ is the set of additive convex functions.

  Suppose assumptions A hold ,

  $$\lambda \leq 512s \max(\sigma, B)\sqrt{\frac{\log^2 np}{n}},$$

  $$\frac{n^{4/5}}{\log np} \geq c' B^4 \max(\sigma, B)^2 s^5$$

  and that the signal-to-noise ratio satisfies

  $$\frac{\alpha_+}{\max(\sigma, B)} \geq cB^2 \sqrt{\frac{s^5}{n^{4/5}} \log^2 np}$$

  and

  $$\frac{\alpha_-}{\max(\sigma, B)} \geq cB^2 \sqrt{\frac{s^5}{n^{4/5}} \log^2 np}$$

  then the solutions to the optimization procedure (with the extra constraint that $\hat{f}_k$ and $\hat{g}_k$ be bounded by B) is such that $\hat{f}_k \neq 0$ or $\hat{g}_k \neq 0$ for all $k \in S$.

The method is illustrated with a simulation where the function f is such that $f(\mathbf{x}) = \mathbf{x}_S^T Q \mathbf{x}_S$ with a symmetric positive definite matrix Q with different values of p and n. The authors also vary the dependence in the design and the sparsity of the matrix Q. The method is also aplied to the Boston housing data.The method is in particular compared to LASSO.

EVALUATION

The results of the present paper are interesting. The paper is rather well written, self-contained and quite complete : first they show that in the case of convex regression, an additive approximation is sufficient, then they develop algorithms for the efficient implementation of the quadratic programs required by the procedure and they give a finite sample statistical analysis, showing in particular that a sample size growing as $n = O((poly(s) \log p)$ is sufficient for variable selection in the convex case. The method is illustrated in simulations (and with real data). However, I found many (unimportant) mistakes in the proofs (I did not mention them all in the minor comments), which need to be read again carefully.

Finally, I think that the paper can be published in AOS after minor revision.

MINOR COMMENTS

- I have a problem with the assertion of uniqueness in Lemma 3.1 and Therorem 3.2. First, in Lemma 3.1, it is clear that f^* is unique. But $(f_1^*, \ldots, f_p^*)$ is not. What is shown is that, for k fixed and given $(f_j^*)_{j \neq k}$, f_k^* is unique. Next, in Theorem 3.2, the proof of uniqueness (bottom of page 16 and top of page 17) is not at all correct.

  Now, to see why I think that, with the assumtions made on X, there is not uniqueness, simply suppose for instance that $p = 2$ and $X_2 = g(X_1)$ for a certain function g. Then if (f_1^*, f_2^*) is a solution, $(f_1^* + f_2^* \circ g, 0)$ is also a solution.

- In the definition of g_k^* on pages 6 and 15, it would be more precise to write $g_k(X_k)$ in the sum instead of g_k.

- On page 9, in the proof of Lemma 3.1, the star has been forgotten twice on $f_{k'}$ in (3.2) and (3.7).

- On page 10, on line 3, "therefore the solution $f_k^*(x_k) = \mathbf{E}[f(X)|x_k] - \mathbf{E}f(X)$ is unique", the term $-\sum_{k' \neq k} f_{k'}(X_{k'})$ has been forgotten.

- On page 10, on line 14, "In particular, if F is the uniform distribution etc", the term $-\int f$ has been forgotten.

- On page 10, on line 16, "additively" should be replaced by "additive".

- On page 11, at the bottom of the page, "before we presenting"

- On page 11, after Definition 3.2, it is written "when the joint density satisfies the property that $\frac{\partial p(\mathbf{x_{-j}}, x_j)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x_{-j}}, x_j)}{\partial x_j^2} = 0$ at boundary points". I think the condition "$p(\mathbf{x}_{-j}, x_j) = 0$ at the boundary points" is missing.

- On page 12, I did not understand why the functions $\frac{\partial^2 p(\mathbf{x_{-j}}|x_j)}{\partial x_j^2}$ and and $\frac{\partial^2 f(\mathbf{x_{-j}}|x_j)}{\partial x_j^2}$ are bounded (they are not supposed to be continuous).

- On page 13, on line 6, "this this".

- On page 13, on line 23, I do not think Σ has been defined. Besides it is not specified that $\mathrm{Var}(X_1) = 1$.
- On page 13, on line 29, "additive" should be replaced by "additively".
- On page 13, the assumption $\mathbf{E}f(X) = 0$ should be made, or else the parameter μ should be introduced as before.
- On page 16, in the expresson of c^*, the denominator should be $\mathbf{E}X_k^2 - m_k^2$.
- On page 16, in the proof of Theorem 3.2, Lemma 8.3 is used. This lemma supposes that ϕ is continuous. Here I do not see why $\phi = \sum_{k' \neq k} f_{k'}^*$ is continuous (it is convex then continuous on the interior of $[0,1]^p$.).
- On page 17, same problem as before with the absence of parameter μ.
- On page 17, on line 16 "for that variable. for each"
- On page 18, at the bottom of the page : " $\beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k}$ for $i = 1, \ldots, n-1$" , $n-1$ must be replaced by $n-2$.
- On page 19, $\sum_i f_{ik}$ instead of $\sum_i f_{ki}$.
- On page 20, in the third paragraph, on the second, third and fourth lines, j and $j-1$ must be replaced by $j+1$ and j respectively in $x_{\pi_k(j)k}$ and $x_{\pi_k(j-1)k}$.
- On page 20, In the central matrix equation, the first vector should be $\begin{bmatrix} f_k(x_{1k} \\ f_k(x_{2k} \\ \vdots \\ f_k(x_{nk} \end{bmatrix}$. Next in "$\mu_k = -\frac{1}{n}\mathbf{1}_n^T \Delta_k d_k$ ", $\mathbf{1}_n$ is missing .
- On page 21, in Section 5, at the end of the second paragraph, "$\hat{mu}$ from our estimation procedure".
- On page 22, in Definition 5.1, the inequalities about d_k must be replaced by the same inequalities as in Proposition 4.1.
- On page 22 (and on page 31), in the statement of Theorem 5.1, in $\max_{i=1,\ldots,n} \frac{X_{k\pi_k(i+1)} - X_{k\pi_k(i)})}{range_k} \geq \frac{1}{16}$ the sign $\geq$ must be replaced by $\leq$. Besides I think it is a bad idea to change notation : $X_{k\pi_k(i)}$ should be replaced by $X_{\pi_k(i)k}$ as before (page 19 for instance).
- On page 23, in equation (5.1), in " $f^*(X)$", the star must be removed.
- On page 24, "$\alpha_+ > 0$ since f is the unique risk minimizer" : I do not understand this statement (there is an infimum, not a minimum in the defintion of α_+.)
- On page 24, in Remark 5.1, "In important direction".
- On page 25, in the statement of Theorem 5.3,the square must be removed on α_-.

- On page 25, after the statement of Corollary 5.1 : I do not understand why " $p = O(\exp(n^c)$" is written instead of " $p = O(\exp(cn)$.

- On page 26, in the first paragraph "the relevant variable set" and in the third paragraph, "seting".

- I really did not understand what was represented in figures 5(a) and 5(b) (axis?). Besides it does not seem to correspond with what is described at the bottom of page 26 and at the top of page 27. The notation $\|.\|_{\infty,1}$ is not defined.

- On page 31, at the top of the page, "see discussion at beginning of Section 5" : I think it should be "at the beginning". In (8.1), same problem as before with the indices : "$\nu_{ki} d_{ki}$ " must be replaced with "$\nu_{ik} d_{ik}$" (see for instance page 20). At the bottom of the page, v_{ki} must be replaced by ν_{ki}.

- On page 32, on the first line of the second paragraph, "$d_k = 0$ for $k \in S$", S must be replaced by S^c and $\|\mu\|$ by $\|\mu\|_1$ and "It clear that". Next, the third paragraph ("to ease notational ....") should be placed at the beginning of the proof (it is used from the beginning of the proof, cf conditions on d_k).

- On page 33, I find that $[\lambda \Delta_k^T \mathbf{u}]_1 = \lambda(X_{kn} - X_{k1})\kappa$. (If κ is be defined in the following way : $\kappa = \frac{1}{\lambda n(X_{kn} - X_{k1})}[\Delta_k^T \mathbf{u}]_1$, then everything goes well).

- On page 33, on the second line, "other rows of stationarity condition holds" and in the third paragraph : "following our strategy, We".

- On page 34, at the bottom of the page, the parentheses in the denominator of the fraction.

- On page 35, in the third paragraph, "as a function of ν" : it is ζ, not ν. In the fourth paragraph, ν is not defined. In the fifth paragraph, in the equation, remove the sum $\sum_{k=1}^p \bar{\Delta}_k^T \mathbf{u}_k$ and replace it by the vector with components $(\bar{\Delta}_k^T \mathbf{u}_k)_{k=1,\dots,p}$. In the sixth paragraph, in $\bar{\Delta}\hat{d}_k = \bar{\Delta}\hat{d}$, replace $\hat{d}_k$ by $\hat{d}'$.

- On page 36, maybe it would be better to replace d by c here. Add "for $k \in S^c$ after " to show that $\hat{c}_k = 0$".

- On page 37, In the first paragraph, I do not understand why p appears here. In the last paragraph, I find 4 instead of 2 on the first term of the second line (and then 12 instead of 8 on the third line) .

- on page 38, In the sixth paragraph, in the inequality, the sign should be $\leq$. In the eighth paragraph, c'' must be replaced by c. In the ninth paragraph, "Taking an union".

- On page 40, on the fifth line, "empricial". In the last paragraph, "whose size is bounded" : it is the log of the size.

- I had a problem with the definition of an ϵ-bracketing. First, in the definition 8.1, on page 49, " we define a bracketing" should be replaced by " we define a ϵ-bracketing". Then it is said that $\rho(f_L, f_U) \leq \epsilon$ and that the corresponding size is $N(\epsilon, \mathcal{C}, \rho)$. But then, in Proposition 8.3 (and almost everywhere else), when $N(2\epsilon, L_2(P), \rho)$ is used, it is written $\|f_L - f_U\|_{L_2} \leq \epsilon$ (instead of 2ϵ.) (same thing in Corollary 8.4 for instance : ϵ-bracketing with a size $N(2\epsilon, ...)$)
- On page 41, problem with ϵ (same as the previous item).
- On page 40, I think the functions in $\mathcal{G}$ are bounded by $2sB$ (not sB.).
- On page 49, in Definition 8.1, f^U and f^L must be replaced by f_U and f_L. On the last line of the page, "additive convex functions with s components" : add "components bounded by B".
- On page 49, "$\int p(x)^2 dx \leq (\int p(x) dx)^2$ "??
- On page 50, a constant 2 is missing in ϵ_n.
- On page 41, in the second paragraph, I think the bound on $\sup_{h_L} |\langle w, h_L \rangle_n|$ is wrong. I would have used the fact that the variables $h_L(X_i)W_i$ are independent, centered and sub-Gaussian with a scale smaller than $2\sigma sB$. It gives a bound of order $sB\sigma\sqrt{\frac{\log \frac{2}{\delta}}{n}}$.
- The term ϵ supposed to balance the two terms is not correct (for instance it does not contain δ). There are many small mistakes in the calculations. In particular, the exponent in s is not the right one.
- In the proof of Theorem 8.4 : too many small mistakes in calculations, but the final result is OK.
- On page 45, the term ϵ supposed to balance the terms is not correct (same remark about δ). In the last equations, on the first line (bottom of the page) n is missing in one of the scalar products.
- On page 46: "taking a union bound and we have that"
- In Lemma 8.3, I do not see why the functions are continuous : the functions $\mathbf{x}_{-k} \to p(\mathbf{x}_{-k} \mid x_k)$ are not supposed to be continuous .
- On page 47, in the proof of Lemma 8.4 : notation ϕ_{-j} has not been defined. In the definition of A_+, the sign should be $>$, not $\geq$. Remove "both" .
- On page 47, "a sub-exponential random is".

References

[1] LaÃńtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40(5):2667–2696, 10 2012.

[2] Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. *arXiv preprint arXiv:1404.2298*, 2014.