Review of "Faithful variable screening for high-dimensional convex regression"

Summary

This paper presents a new method for variable selection in convex nonparametric regression. The authors present a variable selection algorithm, called the AC/DC (additive convex/decoupled concave) procedure, which provably recovers the relevant variables under appropriate assumptions on the underlying regression function and data-generating distribution. Optimization may be performed efficiency, and the efficacy of the algorithm is demonstrated via simulation results.

Comments

This is a well-written paper and the contributions are very interesting and cleanly presented. The optimization algorithm provides a nice complement to the statistical theory, and the experimental results demonstrate that this is indeed a viable method with promising performance even in situations beyond the scope of the present theory.

Conceptually, I have a few questions for the authors to address:

(1) If the true regression function f_0 can actually be written as a sum of univariate convex functions (plus noise), are there natural conditions that one can provide for when additive faithfulness will hold? Similarly, is there a simpler version of Theorem 3 that can be stated when f_0 is actually a sum of univariate convex functions? This would seem to depend on an appropriate statement for Assumption (A3).

(2) If we actually wish to fit smooth convex univariate components, is it easy to add this condition as a regularization term and incorporate it into the analysis?

(3) Assuming model selection consistency, can one derive prediction error bounds using the fitted f_k's and g_k's, as well? Prediction error bounds seem quite relevant due to possible model misspecification when f_0 is not actually a sum of univariate convex functions.

In order to improve clarity, the authors should define the convex additive model at the top of page 5: $f(x_i) = \sum_{k=1}^{p} f_k(x_{ik})$, where the f_k's are convex. Otherwise, the notation in equation (2.3) is initially a bit confusing. Another notational comment: the use of the distribution function F seems unnecessary (e.g., Lemma 3.1 and Corollary 3.1), since everything is already defined in terms of the joint density function $p(x)$. (An additional suggestion would be to use a different letter to represent the joint density function, since p is already used to represent the dimensionality of the distribution.)

I would suggest including the dependence on the positive-valued lower bound on $p(x)$ (as assumed in (A5)) in the statement of Theorem 3. This is a point worth commenting upon, since the earlier theorems only require $p(x)$ to be strictly positive. Assumption (A5) seems like it could also be interpreted as a sort of minimum signal strength requirement for successful variable selection.

It would be good to point to Assumption (A3) at the beginning of Section 5 when discussing the B-boundedness condition. In the third paragraph, the authors might note that the B-boundedness condition arises from the analysis of Theorem 5.2. To clarify, the penultimate paragraph on page 22 should state that Theorem 5.1 controls false positives alone, whereas Theorem 5.3 (to follow) provides finite-sample results controlling the false negative rate. At the top of page 23, the remark regarding mutual incoherence conditions is a bit confusing. Aren't the conditions imposed in Theorem 5.1 on the covariates essentially analogous to the mutual incoherence conditions in this case?

In the simulation section, the authors might consider comparing the set of variables selected after the AC step only with the set of variables selected via the entire AC/DC algorithm. This would help illustrate the necessity of the concave fitting step in addition to the convex additive fit alone.

Minor typos:

(i) In Section 2.3, it would be good to actually state the regularized objective function being analyzed, rather than waiting until Section 4.

(ii) Assumption (A3) on page 7 is missing the subscript ∞.

(iii) The parameter σ is never defined in Section 2.3. Reading ahead, it seems to be the sub-Gaussian parameter of the additive noise.

(iv) Perhaps Section 3 should be titled something other than "Additive faithfulness" (maybe "Population-level results"?).

(v) The first paragraph of Section 5 should read "Figure 3" rather than "Algorithm 3."

(vi) The second paragraph of Section 5 has $\hat{mu}$ instead of $\hat{\mu}$.

(vii) The first sentence of Section 5.1 should read optimization (4.8).

Recommendation

I found this paper quite enjoyable to read and believe the content is very much appropriate for the *Annals of Statistics*. However, the paper would benefit from a proper revision according to some of the comments mentioned earlier.