The Annals of Statistics
2016, Vol. 44, No. 6, 2624–2660
DOI: 10.1214/15-AOS1425

FAITHFUL VARIABLE SCREENING FOR HIGH-DIMENSIONAL CONVEX REGRESSION

BY MIN XU[1,*], MINHUA CHEN[†] AND JOHN LAFFERTY[1,‡]

University of Pennsylvania, Amazon.com*[†] *and University of Chicago*[‡]

We study the problem of variable selection in convex nonparametric regression. Under the assumption that the true regression function is convex and sparse, we develop a screening procedure to select a subset of variables that contains the relevant variables. Our approach is a two-stage quadratic programming method that estimates a sum of one-dimensional convex functions, followed by one-dimensional concave regression fits on the residuals. In contrast to previous methods for sparse additive models, the optimization is finite dimensional and requires no tuning parameters for smoothness. Under appropriate assumptions, we prove that the procedure is faithful in the population setting, yielding no false negatives. We give a finite sample statistical analysis, and introduce algorithms for efficiently carrying out the required quadratic programs. The approach leads to computational and statistical advantages over fitting a full model, and provides an effective, practical approach to variable screening in convex regression.

1. Introduction. Shape restrictions such as monotonicity, convexity and concavity provide a natural way of limiting the complexity of many statistical estimation problems. Shape-constrained estimation is not as well understood as more traditional nonparametric estimation involving smoothness constraints. Even the one-dimensional case is interesting and challenging, and has been of recent interest [10].

In this paper, we study the problem of variable selection in multivariate convex regression. Assuming that the regression function is convex and sparse, our goal is to identify the relevant variables. We show that it suffices to estimate a sum of one-dimensional convex functions, leading to significant computational and statistical advantages. This is in contrast to general nonparametric regression, where fitting an additive model can result in false negatives. Our approach is based on a two-stage quadratic programming procedure. In the first stage, we fit an convex additive model, imposing a sparsity penalty. In the second stage, we fit a concave function on the residual for each variable. As we show, this nonintuitive second stage is in

Received November 2014; revised December 2015.

[1]Supported in part by NSF Grants IIS-1116730 and DMS-15-13594, AFOSR Grant FA9550-09-1-0373, ONR Grant N00014-12-10762 and an Amazon AWS in Education Machine Learning Research grant.

MSC2010 subject classifications. Primary 62G08; secondary 52A41.

Key words and phrases. Nonparametric regression, convex regression, variable selection, quadratic programming, additive model.

general necessary. Our first result is that this procedure is faithful in the population setting, meaning that it results in no false negatives, under mild assumptions on the density of the covariates. Our second result is a finite sample statistical analysis of the procedure, where we upper bound the statistical rate of variable screening consistency. An additional contribution is to show how the required quadratic programs can be formulated to be more scalable. We give simulations to illustrate our method, showing that it performs in a manner that is consistent with our analysis.

Estimation of convex functions arises naturally in several applications. Examples include geometric programming [3], computed tomography [22], target reconstruction [17], image analysis [9] and circuit design [11]. Other applications include queuing theory [4] and economics, where it is of interest to estimate concave utility functions [20]. See [18] for other applications. Beyond cases where the assumption of convexity is natural, convexity can be attractive as a tractable, nonparametric relaxation of the linear model.

Recently, there has been increased research activity on shape-constrained estimation. Guntuboyina and Sen [10] analyze univariate convex regression and show surprisingly that the risk of the MLE is adaptive to the complexity of the true function. Seijo and Sen [26] and Lim and Glynn [18] study maximum likelihood estimation of multivariate convex regression and independently establish its consistency. Cule, Samworth and Stewart [7] and Kim and Samworth [14] analyze log-concave density estimation and prove consistency of the MLE; the latter further show that log-concave density estimation has minimax risk lower bounded by $n^{-2/(d+1)}$ for $d \geq 2$, refuting a common notion that the condition of convexity is equivalent, in estimation difficulty, to the condition of having two bounded derivatives. Additive shape-constrained estimation has also been studied; Pya and Wood [23] propose a penalized B-spline estimator while Chen and Samworth [5] show the consistency of the MLE. To the best of our knowledge, however, there has been no work on variable selection and estimation of high-dimensional convex functions.

Variable selection in general nonparametric regression or function estimation is a notoriously difficult problem. Lafferty and Wasserman [16] develop a greedy procedure for adjusting bandwidths in a local linear regression estimator, and show that the procedure achieves the minimax rate as if the relevant variables were isolated in advance. But the method only provably scales to dimensions p that grow logarithmically in the sample size n, that is, $p = O(\log n)$. This is in contrast to the high-dimensional scaling behavior known to hold for sparsity selection in linear models using ℓ_1 penalization, where n is logarithmic in the dimension p. Bertin and Lecué [1] develop an optimization-based approach in the nonparametric setting, applying the lasso in a local linear model at each test point. Here again, however, the method only scales as $p = O(\log n)$, the low-dimensional regime. An approximation theory approach to the same problem is presented in [8], using techniques based on hierarchical hashing schemes, similar to those used for "junta"

problems [21]. Here, it is shown that the sample complexity scales as $n > \log p$ if one adaptively selects the points on which the high-dimensional function is evaluated.

Comminges and Dalalyan [6] show that the exponential scaling $n = O(\log p)$ is achievable if the underlying function is assumed to be smooth with respect to a Fourier basis. They also give support for the intrinsic difficulty of variable selection in nonparametric regression, giving lower bounds showing that consistent variable selection is not possible if $n < \log p$ or if $n < \exp s$, where s is the number of relevant variables. Variable selection over kernel classes is studied by Koltchinskii and Yuan [15].

Perhaps more closely related to the present work is the framework studied by Raskutti, Wainwright and Yu [24] for sparse additive models, where sparse regression is considered under an additive assumption, with each component function belonging to an RKHS. An advantage of working over an RKHS is that nonparametric regression with a sparsity-inducing regularization penalty can be formulated as a finite dimensional convex cone optimization. On the other hand, smoothing parameters for the component Hilbert spaces must be chosen, leading to extra tuning parameters that are difficult to select in practice. There has also been work on estimating sparse additive models over a spline basis, for instance, the work of [13], but these approaches also require the tuning of smoothing parameters.

While nonparametric, the convex regression problem is naturally formulated using finite dimensional convex optimization, with no additional tuning parameters. The convex additive model can be used for convenience, without assuming it to actually hold, for the purpose of variable selection. As we show, our method scales to high dimensions, with a dependence on the intrinsic dimension s that scales polynomially, rather than exponentially as in the general case analyzed in [6].

In the following section, we give a high-level summary of our technical results, including additive faithfulness, variable selection consistency and high-dimensional scaling. In Section 3, we give a detailed account of our method and the conditions under which we can guarantee consistent variable selection. In Section 4, we show how the required quadratic programs can be reformulated to be more efficient and scalable. In Section 5, we give the details of our finite sample analysis, showing that a sample size growing as $n = O(\text{poly}(s) \log p)$ is sufficient for variable selection. In Section 6, we report the results of simulations that illustrate our methods and theory. The full proofs are given in the supplementary material [28].

2. Overview of results. In this section, we provide a high-level description of our technical results. The full technical details, the precise statement of the results and their detailed proofs are provided in following sections.

Our main contribution is an analysis of an additive approximation for identifying relevant variables in convex regression. We prove a result that shows when

and how the additive approximation can be used without introducing false negatives in the population setting. In addition, we develop algorithms for the efficient implementation of the quadratic programs required by the procedure.

We first establish some notation, to be used throughout the paper. If $\mathbf{x}$ is a vector, we use $\mathbf{x}_{-k}$ to denote the vector with the kth coordinate removed. If $\mathbf{v} \in \mathbb{R}^n$, then $v_{(1)}$ denotes the smallest coordinate of $\mathbf{v}$ in magnitude, and $v_{(j)}$ denotes the jth smallest; $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector. If $X \in \mathbb{R}^p$ is a random variable and $S \subset \{1, \ldots, p\}$, then X_S is the subvector of X restricted to the coordinates in S. Given n samples $X^{(1)}, \ldots, X^{(n)}$, we use $\overline{X}$ to denote the sample mean. Given a random variable X_k and a scalar x_k, we use $\mathbb{E}[\cdot \mid x_k]$ as a shorthand for $\mathbb{E}[\cdot \mid X_k = x_k]$. If we say a function is integrable, we mean it is Lebesgue integrable.

2.1. *Faithful screening.* The starting point for our approach is the observation that least squares nonparametric estimation under convexity constraints is equivalent to a finite dimensional quadratic program. Specifically, the infinite dimensional optimization

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} \big(Y_i - f(\mathbf{x}_i)\big)^2 \\ &\text{subject to} && f : \mathbb{R}^p \to \mathbb{R} \qquad \text{is convex} \end{aligned} \tag{2.1}$$

is equivalent to the finite dimensional quadratic program

$$\begin{aligned} &\text{minimize}_{f,\beta} && \sum_{i=1}^{n} (Y_i - f_i)^2 \\ &\text{subject to} && f_j \geq f_i + \beta_i^\top (\mathbf{x}_j - \mathbf{x}_i) \qquad \text{for all } i, j. \end{aligned} \tag{2.2}$$

Here, f_i is the estimated function value $f(\mathbf{x}_i)$, and the vectors $\beta_i \in \mathbb{R}^d$ represent supporting hyperplanes to the epigraph of f. See [3], Section 6.5.5. Importantly, this finite dimensional quadratic program does not have tuning parameters for smoothing the function.

This formulation of convex regression is subject to the curse of dimensionality. Moreover, attempting to select variables by regularizing the subgradient vectors β_i with a group sparsity penality is not effective. Intuitively, the reason is that all p components of the subgradient β_i appear in every convexity constraint $f_j \geq f_i + \beta_i^T (\mathbf{x}_j - \mathbf{x}_i)$; small changes to the subgradients may not violate the constraints. Experimentally, we find that regularization with a group sparsity penality will make the subgradients of irrelevant variables small, but may not zero them out completely.

This motivates us to consider an additive approximation. As we show, this leads to an effective variable selection procedure. The shape constraints play an essential role. For general regression, using an additive approximation for variable selection

may make errors. In particular, the nonlinearities in the regression function may result in an additive component being wrongly zeroed out. We show that this cannot happen for convex regression under appropriate conditions.

We say that a differentiable function f depends on variable x_k if $\partial_{x_k} f \neq 0$ with probability greater than zero. An additive approximation is given by

$$\{f_k^*\}, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_p,\mu} \left\{ \mathbb{E}\left(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \right)^2 : \mathbb{E} f_k(X_k) = 0 \right\}. \tag{2.3}$$

We say that f is *additively faithful* in case $f_k^* = 0$ implies that f does not depend on coordinate k. Additive faithfulness is a desirable property since it implies that an additive approximation may allow us to screen out irrelevant variables.

Our first result shows that convex multivariate functions are additively faithful under the following assumption on the distribution of the data.

DEFINITION 2.1. Let $p(\mathbf{x})$ be a density supported on $[0,1]^p$. Then p satisfies the *boundary flatness condition* if it satisfies certain regularity conditions (see the precise statement in Definition 3.2) and if for all j, and for all $\mathbf{x}_{-j}$,

$$\frac{\partial p(\mathbf{x}_{-j} \mid x_j)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x}_{-j} \mid x_j)}{\partial x_j^2} = 0 \qquad \text{at } x_j = 0 \text{ and } x_j = 1.$$

As discussed in Section 3, this is a relatively weak condition. Our first result is that this condition suffices in the population setting of convex regression.

THEOREM 1. *Let $p(\mathbf{x})$ be a positive density supported on $C = [0,1]^p$ that satisfies the boundary flatness property. If f is convex with a bounded second derivative on an open set around C, then f is additively faithful under p.*

Intuitively, an additive approximation zeroes out variable k when, fixing x_k, every "slice" of f integrates to zero. We prove this result by showing that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

While this shows that convex functions are additively faithful, it is difficult to estimate the optimal additive functions. The difficulty is that f_k^* need not be a convex function, as we show through a counterexample in Section 3. It may be possible to estimate f_k^* with smoothing parameters, but for the purpose of variable screening, it is sufficient in fact to approximate f_k^* by a *convex* additive model.

Our next result states that a convex additive fit, combined with a series of univariate concave fits, is faithful. We abuse notation in Theorem 2 and let the notation f_k^* represent convex additive components.

THEOREM 2. *Suppose $p(\mathbf{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary flatness condition. Suppose that f is convex and continuously twice-differentiable on an open set around C and that the derivatives $\partial_{x_k} f$, $\partial_{x_k} p(\mathbf{x}_{-k} \mid x_k)$, and $\partial^2_{x_k} p(\mathbf{x}_{-k} \mid x_k)$ are all continuous as functions on C. Define*

$$\{f_k^*\}_{k=1}^p, \mu^* = \arg\min_{\{f_k\},\mu} \left\{ \mathbb{E}\left(f(X) - \mu - \sum_{k=1}^s f_k(X_k) \right)^2 : f_k \in \mathcal{C}^1, \right. \\ \left. \mathbb{E} f_k(X_k) = 0 \right\}, \tag{2.4}$$

where $\mathcal{C}^1$ is the set of univariate convex functions. Using the f_k^*s above, define

$$g_k^* = \arg\min_{g_k} \left\{ \mathbb{E}\left(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k(X_k) \right)^2 : g_k \in \text{-}\mathcal{C}^1, \right. \\ \left. \mathbb{E} g_k(X_k) = 0 \right\}, \tag{2.5}$$

with -$\mathcal{C}^1$ denoting the set of univariate concave functions. Then $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k, that is, $\partial_{x_k} f(\mathbf{x}) = 0$ with probability one.

This result naturally suggests a two-stage screening procedure for variable selection. In the first stage, we fit a sparse convex additive model $\{\widehat{f}_k\}$. In the second stage, we fit a concave function $\widehat{g}_k$ to the residual for each variable having a zero convex component $\widehat{f}_k$. If both $\widehat{f}_k = 0$ and $\widehat{g}_k = 0$, we can safely discard variable x_k. As a shorthand, we refer to this two-stage procedure as AC/DC. In the AC stage, we fit an additive convex model. In the DC stage, we fit decoupled concave functions on the residuals. The decoupled nature of the DC stage allows all of the fits to be carried out in parallel. The entire process involves no smoothing parameters. Our next result concerns the required optimizations, and their finite sample statistical performance.

2.2. *Optimization.* Given samples (y_i, X_i), AC/DC becomes the following optimization:

$$\{\widehat{f}_k\}_{k=1}^p = \arg\min_{\{f_k \in \mathcal{C}^1\}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \overline{y} - \sum_{k=1}^p f_k(X_{ik}) \right)^2 + \lambda \sum_{k=1}^p \| f_k \|_\infty,$$

$$\forall k, \widehat{g}_k = \arg\min_{g_k \in \mathcal{C}^1} \frac{1}{n} \sum_{i=1}^n \left(y_i - \overline{y} - \sum_{k' \neq k} \widehat{f}_{k'}(X_{ik'}) - g_k(X_{ik}) \right)^2 + \lambda \| g_k \|_\infty,$$

where $\overline{y}$ is the empirical mean of y. Our estimate of the relevant variables is $\widehat{S} = \{k : \| \widehat{f}_k \| > 0 \text{ or } \| \widehat{g}_k \| > 0\}$.

We present the optimization algorithms in Section 4. The convex constraints for the additive functions, analogous to the multivariate constraints (2.2), are that each component $f_k(\cdot)$ can be represented by its supporting hyperplanes, that is,

$$f_{ki'} \geq f_{ki} + \beta_{ki}(x_{ki'} - x_{ki}) \qquad \text{for all } i, i', \tag{2.6}$$

where $f_{ki} := f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. While this apparently requires $O(n^2 p)$ equations to impose the supporting hyperplane constraints, in fact, only $O(np)$ constraints suffice. This is because univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to a reduced quadratic program with $O(np)$ variables and $O(np)$ constraints.

Directly applying a QP solver to this optimization is still computationally expensive for relatively large n and p. We thus develop a block coordinate descent method, where in each step we solve a sparse quadratic program involving $O(n)$ variables and $O(n)$ constraints. This is efficiently solved using optimization packages such as MOSEK. The details of these optimizations are given in Section 4.

2.3. *Finite sample analysis.* In Section 5, we analyze the finite sample variable selection consistency of AC/DC under the model

$$y_i = f_0(X_i) + w_i \qquad \text{for } i = 1, \ldots, n,$$

without assuming that the true regression function f_0 is additive. Our analysis first establishes a sufficient deterministic condition for variable selection consistency, and then considers a stochastic setting. Our proof technique decomposes the KKT conditions for the optimization in a manner that is similar to the now standard *primal-dual witness* method [27].

We prove separate results that allow us to analyze false negative rates and false positive rates. To control false positives, we analyze scaling conditions on the regularization parameter λ_n for group sparsity needed to zero out irrelevant variables $k \in S^c$, where $S \subset \{1, \ldots, p\}$ is the set of variables selected by the AC/DC algorithm in the population setting. To control false negatives, we analyze the restricted regression where the variables in S^c are zeroed out, following the primal-dual strategy.

Each of our theorems uses a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.
A2: f_0 is convex with a bounded second derivative. $\mathbb{E} f_0(X) = 0$.
A3: $\|f_0\|_\infty \leq sB$ and $\|f_k^*\|_\infty \leq B$ for all k.
A4: The noise is mean-zero sub-Gaussian with scale σ, independent of X.
A5: The density $p(\mathbf{x})$ is bounded away from $0/\infty$ and satisfies the boundary flatness condition.

In assumption A3, $f^* = \sum_k f_k^*$ denotes the optimal additive projection of f_0 in the population setting.

Our analysis involves parameters α_+ and α_-, which are measures of the signal strength of the weakest variable:

$$\alpha_+ = \inf_{f \in \mathcal{C}^p : \mathrm{supp}(f) \subsetneq \mathrm{supp}(f^*)} \{\mathbb{E}(f_0(X) - f(X))^2 - \mathbb{E}(f_0(X) - f^*(X))^2\},$$

$$\alpha_- = \min_{k \in S : g_k^* \neq 0} \{\mathbb{E}(f_0(X) - f^*(X))^2 - \mathbb{E}(f_0(X) - f^*(X) - g_k^*(X_k))^2\}.$$

Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure.

We make strong assumptions on the covariates in A1 in order to make very weak assumptions on the true regression function f_0 in A2; in particular, we do not assume that f_0 is additive. Relaxing this condition is an important direction for future work. We also include an extra boundedness constraint to use new bracketing number results [14].

Our main result is the following.

THEOREM 3. *Suppose assumptions* A1–A5 *hold. Let* $\{\widehat{f}_i\}$ *be any AC solution and let* $\{\widehat{g}_k\}$ *be any DC solution, both estimated with regularization parameter* λ *scaling as* $\lambda = \Theta(s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np})$. *Suppose in addition that*

$$\alpha_f/\widetilde{\sigma} \geq cB^2\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}, \tag{2.7}$$

$$\alpha_g^2/\widetilde{\sigma} \geq cB^4\sqrt{\frac{s^5}{n^{4/5}}\log^2 2np}, \tag{2.8}$$

where $\widetilde{\sigma} \equiv \max(\sigma, B)$ *and* c *is a constant dependent only on* b, c_1. *Then, for sufficiently large* n, *with probability at least* $1 - \frac{1}{n}$,

$$\widehat{f}_k \neq 0 \quad \text{or} \quad \widehat{g}_k \neq 0 \qquad \text{for all } k \in S,$$

$$\widehat{f}_k = 0 \quad \text{and} \quad \widehat{g}_k = 0 \qquad \text{for all } k \notin S.$$

This shows that variable selection consistency is achievable under exponential scaling of the ambient dimension, $p = O(\exp(cn))$ for some $0 < c < 1$, as for linear models. The cost of nonparametric estimation is reflected in the scaling with respect to $s = |S|$, which can grow only as $o(n^{4/25})$.

We remark that Comminges and Dalayan [6] show that, even under the product distribution, variable selection is achievable under traditional smoothness constraints only if $n > O(e^s)$. Here, we demonstrate that convexity yields the scaling $n = O(\mathrm{poly}(s))$.

3. Population level analysis: Additive faithfulness. For a general regression function, an additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent in the approximation and may persist even in the population setting. In this section, we give examples of this phenomenon, and then show how the convexity assumption changes the behavior of the additive approximation. We work with $C = [0,1]^p$ as the support of the distribution in this section but all of our results apply to general hypercubes. We begin with a lemma that characterizes the components of the additive approximation under mild conditions.

LEMMA 3.1. *Let P be a distribution on $C = [0,1]^p$ with a positive density function $p(\mathbf{x})$. Let $f : C \to \mathbb{R}$ be in $L^2(P)$. Let*

$$f_1^*, \ldots, f_p^*, \mu^* := \arg\min \left\{ \mathbb{E}\left(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \right)^2 : f_k \in L^2(P), \mathbb{E} f_k(X_k) = 0, \forall k \right\}.$$

With $\mu^* = \mathbb{E} f(X)$,

$$f_k^*(x_k) = \mathbb{E}\left[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,\middle|\, x_k \right] - \mathbb{E} f(X), \tag{3.1}$$

and this solution is unique.

Lemma 3.1 follows from the stationarity conditions of the optimal solution. This result is known, and criterion (3.1) is used in the backfitting algorithm for fitting additive models. We include a proof as our results build on it.

PROOF OF LEMMA 3.1. Let $f_1^*, \ldots, f_p^*, \mu^*$ be the minimizers as defined; they exist since the set of mean zero additive functions is a closed subspace of $L^2(P)$. We first show that the optimal μ is $\mu^* = \mathbb{E} f(X)$ for any $f_1, \ldots, f_k$ such that $\mathbb{E} f_k(X_k) = 0$. This follows from the stationarity condition, which states that $\mu^* = \mathbb{E}[f(X) - \sum_k f_k(X_k)] = \mathbb{E}[f(X)]$. Uniqueness is apparent because the second derivative is strictly larger than zero and strong convexity is guaranteed.

We now turn our attention toward the f_k^*'s. It must be that f_k^* minimizes

$$\min_{f_k} \mathbb{E}\left(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k) \right)^2 \tag{3.2}$$

subject to $\mathbb{E} f_k(X_k) = 0$. Fixing x_k, we will show that the value

$$\mathbb{E}\left[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,\middle|\, x_k \right] - \mu^* \tag{3.3}$$

uniquely minimizes

$$\min_{f_k(x_k)} \int_{\mathbf{x}_{-k}} p(\mathbf{x}) \Big(f(\mathbf{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - f_k(x_k) - \mu^* \Big)^2 d\mathbf{x}_{-k}. \tag{3.4}$$

The first-order optimality condition gives us

$$\begin{aligned} &\int_{\mathbf{x}_{-k}} p(\mathbf{x}) f_k(x_k)\, d\mathbf{x}_{-k} \\ &\quad = \int_{\mathbf{x}_{-k}} p(\mathbf{x}) \Big(f(\mathbf{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^* \Big) d\mathbf{x}_{-k}, \end{aligned} \tag{3.5}$$

$$p(x_k) f_k(x_k) = \int_{\mathbf{x}_{-k}} p(x_k) p(\mathbf{x}_{-k} \mid x_k) \Big(f(\mathbf{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^* \Big) d\mathbf{x}_{-k}, \tag{3.6}$$

$$f_k(x_k) = \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \mid x_k) \Big(f(\mathbf{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^* \Big) d\mathbf{x}_{-k}. \tag{3.7}$$

To prove uniqueness, suppose $\widetilde{f} = \sum_{j=1}^p \widetilde{f}_j$ is another additive function that achieves the same square error. Let $\nu \in [0,1]$, and consider the expectation $\mathbb{E}(f(X) - \mu^* - (f^* + \nu(\widetilde{f} - f^*)))^2$ as a function of ν. The objective is strongly convex if $\mathbb{E}(\widetilde{f} - f^*)^2$, and so $\mathbb{E}(\widetilde{f} - f^*)^2 = 0$ by the assumption that f^* and $\widetilde{f}$ are both optimal solutions. By Lemma 1.3 in the supplement, we conclude that $\mathbb{E}(f_j^* - \widetilde{f}_j)^2 = 0$ as well, and thus, $f_j^* = \widetilde{f}_j$ almost everywhere.

We note that $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \mid x_k] - \mathbb{E} f(X)$ has mean zero as a function of x_k, which shows that the f_k^*s are feasible. □

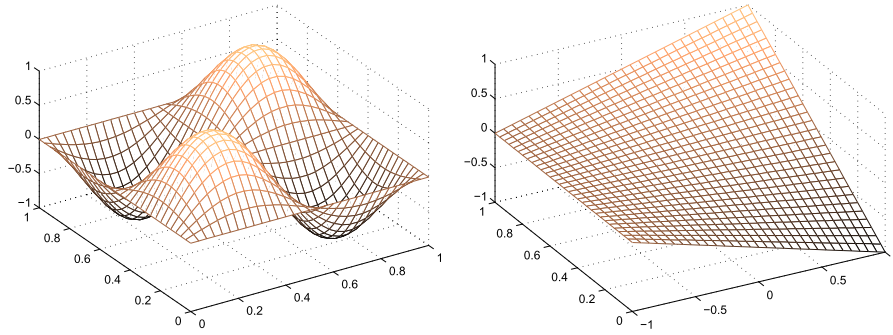
In the case that the distribution in Lemma 3.1 is a product distribution, the additive components take on a simple form.

COROLLARY 3.1. *Let $p(\mathbf{x})$ be a positive density on $C = [0,1]^p$. Let μ^*, $f_k^*(x_k)$ be defined as in Lemma 3.1. Then $\mu^* = \mathbb{E} f(X)$ and $f_k^*(x_k) = \mathbb{E}[f(X) \mid x_k] - \mathbb{E} f(X)$ and this solution is unique.*

In particular, under the uniform distribution, $f_k^*(x_k) = \int f(x_k, \mathbf{x}_{-k})\, d\mathbf{x}_{-k} - \int f(\mathbf{x})\, d\mathbf{x}$.

EXAMPLE 3.1. Using Corollary 3.1, we give two examples of *additive unfaithfulness* under the uniform distribution—where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

$$f(x_1, x_2) = \sin(2\pi x_1) \sin(2\pi x_2) \qquad \text{(egg carton)} \tag{3.8}$$



(a) egg carton (b) tilting slope

FIG. 1. *Two additively unfaithful functions. Relevant variables are zeroed out under an additive approximation because every "slice" of the function integrates to zero.*

defined for $(x_1, x_2) \in [0,1]^2$. Then $\int_{x_2} f(x_1, x_2)\, dx_2 = 0$ and $\int_{x_1} f(x_1, x_2)\, dx_1 = 0$ for each x_1 and x_2. An additive approximation would set $f_1 = 0$ and $f_2 = 0$. Next, consider the function

$$f(x_1, x_2) = x_1 x_2 \qquad \text{(tilting slope)} \tag{3.9}$$

defined for $x_1 \in [-1, 1]$, $x_2 \in [0, 1]$. In this case $\int_{x_1} f(x_1, x_2)\, dx_1 = 0$ for each x_2; therefore, we expect $f_2 = 0$ under the additive approximation. This function, for every fixed x_2, is a zero-intercept linear function of x_1 with slope x_2. See Figure 1.

In order to exploit additive models in variable selection, it is important to understand when the additive approximation accurately captures all of the relevant variables. We call this property *additive faithfulness*. We first formalize the concept that a multivariate function f *does not depend on* a coordinate x_k.

DEFINITION 3.1. Let $C = [0,1]^p$ and let $f : C \to \mathbb{R}$. We say that f *does not depend on coordinate* k if for all $\mathbf{x}_{-k}$, $f(x_k, \mathbf{x}_{-k})$ is a constant as a function of x_k. If f is differentiable, then f does not depend on k if $\partial_{x_k} f(x_k, \mathbf{x}_{-k})$ is 0 for all $\mathbf{x}_{-k}$.

In addition, suppose we have a distribution P over C and the additive approximation

$$f_k^*, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_p,\mu} \left\{ \mathbb{E}\left[\left(f(X) - \sum_{k=1}^{p} f_k(X_k) - \mu \right)^2 \right] : \mathbb{E} f_k(X_k) = 0 \right\}. \tag{3.10}$$

We say that f is *additively faithful* under P if $f_k^* = 0$ implies that f does not depend on coordinate k.

Additive faithfulness is an attractive property because it implies that, in the population setting, the additive approximation yields a consistent variable screening.

3.1. *Additive faithfulness of convex functions.* We now show that under a general class of distributions which we characterize below, convex multivariate functions are additively faithful. To simplify the presentation, we restrict our attention to densities bounded away from $0/\infty$, that is, $0 < \inf p(\mathbf{x}) \leq \sup p(\mathbf{x}) < \infty$.

DEFINITION 3.2. Let $p(\mathbf{x})$ be a density supported on $[0,1]^p$. We say that $p(\mathbf{x})$ satisfies the *boundary flatness condition* if for all j and for all $\mathbf{x}_{-j}$:

(i) the derivatives $\frac{\partial p(\mathbf{x}_{-j}|x_j)}{\partial x_j}$, $\frac{\partial^2 p(\mathbf{x}_{-j}|x_j)}{\partial^2 x_j}$ exist and are bounded, for all $x_j \in [0,\varepsilon) \cup (1-\varepsilon, 1]$ with $\varepsilon > 0$ arbitrarily small,

(ii) for $x_j = 0$ and $x_j = 1$, $\frac{\partial p(\mathbf{x}_{-j}|x_j)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x}_{-j}|x_j)}{\partial x_j^2} = 0$.

The boundary flatness condition intuitively states that two conditional densities $p(\mathbf{x}_{-j} \mid x_j)$ and $p(\mathbf{x}_{-j} \mid x_j')$ are similar when x_j and x_j' are both close to the same boundary point. It is thus much more general than requiring the density to be a product density. Boundary flatness is a weak condition because it affects only an ε-small region around the boundary; $p(\mathbf{x}_{-j} \mid x_j)$ can take arbitrary shapes away from the boundary. Boundary flatness also allows arbitrary correlation structure between the variables [provided $p(\mathbf{x}) > 0$]. In Section 3.2, we give a detailed discussion of the boundary flatness condition and show examples of boundary flat densities. In particular, we show that any density supported on a compact set can be approximated arbitrarily well by boundary flat densities.

The following theorem is the main result of this section.

THEOREM 3.1. *Let $p(\mathbf{x})$ be a density supported on $C = [0,1]^p$ and bounded away from $0/\infty$ that satisfies the boundary flatness property. Suppose f is a convex function with bounded second derivatives on an open set containing C. Then f is additively faithful under $p(\mathbf{x})$.*

We let the domain of f be slightly larger than C for a technical reason—it is so we can say in the proof that the Hessian of f is positive semidefinite even at the boundary of C.

We pause to give some intuition before we present the full proof. Suppose that the underlying density is a product density. We know from Lemma 3.1 that the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero, but "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity. Since the behavior of the whole convex function is constrained by its behavior at the boundary, the same result holds even if the underlying density is not a product density but merely resembles a product density at the boundary, which is exactly the notion formalized by the boundary flatness condition.

PROOF OF THEOREM 3.1. Fixing k and using the result of Lemma 3.1, we need only show that for all x_k, $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \mid x_k] - \mathbb{E}f(X) = 0$ implies that f does not depend on coordinate k, that is, $\partial_{x_k} f(\mathbf{x}) = 0$ for all $\mathbf{x}$.

Let us use the shorthand notation that $r(\mathbf{x}_{-k}) = \sum_{k' \neq k} f_{k'}(x_{k'})$ and assume without loss of generality that $\mu^* = E[f(X)] = 0$. We then assume that, for all x_k,

$$\mathbb{E}\big[f(X) - r(X_{-k}) \mid x_k\big] \equiv \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \mid x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) = 0. \tag{3.11}$$

We let $p'(\mathbf{x}_{-k} \mid x_k)$ denote $\frac{\partial p(\mathbf{x}_{-k}|x_k)}{\partial x_k}$ and $p''(\mathbf{x}_{-k} \mid x_k)$ denote $\frac{\partial^2 p(\mathbf{x}_{-k}|x_k)}{\partial x_k^2}$ and likewise for $f'(x_k, \mathbf{x}_{-k})$ and $f''(x_k, \mathbf{x}_{-k})$.

We differentiate with respect to x_k at $x_k = 0, 1$ under the integral. The details necessary to verify the validity of this operation are technical and given in Section 1.4.1 of the supplementary material.

$$\int_{\mathbf{x}_{-k}} p'(\mathbf{x}_{-k} \mid x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + p(\mathbf{x}_{-k} \mid x_k) f'(x_k, \mathbf{x}_{-k})\, d\mathbf{x}_{-k} = 0, \tag{3.12}$$

$$\begin{aligned}\int_{\mathbf{x}_{-k}} & p''(\mathbf{x}_{-k} \mid x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + 2p'(\mathbf{x}_{-k} \mid x_k) f'(x_k, \mathbf{x}_{-k}) \\ & + p(\mathbf{x}_{-k} \mid x_k) f''(x_k, \mathbf{x}_{-k})\, d\mathbf{x}_{-k} = 0.\end{aligned} \tag{3.13}$$

By the boundary flatness condition, we have that $p''(\mathbf{x}_{-k} \mid x_k)$ and $p'(\mathbf{x}_{-k} \mid x_k)$ are zero at $x_k = x_k^0 \equiv 0$. The integral equations then reduce to the following:

$$\int_{\mathbf{x}_{-k}} p\big(\mathbf{x}_{-k} \mid x_k^0\big) f'\big(x_k^0, \mathbf{x}_{-k}\big)\, d\mathbf{x}_{-k} = 0, \tag{3.14}$$

$$\int_{\mathbf{x}_{-k}} p\big(\mathbf{x}_{-k} \mid x_k^0\big) f''\big(x_k^0, \mathbf{x}_{-k}\big)\, d\mathbf{x}_{-k} = 0. \tag{3.15}$$

Because f is convex, $f(x_k, \mathbf{x}_{-k})$ must be a convex function of x_k for all $\mathbf{x}_{-k}$. Therefore, for all $\mathbf{x}_{-k}$, $f''(x_k^0, \mathbf{x}_{-k}) \geq 0$. Since $p(\mathbf{x}_{-k} \mid x_k^0) > 0$ by the assumption that $p(\mathbf{x})$ is a positive density, we have that $\forall \mathbf{x}_{-k}$, $f''(x_k^0, \mathbf{x}_{-k}) = 0$ necessarily.

The Hessian of f at $(x_k^0, \mathbf{x}_{-k})$ then has a zero at the kth main diagonal entry. A positive semidefinite matrix with a zero on the kth main diagonal entry must have only zeros on the kth row and column; see Proposition 7.1.10 of [12]. Thus, at all $\mathbf{x}_{-k}$, the gradient of $f'(x_k^0, \mathbf{x}_{-k})$ with respect to $\mathbf{x}_{-k}$ must be zero. Therefore, $f'(x_k^0, \mathbf{x}_{-k})$ must be constant for all $\mathbf{x}_{-k}$. By equation (3.14), we conclude that $f'(x_k^0, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$. We can use the same reasoning for the case where $x_k = x_k^1$ and deduce that $f'(x_k^1, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$.

Because $f(x_k, \mathbf{x}_{-k})$ as a function of x_k is convex, it must be that, for all $x_k \in (0, 1)$ and for all $\mathbf{x}_{-k}$,

$$0 = f'\big(x_k^0, \mathbf{x}_{-k}\big) \leq f'(x_k, \mathbf{x}_{-k}) \leq f'\big(x_k^1, \mathbf{x}_{-k}\big) = 0. \tag{3.16}$$

Therefore, f does not depend on x_k. $\square$

Theorem 3.1 plays an important role in our finite sample analysis, where we show that the additive approximation is variable screening consistent, even when the true function is not additive.

REMARK 3.1. We assume twice differentiability in Theorems 3.1 to simplify the proof. We expect, however, that this smoothness condition is not necessary—every convex function can be approximated arbitrarily well by a smooth convex function.

REMARK 3.2. We have not found natural conditions under which the opposite direction of additive faithfulness holds—conditions implying that if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation. Suppose, for example, that f is only a function of X_1, X_2, and that (X_1, X_2, X_3) follows a degenerate three-dimensional distribution where $X_3 = f(X_1, X_2) - f^*(X_1) - f_2^*(X_2)$. In this case X_3 exactly captures the additive approximation error. The best additive approximation of f would have a component $f_3^*(x_3) = x_3$ even though f does not depend on x_3.

REMARK 3.3. In Theorem 3.1, we do not assume a parametric form for the additive components; the additive approximations may not be faithful if we take a parametric form. For example, suppose we approximate a mean-zero convex function $f(X)$ by a linear form $X\beta$. The optimal linear function in the population setting is $\beta^* = \Sigma^{-1}\operatorname{Cov}(X, f(X))$ where $\Sigma = \mathbb{E}X^\top X$ is the covariance matrix. Suppose the X's are independent, centered and have a symmetric distribution with unit variance, and suppose $f(\mathbf{x}) = x_1^2 - \mathbb{E}[X_1^2]$. Then $\beta_1^* = \mathbb{E}[X_1 f(X)] = \mathbb{E}[X_1^3 - X_1\mathbb{E}[X_1^2]] = 0$.

3.2. *Boundary flatness examples*. In this section, we give more examples of boundary flat densities (see Definition 3.2) and discuss extending the notion of boundary flatness to densities with a more general support. We first start with a sufficient condition on the *joint density* that ensures boundary flatness.

EXAMPLE 3.2. Boundary flatness is satisfied if the joint density becomes flat at the boundary. To be precise, let $p(\mathbf{x})$ be a joint density bounded away from $0/\infty$ with a bounded second derivative. Suppose also, for all j,

$$\partial_{x_j} p(x_j, \mathbf{x}_{-j}) = \partial_{x_j}^2 p(x_j, \mathbf{x}_{-j}) = 0 \qquad \text{at } x_j = 0, 1.$$

It is then straightforward to show boundary flatness. One can first verify that the derivatives of the marginal density $p(x_j)$ vanish at $x_j = 0, 1$ and then apply the quotient rule on $\frac{p(x_j, \mathbf{x}_{-j})}{p(x_j)}$ to show that $\partial_{x_j} p(\mathbf{x}_{-j} \mid x_j) = \partial_{x_j}^2 p(\mathbf{x}_{-j} \mid x_j) = 0$ at $x_j = 0, 1$ as well.

The next example shows that any bounded density over a hypercube can be approximated arbitrarily well by boundary flat densities.

EXAMPLE 3.3. Suppose $p_\varepsilon(\mathbf{x})$ is a bounded density over $[\varepsilon, 1-\varepsilon]^p$ for some $0 < \varepsilon < 1/2$. Let $q(\mathbf{x})$ be an arbitrary boundary flat density over $[0,1]^p$ (one can take the uniform density for instance). Define a mixture $p_{\lambda,\varepsilon}(\mathbf{x}) = \lambda q(\mathbf{x}) + (1-\lambda)p_\varepsilon(\mathbf{x})$ where $0 < \lambda \le 1$; then $p_{\lambda,\varepsilon}(\mathbf{x})$ is boundary flat over $[0,1]^p$.

Now, let $p(\mathbf{x})$ be a bounded density over $[0,1]^p$. Let $p_\varepsilon(\mathbf{x})$ be the density formed from truncating $p(\mathbf{x})$ in $[\varepsilon, 1-\varepsilon]^p$ and proper re-normalization. The corresponding mixture $p_{\lambda,\varepsilon}(\mathbf{x})$ then approximates $p(\mathbf{x})$ when λ and ε are both small.

Since $p_{\lambda,\varepsilon}(\mathbf{x})$ remains boundary flat for arbitrarily small ε and λ, $p(\mathbf{x})$ can be approximated arbitrarily well (e.g., in L_1) by boundary flat densities.

In the discussion so far we have restricted our attention to densities supported and positive on the hypercube $[0,1]^p$ to minimize extraneous technical details. It may also be possible to extend the analysis to densities whose support is a convex and compact set so long as the marginal density $p(x_j) > 0$ for all x_j in the support. A rigorous analysis of this, however, is beyond the scope of this paper.

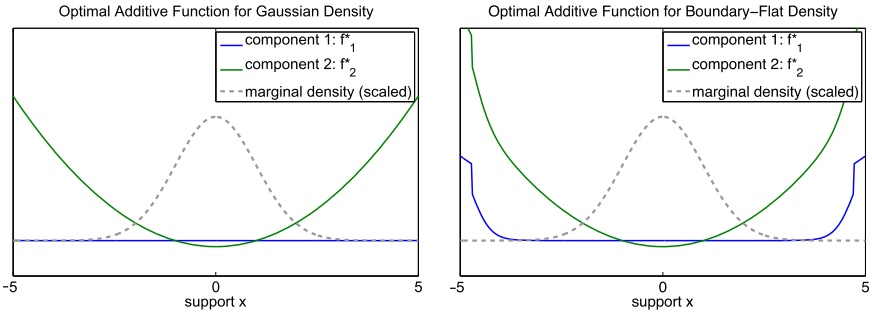
It may also be possible to formulate a similar result to densities with unbounded support, by using a limit condition $\lim_{|x_k|\to\infty} \frac{\partial p(\mathbf{x}_{-k}|x_k)}{\partial x_k} = 0$. Such a limit condition, however, is not obeyed by a correlated multivariate Gaussian distribution. The next example shows that certain convex functions are not additively faithful under general multivariate Gaussian distributions.

EXAMPLE 3.4. Consider a two-dimensional quadratic function $f(\mathbf{x}) = \mathbf{x}^\top H\mathbf{x} + c$ with zero mean where $H = \left(\begin{smallmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{smallmatrix}\right)$ is positive definite and a Gaussian distribution $X \sim N(0, \Sigma)$ where $\Sigma = \left(\begin{smallmatrix} 1 & \alpha \\ \alpha & 1 \end{smallmatrix}\right)$. As we show in Section 2 of the supplementary material [28], the additive approximation has the following closed form:

$$f_1^*(x_1) = \left(\frac{T_1 - T_2\alpha^2}{1-\alpha^4}\right)x_1^2 + c_1,$$

$$f_2^*(x_2) = \left(\frac{T_2 - T_1\alpha^2}{1-\alpha^4}\right)x_2^2 + c_2,$$

where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$, $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$, c_1, c_2 are constants such that f_1^* and f_2^* both have mean zero. Let $H = \left(\begin{smallmatrix} 1.6 & 2 \\ 2 & 5 \end{smallmatrix}\right)$, then it is easy to check that if $\alpha = -\frac{1}{2}$, then $f_1^* = 0$ and additive faithfulness is violated, if $\alpha > \frac{1}{2}$, then f_1^* is a concave function. We take the setting where $\alpha = -0.5$, compute the optimal additive functions via numerical simulation, and show the results in Figure 2(a). Here, f_1^* is zero as expected.



(a) Gaussian distribution (b) Mixture approximation

FIG. 2. *Optimal additive projection of the quadratic function described in Example 3.4 under both the Gaussian distribution described in Example 3.4 and the approximately Gaussian mixture distribution described in Example 3.5. For the mixture approximation, we used* $b = 5, \varepsilon = 0.3, \lambda = 0.0001$ *where the parameters are defined in Example 3.5. This example shows the effect and the importance of the boundary flatness condition.*

Although the Gaussian distribution does not satisfy the boundary flatness condition, it is possible to approximate the Gaussian distribution arbitrarily well with distributions that do satisfy the boundary flatness condition. We use an idea similar to that of Example 3.3.

EXAMPLE 3.5. Let Σ be as in Example 3.4 with $\alpha = -0.5$ so that $f_1^* = 0$. Consider a mixture $\lambda U[-(b+\varepsilon), b+\varepsilon]^2 + (1-\lambda)N_b(0, \Sigma)$ where $N_b(0, \Sigma)$ is the density of a *truncated* bivariate Gaussian bounded in $[-b, b]^2$ and $U[-(b+\varepsilon), b+\varepsilon]^2$ is the uniform distribution over a square. The uniform distribution is supported over a slightly larger square to satisfy the boundary flatness condition.

When b is large, ε is small and λ is small, the mixture closely approximates the Gaussian distribution but is still additively faithful for convex functions. Figure 2(b) shows the optimal additive components under the mixture distribution, computed by numerical integration with $b = 5, \varepsilon = 0.3, \lambda = 0.0001$. True to our theory, f_1^*, which is zero under the Gaussian distribution, is nonzero under the mixture approximation to the Gaussian distribution. We note that the magnitude $\mathbb{E}f_1^*(X_1)^2$, although nonzero, is very small, consistent with the fact that the mixture distribution closely approximates the Gaussian distribution.

3.3. *Convex additive models.* Although convex functions are additively faithful—under appropriate conditions—it is difficult to estimate the optimal additive functions f_k^*s as defined in equation (3.10). The reason is that f_k^* need not be a convex function, as Examples 3.4 and Example 3.5 show. It may be possible to estimate f_k^* via smoothing, but we prefer an approach that is free of smoothing parameters. Since the true regression function f is convex, we approximate the additive model with a *convex* additive model. Without loss of generality, we assume in this section that $\mathbb{E}f(X) = 0$.

We abuse notation and, for the rest of the paper, use the notation f_k^* to represent convex additive fits:

$$\{f_k^*\}_{k=1}^p = \arg\min\left\{\mathbb{E}\left(f(X) - \sum_{k=1}^p f_k(X_k)\right)^2 : f_k \in \mathcal{C}^1, \mathbb{E} f_k(X_k) = 0\right\}, \tag{3.17}$$

where $\mathcal{C}^1$ is the set of univariate convex functions.

If $p(\mathbf{x})$ is a product density, then $\mathbb{E}[f(X) \mid x_k]$ is convex in x_k and the additive projection is simultaneously the convex additive projection. Thus, in this case, additive faithfulness trivially holds for the convex additive projection. For a general boundary flat density $p(\mathbf{x})$, however, the additive projection need not be convex and we thus cannot say anything about additive faithfulness of the convex additive projection.

Luckily, we can restore faithfulness by coupling the f_k^*s with a set of univariate concave fits on the *residual* $f - f^*$:

$$\begin{aligned} g_k^* = \arg\min\Bigg\{&\mathbb{E}\left(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k(X_k)\right)^2 : g_k \in \text{-}\mathcal{C}^1, \\ &\mathbb{E} g_k(X_k) = 0\Bigg\}. \end{aligned} \tag{3.18}$$

THEOREM 3.2. *Suppose $p(\mathbf{x})$ is a density on $C = [0,1]^p$ bounded away from $0/\infty$ that satisfies the boundary flatness condition. Suppose that f is convex with a bounded second derivative on an open set around C. Let f_k^* and g_k^* be as defined in equations (3.17) and (3.18), then the f_k^*'s and the g_k^*'s are unique. Furthermore, $f_k^* = 0$ and $g_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$, that is, f does not depend on x_k.*

Before we can prove the theorem, we need a lemma that generalizes Theorem 3.1.

LEMMA 3.2. *Suppose $p(\mathbf{x})$ is a density on $C = [0,1]^p$ bounded away from 0 and ∞ satisfying the boundary flatness condition. Let $f(\mathbf{x})$ be a convex function with a bounded second derivative on an open set around C. Let $\phi(\mathbf{x}_{-k})$ be a bounded function that does not depend on x_k. Then the unconstrained univariate function*

$$h_k^* = \arg\min_{h_k} \mathbb{E}\big[\big(f(X) - \phi(X_{-k}) - h_k(X_k)\big)^2\big] \tag{3.19}$$

is given by $h_k^*(x_k) = \mathbb{E}[f(X) - \phi(X_{-k}) \mid x_k]$, and $h_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$.

PROOF. In the proof of Theorem 3.1, the only property of $r(\mathbf{x}_{-k})$ we used was the fact that $\partial_{x_k} r(\mathbf{x}_{-k}) = 0$. Therefore, the proof here is identical to that of Theorem 3.1 except that we replace $r(\mathbf{x}_{-k})$ with $\phi(\mathbf{x}_{-k})$. □

PROOF OF THEOREM 3.2. Fix k. Let f_k^* and g_k^* be defined as in equation (3.17) and equation (3.18). Let $\phi(\mathbf{x}_{-k}) \equiv \sum_{k' \neq k} f_{k'}^*(x_{k'})$. Each $f_{k'}^*$ is convex and thus continuous on $(0, 1)$. The function $f_{k'}^*(x_{k'})$ is defined at $x_{k'} = 0, 1$; thus, $f_{k'}^*$ must be bounded and $\phi(\mathbf{x}_{-k})$ is bounded.

We have that

$$f_k^* = \arg\min_{f_k} \Bigg\{ \mathbb{E}\Bigg(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k \Bigg)^2 : f_k \in \mathcal{C}^1, \mathbb{E} f_k(X_k) = 0 \Bigg\}, \tag{3.20}$$

$$g_k^* = \arg\min_{g_k} \Bigg\{ \mathbb{E}\Bigg(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Bigg)^2 : g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Bigg\}. \tag{3.21}$$

Let us suppose that $f_k^* = g_k^* = 0$. It must be then that

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\big(f(X) - \phi(X_{-k}) - c\big(X_k^2 - m_k^2\big)\big)^2 = 0,$$

where $m_k^2 \equiv \mathbb{E} X_k^2$; this is because $c(x_k^2 - m_k^2)$ is either convex or concave in x_k and it is centered, that is, $\mathbb{E}[X_k^2 - m_k^2] = 0$. Since the optimum has a closed form $c^* = \frac{\mathbb{E}[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)]}{(\mathbb{E} X_k^2 - m_k^2)^2}$, we deduce that

$$\begin{aligned} \mathbb{E}\big[\big(f(X) - \phi(X_{-k})\big)\big(X_k^2 - m_k^2\big)\big] &= \mathbb{E}\big[\big(f(X) - \phi(X_{-k})\big) X_k^2\big] \\ &= \mathbb{E}\big[\mathbb{E}\big[f(X) - \phi(X_{-k}) \mid X_k\big] X_k^2\big] = 0. \end{aligned}$$

We denote $h_k^*(x_k) = \mathbb{E}[f(X) - \phi(X_{-k}) \mid x_k]$. Since $f(\mathbf{x})$ and $\phi(\mathbf{x}_{-k})$ are both bounded, $h_k^*(x_k)$ is bounded as well. Therefore, h_k^* is square integrable and there exists a Fourier series $s_n(x_k)$ convergent to h_k^* in L_2. Since $p(\mathbf{x})$ is bounded,

$$\lim_{n \to \infty} \mathbb{E}\big(s_n(X_k) - h_k^*(X_k)\big)^2 \to 0$$

as well.

If we can show that $\mathbb{E} h_k^*(X_k)^2 = 0$, we can apply Lemma 3.2 and complete the proof. So let us suppose for sake of contradiction that $\mathbb{E} h_k^*(X_k)^2 > 0$.

Let $0 < \varepsilon < 1$ be fixed and let n be large enough such that $\mathbb{E}(s_n(X_k) - h_k^*(X_k))^2 \leq \varepsilon \mathbb{E} h_k^*(X_k)^2$. Since $s_n(x_k)$ is twice-differentiable and has a second derivative bounded away from $-\infty$, there exists a positive scalar α such that $s_n(x_k) + \alpha(x_k^2 - m_k^2)$ has a nonnegative second derivative and is thus convex.

Because we assumed $f^* = g^* = 0$, it must be that

$$\arg\min_{c\in\mathbb{R}} \mathbb{E}\big(f(X) - \phi(X_{-k}) - c\big(s_n(X_k) - \mathbb{E}s_n(X_k) + \alpha\big(X_k^2 - m_k^2\big)\big)\big)^2 = 0.$$

This is because $c(s_n(x_k) - \mathbb{E}s_n(X_k) + \alpha(x_k^2 - m_k^2))$ is convex for $c \geq 0$ and concave for $c \leq 0$ and it is a centered function.

Again, $c^* = \frac{\mathbb{E}[(f(X)-\phi(X_{-k}))(s_n(X_k)-\mathbb{E}s_n(X_k)+\alpha(X_k^2-m_k^2))]}{\mathbb{E}(s_n(X_k)-\mathbb{E}s_n(X_k)+\alpha(X_k^2-m_k^2))^2} = 0$, so

$$\begin{aligned}
&\mathbb{E}\big[\big(f(X) - \phi(X_{-k})\big)\big(s_n(X_k) - \mathbb{E}s_n(X_k) + \alpha\big(X_k^2 - m_k^2\big)\big)\big] \\
&\quad = \mathbb{E}\big[\big(f(X) - \phi(X_{-k})\big)s_n(X_k)\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}\big[f(X) - \phi(X_{-k}) \mid X_k\big]s_n(X_k)\big] \\
&\quad = \mathbb{E}h_k^*(X_k)s_n(X_k) = 0,
\end{aligned}$$

where the first equality follows because $\mathbb{E}[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)] = 0$.

We have chosen s_n such that $\mathbb{E}(h_k^*(X_k) - s_n(X_k))^2 \leq \varepsilon \mathbb{E}h_k^*(X_k)^2$ for some $\varepsilon < 1$. But, $\mathbb{E}(h_k^*(X_k) - s_n(X_k))^2 = \mathbb{E}h_k^*(X_k)^2 - 2\mathbb{E}h_k^*(X_k)s_n(X_k) + \mathbb{E}s_n(X_k)^2 \geq \mathbb{E}h_k^*(X_k)^2$. This is a contradiction and therefore, $\mathbb{E}h_k^*(X_k)^2 = 0$.

Now we use Lemma 3.2 with $\phi(\mathbf{x}_{-k}) = \sum_{k'\neq k} f_{k'}^*(x_{k'})$ and conclude that $f_k^* = 0$ and $g_k^* = 0$ together imply that f does not depend on x_k.

Now we turn to uniqueness. Suppose for sake of contradiction that f^* and $\widetilde{f}$ are optimal solutions to (3.17) and $\mathbb{E}(\widetilde{f} - f^*)^2 > 0$. $f^* + \lambda(\widetilde{f} - f^*)$ for any $\lambda \in [0, 1]$ must then also be an optimal solution by convexity of the objective and constraint. However, the second derivative of the objective $\mathbb{E}(f - f^* - \lambda(\widetilde{f} - f^*))^2$ with respect to λ is $2\mathbb{E}(\widetilde{f} - f^*)^2 > 0$. The objective is thus strongly convex and $\mathbb{E}(f^* - \widetilde{f})^2 = 0$. We now apply Lemma 1.3 in the supplement by letting $\phi_k = f_k^* - \widetilde{f}_k$. We conclude that $\mathbb{E}(f_k^* - \widetilde{f}_k)^2 = 0$ for all k. The uniqueness of g^* is proved similarly. □

3.4. *Estimation procedure.* Theorem 3.2 naturally suggests a two-stage screening procedure for variable selection in the population setting. In the first stage, we fit a convex additive model,

$$f_1^*, \ldots, f_p^* = \arg\min_{f_1,\ldots,f_p\in\mathcal{C}_0^1,\mu} \mathbb{E}\left(f(X) - \mu - \sum_{k=1}^p f_k(X_k)\right)^2, \tag{3.22}$$

where we denote $\mathcal{C}_0^1$ ($\text{-}\mathcal{C}_0^1$) as the set of one-dimensional convex (resp. concave) functions with population mean zero. In the second stage, for every variable marked as irrelevant in the first stage, we fit a univariate *concave* function separately on the residual for that variable. For each k such that $f_k^* = 0$,

$$g_k^* = \arg\min_{g_k\in\text{-}\mathcal{C}_0^1} \mathbb{E}\left(f(X) - \mu^* - \sum_{k'} f_{k'}^*(X_{k'}) - g_k(X_k)\right)^2. \tag{3.23}$$

AC/DC ALGORITHM FOR VARIABLE SELECTION IN CONVEX REGRESSION

Input: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, regularization parameter λ.

AC Stage: Estimate a sparse additive convex model:

$$\widehat{f}_1, \ldots, \widehat{f}_p, \widehat{\mu} = \mathop{\arg\min}_{f_1, \ldots, f_p \in \mathcal{C}_0^1} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \right)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty. \tag{3.25}$$

DC Stage: Estimate concave functions for each k such that $\|\widehat{f}_k\|_\infty = 0$:

$$\widehat{g}_k = \mathop{\arg\min}_{g_k \in -\mathcal{C}_0^1} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \right)^2 + \lambda \|g_k\|_\infty. \tag{3.26}$$

Output: Component functions $\{\widehat{f}_k\}$ and relevant variables $\widehat{S}$ where

$$\widehat{S}^c = \{k : \|\widehat{f}_k\| = 0 \text{ and } \|\widehat{g}_k\| = 0\}. \tag{3.27}$$

FIG. 3. *The AC/DC algorithm for variable selection in convex regression. The AC stage fits a sparse additive convex regression model, using a quadratic program that imposes an group sparsity penalty for each component function. The DC stage fits decoupled concave functions on the residuals, for each component that is zeroed out in the AC stage.*

We screen out S^C, any variable k that is zero after the second stage, and output S:

$$S^c = \{k : f_k^* = 0 \text{ and } g_k^* = 0\}. \tag{3.24}$$

We refer to this procedure as AC/DC (additive convex/decoupled concave). Theorem 3.2 guarantees that the true set of relevant variables S_0 must be a subset of S.

It is straightforward to construct a finite sample variable screening procedure, which we describe in Figure 3. We use an ℓ_∞/ℓ_1 penalty in equation (3.25) and an ℓ_∞ penalty in equation (3.23) to encourage sparsity. Other penalties can also produce sparse estimates, such as a penalty on the derivative of each of the component functions. The $\|\cdot\|_\infty$ norm is convenient for both theoretical analysis and implementation.

After selecting the variable set $\widehat{S}$, one can refit a low-dimensional nonadditive convex function to build the best predictive model. If refitting is undesirable for whatever reason, the AC/DC outputs can also be used for prediction. Given a new sample $\mathbf{x}$, we let $\widehat{y} = \sum_k \widehat{f}_k(\mathbf{x}_k) + \sum_k \widehat{g}_k(\mathbf{x}_k)$. Note that $\widehat{g}_k = 0$ for k such that $\widehat{f}_k \neq 0$ in AC/DC. The next section describes how to compute this function evaluation.

The optimization in (3.25) appears to be infinite dimensional, but it is equivalent to a finite dimensional quadratic program. In the following section, we give the details of this optimization, and show how it can be reformulated to be more computationally efficient.

4. Optimization. We now describe in detail the optimization algorithm for the additive convex regression stage. The second decoupled concave regression stage follows a very similar procedure.

Let $\mathbf{x}_i \in \mathbb{R}^p$ be the covariate, let y_i be the response and let ε_i be the mean zero noise. The regression function $f(\cdot)$ we estimate is the sum of univariate functions $f_k(\cdot)$ in each variable dimension and a scalar offset μ. We impose additional constraints that each function $f_k(\cdot)$ is convex, which can be represented by its supporting hyperplanes, that is,

$$f_{i'k} \geq f_{ik} + \beta_{ik}(x_{i'k} - x_{ik}) \qquad \text{for all } i, i' = 1, \ldots, n, \tag{4.1}$$

where $f_{ik} := f_k(x_{ik})$ is the function value and β_{ik} is a subgradient at point x_{ik}. This ostensibly requires $O(n^2 p)$ constraints to impose the supporting hyperplane constraints. In fact, only $O(np)$ constraints suffice, since univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to the optimization

$$\begin{aligned}
&\min_{\{f_k, \beta_k\}, \mu} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{k=1}^{p} f_{ik} \right)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \\
&\text{subject to} \qquad \text{for all } k = 1, \ldots, p: \\
&\qquad f_{\pi_k(i+1)k} = f_{\pi_k(i)k} + \beta_{\pi_k(i)k}(x_{\pi_k(i+1)k} - x_{\pi_k(i)k}), \\
&\qquad \text{for } i = 1, \ldots, n-1 \\
&\qquad \sum_{i=1}^{n} f_{ik} = 0, \\
&\qquad \beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k} \qquad \text{for } i = 1, \ldots, n-2.
\end{aligned} \tag{4.2}$$

Here, f_k denotes the vector $f_k = (f_{1k}, f_{2_k}, \ldots, f_{nk})^T \in \mathbb{R}^n$ and the indices $\{\pi_k(1), \pi_k(2), \ldots, \pi_k(n)\}$ are from the sorted ordering of the values of coordinate k:

$$x_{\pi_k(1)k} \leq x_{\pi_k(2)k} \leq \cdots \leq x_{\pi_k(n)k}. \tag{4.3}$$

We can solve for μ explicitly as $\mu = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$. This follows from the KKT conditions and the constraints $\sum_i f_{ik} = 0$.

The sparse convex additive model optimization in (4.2) is a quadratic program with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for f and β is computationally expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the squared error term $(y_i - \mu - \sum_{k=1}^{p} f_{ik})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{f_k, \beta_k\}$

with the other variables fixed. In matrix notation, the optimization is

$$\begin{aligned} \min_{f_k,\beta_k,\gamma_k} \ & \frac{1}{2n}\|r_k - f_k\|_2^2 + \lambda\gamma_k \\ \text{such that} \quad & P_k f_k = \operatorname{diag}(P_k \mathbf{x}_k)\beta_k, \\ & D_k\beta_k \le 0, \\ & -\gamma_k \mathbf{1}_n \le f_k \le \gamma_k \mathbf{1}_n, \\ & \mathbf{1}_n^\top f_k = 0, \end{aligned} \tag{4.4}$$

where $\beta_k \in \mathbb{R}^{n-1}$ is the vector $\beta_k = (\beta_{1k}, \ldots, \beta_{(n-1)k})^T$, and $r_k \in \mathbb{R}^n$ is the residual vector $r_k = (y_i - \widehat{\mu} - \sum_{k' \neq k} f_{ik'})^T$. In addition, $P_k \in \mathbb{R}^{(n-1)\times n}$ is a permutation matrix where the ith row is all zeros except for the value -1 in position $\pi_k(i)$ and the value 1 in position $\pi_k(i+1)$, and $D_k \in \mathbb{R}^{(n-2)\times(n-1)}$ is another permutation matrix where the ith row is all zeros except for a value 1 in position $\pi_k(i)$ and a value -1 in position $\pi_k(i+1)$. We denote by $\operatorname{diag}(v)$ the diagonal matrix with diagonal entries v. The extra variable γ_k is introduced to impose the regularization penalty involving the ℓ_∞ norm.

This QP subproblem involves $O(n)$ variables, $O(n)$ constraints and a sparse structure, which can be solved efficiently using optimization packages. In our experiments, we use MOSEK (www.mosek.com). We cycle through all covariates k from 1 to p multiple times until convergence. Empirically, we observe that the algorithm converges in only a few cycles. We also implemented an ADMM solver for (4.2) [2], but found that it is not as efficient as this blockwise QP solver.

After optimization, the function estimate for an input vector $\mathbf{x}$ is, according to (4.1),

$$\widehat{f}(\mathbf{x}) = \sum_{k=1}^{p} \widehat{f}_k(x_k) + \widehat{\mu} = \sum_{k=1}^{p} \max_i \{\widehat{f}_{ik} + \widehat{\beta}_{ik}(x_k - x_{ik})\} + \widehat{\mu}. \tag{4.5}$$

The univariate concave function estimation required in the DC stage is a straightforward modification of optimization (4.4). It is only necessary to modify the linear inequality constraints so that the subgradients are nonincreasing: $\beta_{\pi_k(i+1)k} \le \beta_{\pi_k(i)k}$.

4.1. *Alternative formulation.* Optimization (4.2) can be reformulated in terms of the second derivatives. The alternative formulation replaces the order constraints $\beta_{\pi_k(i+1)k} \ge \beta_{\pi_k(i)k}$ with positivity constraints, which simplifies the analysis.

Define $d_{\pi_k(i)k}$ as the second derivative: $d_{\pi_k(1)k} = \beta_{\pi_k(1)k}$, and $d_{\pi_k(i)k} = \beta_{\pi_k(i)k} - \beta_{\pi_k(i-1)k}$ for $i > 1$. The convexity constraint is equivalent to the constraint that $d_{\pi_k(i)k} \ge 0$ for all $i > 1$.

It is easy to verify that $\beta_{\pi_k(i)k} = \sum_{j \le i} d_{\pi_k(j)k}$ and

$$\begin{aligned} f_k(x_{\pi_k(i)k}) &= f_k(x_{\pi_k(i-1)k}) + \beta_{\pi_k(i-1)k}(x_{\pi_k(i)k} - x_{\pi_k(i-1)k}) \\ &= f_k(x_{\pi_k(1)k}) + \sum_{j<i} \beta_{\pi_k(j)k}(x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\ &= f_k(x_{\pi_k(1)k}) + \sum_{j<i} \sum_{j' \le j} d_{\pi_k(j')k}(x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\ &= f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k} \sum_{i>j\ge j'} (x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\ &= f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k}(x_{\pi_k(i)k} - x_{\pi_k(j')k}). \end{aligned}$$

We can write this more compactly in matrix notation as

$$\begin{aligned} \begin{bmatrix} f_k(x_{1k}) \\ f_k(x_{2k}) \\ \vdots \\ f_k(x_{nk}) \end{bmatrix} &= \begin{bmatrix} (x_{1k} - x_{\pi_k(1)k})_+ & \cdots & (x_{1k} - x_{\pi_k(n-1)k})_+ \\ \cdots & & \\ (x_{nk} - x_{\pi_k(1)k})_+ & \cdots & (x_{nk} - x_{\pi_k(n-1)k})_+ \end{bmatrix} \\ &\quad \times \begin{bmatrix} d_{\pi_k(1)k} \\ \cdots \\ d_{\pi_k(n-1)k} \end{bmatrix} + \mu_k \\ &\equiv \Delta_k d_k + \mu_k, \end{aligned} \tag{4.6}$$

where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = (x_{ik} - x_{\pi_k(j)k})_+$, $d_k = (d_{\pi_k(1)k}, \ldots, d_{\pi_k(n-1)k})$, and $\mu_k = f_k(x_{\pi_k(1)k})\mathbf{1}_n$. Because f_k has to be centered, $\mu_k = -\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \Delta_k d_k$ and, therefore,

$$\Delta_k d_k + \mu_k = \Delta_k d_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \Delta_k d_k = \overline{\Delta}_k d_k, \tag{4.7}$$

where $\overline{\Delta}_k \equiv \Delta_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \Delta_k$ is Δ_k with the mean of each column subtracted.

The above derivations prove the following proposition, which states that (4.2) has an alternative formulation.

PROPOSITION 4.1. *Let $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\ldots,p}$ be an optimal solution to (4.2) and suppose $\overline{Y} = 0$. Define vectors $\widehat{d}_k \in \mathbb{R}^{n-1}$ such that $\widehat{d}_{\pi_k(1)k} = \widehat{\beta}_{\pi_k(1)k}$ and $\widehat{d}_{\pi_k(i)k} = \widehat{\beta}_{\pi_k(i)k} - \widehat{\beta}_{\pi_k(i-1)k}$ for $i > 1$. Then $\widehat{f}_k = \overline{\Delta}_k \widehat{d}_k$ and $\widehat{d}_k$ is an optimal solution to the following optimization*:

$$\begin{aligned} &\min_{\{d_k \in \mathbb{R}^{n-1}\}_{k=1,\ldots,p}} \frac{1}{2n}\left\| Y - \sum_{k=1}^p \overline{\Delta}_k d_k \right\|_2^2 + \lambda_n \sum_{k=1}^p \|\overline{\Delta}_k d_k\|_\infty \\ &\quad \textit{such that} \quad d_{\pi_k(2)k}, \ldots, d_{\pi_k(n-1)k} \ge 0 \qquad (\textit{convexity}). \end{aligned} \tag{4.8}$$

Likewise, suppose $\{\widehat{d}_k\}_{k=1,\dots,p}$ *is a solution to* (4.8), *define* $\widehat{\beta}_{\pi_k(i)k} = \sum_{j\le i} \widehat{d}_{\pi_k(j)k}$ *and* $\widehat{f}_k = \overline{\Delta}_k \widehat{d}_k$. *Then* $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\dots,p}$ *is an optimal solution to* (4.2). $\overline{\Delta}$ *is the* n *by* $n-1$ *matrix defined by* (4.7).

The decoupled concave postprocessing stage optimization is again similar. Specifically, suppose $\widehat{d}_k$ is the output of optimization (4.8), and define the residual vector

$$\widehat{r} = Y - \sum_{k=1}^{p} \overline{\Delta}_k \widehat{d}_k. \tag{4.9}$$

Then for all k such that $\widehat{d}_k = 0$, the DC stage optimization is formulated as

$$\begin{aligned} &\min_{c_k} \frac{1}{2n} \|\widehat{r} - \Delta_k c_k\|_2^2 + \lambda_n \|\Delta_k c_k\|_\infty \\ &\text{such that} \qquad c_{\pi_k(2)k}, \dots, c_{\pi_k(n-1)k} \le 0 \qquad \text{(concavity)}. \end{aligned} \tag{4.10}$$

We can use either the off-centered Δ_k matrix or the centered $\overline{\Delta}_k$ matrix because the concave estimations are decoupled, and hence are not subject to nonidentifiability under additive constants.

5. Analysis of variable screening consistency. Our goal is to show variable screening consistency. That is, as $n, p \to \infty$, $\mathbb{P}(\widehat{S} = S)$ approaches 1 where $\widehat{S}$ is the set of variables output by AC/DC in the finite sample setting (Figure 3) and S is the set of variables output in the population setting (3.24).

We divide our analysis into two parts. We first establish a sufficient deterministic condition for consistency of the sparsity pattern screening procedure. We then consider the stochastic setting and argue that the deterministic conditions hold with high probability. Note that in all of our results and analysis, we let c, C represent absolute constants; the actual values of c, C may change from line to line. We derived two equivalent optimizations for AC/DC: (4.2) outputs $\widehat{f}_k, \widehat{g}_k$ and (4.8) outputs the second derivatives $\widehat{d}_k$. Their equivalence is established in Proposition 4.1 and we use both $\widehat{d}_k$ and $\widehat{f}_k$ in our analysis. We will also assume in this section that the true regression function f_0 has mean-zero and, therefore, we will omit the intercept term $\widehat{\mu}$ in our estimation procedure.

In our analysis, we assume that an upper bound B to $\|\widehat{f}_k\|_\infty$ is imposed in the optimization procedure, where B is chosen to also upper bound $\|f_k^*\|_\infty$ (same B as in assumption A3 in Section 5.2). This B-boundedness constraint is added so that we may use the convex function bracketing results from [14] to establish uniform convergence between the empirical risk and the population risk. We emphasize that this constraint is not needed in practice and we do not use it for any of our simulations.

5.1. *Deterministic setting.* We analyze optimization (4.8) and construct an additive convex solution $\{\widehat{d}_k\}_{k=1,\dots,p}$ that is zero for $k \in S^c$, where S is the set of relevant variables, and show that it satisfies the KKT conditions for optimality of optimization (4.8). We define $\widehat{d}_k$ for $k \in S$ to be a solution to the restricted regression (defined below). We also show that $\widehat{c}_k = 0$ satisfies the optimality condition of optimization (4.10) for all $k \in S^c$.

DEFINITION 5.1. We define the *restricted regression* problem

$$\min_{d_k} \frac{1}{n}\Big\| Y - \sum_{k\in S} \overline{\Delta}_k d_k \Big\|_2^2 + \lambda_n \sum_{k\in S} \|\overline{\Delta}_k d_k\|_\infty$$
$$\text{such that} \qquad d_{\pi_k(2)k}, \dots, d_{\pi_k(n-1)k} \geq 0,$$

where we restrict the indices k in optimization (4.8) to lie in some set S which contains the true relevant variables.

THEOREM 5.1 (Deterministic setting). *Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression as defined above. Let $\widehat{r} := Y - \sum_{k\in S} \overline{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

Let $\pi_k(i)$ be a reordering of X_k in ascending order so that $X_{\pi_k(n)k}$ is the largest entry. Let $\mathbf{1}_{\pi_k(i:n)}$ be 1 on the coordinates $\pi_k(i), \pi_k(i+1), \dots, \pi_k(n)$ and 0 elsewhere. Define $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$.

Suppose for all $k \in S^c$, for all $i = 1, \dots, n$, $\lambda_n > \mathsf{range}_k |\frac{32}{n}\widehat{r}^\top \mathbf{1}_{\pi_k(i:n)}|$. Suppose also that for all $k \in S^c$, $\max_{i=1,\dots,n-1} \frac{X_{\pi_k(i+1)k} - X_{\pi_k(i)k}}{\mathsf{range}_k} \leq \frac{1}{16}$, and $\mathsf{range}_k \geq 1$.

Then the following two statements hold:

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$. Then $\{\widehat{d}_k\}_{k=1,\dots,p}$ is an optimal solution to optimization (4.8). Furthermore, any solution to the optimization program (4.8) must be zero on S^c.*
2. *For all $k \in S^c$, the solution $\widehat{c}_k$ to optimization (4.10) must be zero and unique.*

Theorem 5.1 states that the estimator produces no false positive so long as λ_n upper bounds the partial sums of the residual $\widehat{r}$ and that the maximum gap between ordered values of X_k is small.

This result holds regardless of whether or not we impose the boundedness conditions in optimization (4.8) and (4.10). The full proof of Theorem 5.1 is in Section 1.1 of the supplementary material [28]. We allow S in Theorem 5.1 to be any set containing the relevant variables; in Lasso analysis, S is taken to be the set of relevant variables; we will take S to be the set of variables chosen by the additive convex and decoupled concave procedure in the population setting, which is guaranteed to contain the relevant variables because of additive faithfulness.

Theorem 5.1 allows us to separately analyze the false negative rates and false positive rates. To control false positives, Theorem 5.2 verifies that the conditions in

Theorem 5.1 hold in a stochastic setting. To control false negatives, Theorem 5.3 analyzes the restricted regression with only $|S|$ variables.

The proof of Theorem 5.1 analyses the KKT conditions of optimization (4.8). This parallels the now standard *primal-dual witness* technique [27]. The conditions in Theorem 5.1 are analogues of the *mutual incoherence* conditions. Our conditions are much more strict, however, because the estimation is nonparametric—even the low-dimensional restricted regression has $s(n-1)$ variables.

The details of the proof are given in Section 1.1 of the supplementary material [28].

5.2. *Probabilistic setting.* In the probabilistic setting, we treat the covariates as random. We adopt the following standard setup:

1. The data $X^{(1)}, \dots, X^{(n)} \sim P$ are i.i.d. from a distribution P with a density $p(\mathbf{x})$ that is supported on $\mathcal{X} = [-1, 1]^p$.
2. The response is $Y = f_0(X) + W$ where W is independent, zero-mean noise; thus $Y^{(i)} = f_0(X^{(i)}) + W^{(i)}$.
3. The regression function f_0 satisfies $f_0(X) = f_0(X_{S_0})$, where $S_0 = \{1, \dots, s_0\}$ is the set of relevant variables.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-1, 1]$, and let $\mathcal{C}_1^p$ denote the set of convex additive functions $\mathcal{C}_1^p \equiv \{f : f = \sum_{k=1}^p f_k, f_k \in \mathcal{C}^1\}$. Let $f^*(\mathbf{x}) = \sum_{k=1}^p f_k^*(x_k)$ be the population risk minimizer in $\mathcal{C}_1^p$,

$$f^* = \arg\min_{f \in \mathcal{C}_1^p} \mathbb{E}\big(f_0(X) - f(X)\big)^2. \tag{5.1}$$

f^* is the unique minimizer by Theorem 3.2. Similarly, we define -$\mathcal{C}^1$ as the set of univariate concave functions supported on $[-1, 1]$ and define

$$g_k^* = \arg\min_{g_k \in \text{-}\mathcal{C}^1} \mathbb{E}\big(f_0(X) - f^*(X) - g_k(X_k)\big)^2. \tag{5.2}$$

The $\widehat{g}_k$'s are unique minimizers as well. We let $S = \{k = 1, \dots, p : f_k^* \neq 0 \text{ or } g_k^* \neq 0\}$ and let $s = |S|$. By additive faithfulness (Theorem 3.2), it must be that $S_0 \subset S$, and thus $s \geq s_0$. In some cases, such as when $X_{S_0}, X_{S_0^c}$ are independent, we have $S = S_0$. Each of our theorems will use a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.
A2: f_0 is convex with a bounded second derivative on an open set around $[-1, 1]^p$. $\mathbb{E} f_0(X) = 0$.
A3: $\|f_0\|_\infty \leq sB$ and $\|f_k^*\|_\infty \leq B$ for all k.
A4: W is mean-zero sub-Gaussian, independent of X, with scale σ; that is, for all $t \in \mathbb{R}$, $\mathbb{E}e^{tW} \leq e^{\sigma^2 t^2/2}$.
A5: The density $p(\mathbf{x})$ satisfies the boundary flatness condition (Definition 3.2), and $0 < c_l \leq \inf p(\mathbf{x}) \leq \sup p(\mathbf{x}) \leq c_u < \infty$ for two constants c_l, c_u.

By assumption A1, f_k^* must be zero for $k \notin S$. We define α_+, α_- as a measure of the signal strength of the weakest variable:

$$\begin{aligned}\alpha_+ &= \min_{f \in \mathcal{C}_1^p : \operatorname{supp}(f) \subsetneq \operatorname{supp}(f^*)} \{\mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2\}, \\ \alpha_- &= \min_{k \in S : g_k^* \neq 0} \{\mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2\}.\end{aligned} \tag{5.3}$$

The term α_+ is a lower bound on the excess risk incurred by any additive convex function whose support is strictly smaller than f^*. α_+ is achieved by some $f \neq f^*$ because the set $\{f \in \mathcal{C}_1^p : \operatorname{supp}(f) \subsetneq \operatorname{supp}(f^*)\}$ is a finite union of closed convex sets. $\alpha_+ > 0$ since f^* is the unique risk minimizer. Likewise, α_- lower bounds the excess risk of any decoupled concave fit of the residual $f_0 - f^*$ that is strictly more sparse than the optimal decoupled concave fit $\{\widehat{g}_k^*\}$; $\alpha_- > 0$ by the uniqueness of $\{g_k^*\}$ as well. These quantities play the role of the absolute value of the smallest nonzero coefficient in the true linear model in lasso theory. Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure. If $p(\mathbf{x})$ is a product density, then α_+ can be simplified to $\min_{k: f_k^* \neq 0} \mathbb{E} f_k^*(X)^2$ and α_- becomes unnecessary (see Section 3 in the supplementary material).

REMARK 5.1. We make strong assumptions on the covariates in A1 in order to make weak assumptions on the true regression function f_0 in A2. In particular, we do not assume that f_0 is additive. An important direction for future work is to weaken assumption A1. Our simulation experiments indicate that the procedure can be effective even when the relevant and irrelevant variables are correlated.

THEOREM 5.2 (Controlling false positives). *Suppose assumptions* A1–A5 *hold. Define* $\widetilde{\sigma} \equiv \max(\sigma, B)$ *and define* $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$. *Suppose* $p \leq O(\exp(cn))$ *and* $n \geq C$ *for some positive constants* C *and* $0 < c < \frac{c_l}{32}$. *Suppose also*

$$\lambda_n \geq 768 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}}. \tag{5.4}$$

Then with probability at least $1 - \frac{24}{n}$, *for all* $k \in S^c$, *for all* $i' = 1, \ldots, n$,

$$\lambda_n \geq \mathsf{range}_k \left| \frac{32}{n} \widehat{r}^\top \mathbf{1}_{(i':n)_k} \right|, \tag{5.5}$$

$\max_{i'} \frac{X_{\pi_k(i'+1)k} - X_{\pi_k(i')k}}{\mathsf{range}_k} \leq \frac{1}{16}$, $\mathsf{range}_k \geq 1$, *and both the AC solution* $\widehat{f}_k$ *from optimization* (4.8) *and the DC solution* $\widehat{g}_k$ *from optimization* (4.10) *are zero.*

Here, we use $\mathbf{1}_{(i':n)}$ to denote a vector that is 1 on the i'th to the nth coordinates and 0 elsewhere.

The proof of Theorem 5.2 exploits independence of $\widehat{r}$ and X_k under assumption A1; when $\widehat{r}$ and X_k are independent, $\widehat{r}^\top \mathbf{1}_{(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We can then use concentration of measure results for sampling without replacement to argue that $|\frac{1}{n}\widehat{r}^\top \mathbf{1}_{(i':n)}|$ is small with high probability. The result of Theorem 5.1 is then used. The full proof of Theorem 5.2 is in Section 1.2 of the supplementary material [28].

THEOREM 5.3 (Controlling false negatives). *Suppose assumptions* A1–A5 *hold. Let* $\widehat{f}$ *be any AC solution to the restricted regression with* B*-boundedness constraint, and let* $\widehat{g}_k$ *be any DC solution to the restricted regression with* B*-boundedness constraint. Let* $\widetilde{\sigma}$ *denote* $\max(\sigma, B)$. *Suppose*

$$\lambda_n \leq 768 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}} \tag{5.6}$$

and that n *is sufficiently large so that, for some constant* $c' > 1$,

$$\frac{n^{4/5}}{\log np} \geq c' B^4 \widetilde{\sigma}^2 s^5. \tag{5.7}$$

Assume that the signal-to-noise ratio satisfies

$$\frac{\alpha_+}{\widetilde{\sigma}} \geq c B^2 \sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}} \log^2 np}, \tag{5.8}$$

$$\frac{\alpha_-^2}{\widetilde{\sigma}} \geq c B^2 \sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}} \log^2 np}, \tag{5.9}$$

where c *is a constant. Then with probability at least* $1 - \frac{C}{n}$ *for some constant* C, $\widehat{f}_k \neq 0$ *or* $\widehat{g}_k \neq 0$ *for all* $k \in S$.

This is a finite sample version of Theorem 3.1. We need stronger assumptions in Theorem 5.3 to use our additive faithfulness result, Theorem 3.1. The full proof of Theorem 5.3 is in Section 1.3 of the supplement [28].

Combining Theorems 5.2 and 5.3, we obtain the following result.

COROLLARY 5.1. *Suppose the assumptions of Theorem* 5.2 *and Theorem* 5.3 *hold. Then with probability at least* $1 - \frac{C}{n}$

$$\widehat{f}_k \neq 0 \quad \text{or} \quad \widehat{g}_k \neq 0 \qquad \text{for all } k \in S, \tag{5.10}$$

$$\widehat{f}_k = 0 \quad \text{and} \quad \widehat{g}_k = 0 \qquad \text{for all } k \notin S \tag{5.11}$$

for some constant C.

The above corollary implies that consistent variable selection is achievable with an exponential scaling of the ambient dimension scaling, $p = O(\exp(cn))$ for some $0 < c < 1$, just as in parametric models. The cost of nonparametric modeling through shape constraints is reflected in the scaling with respect to the number of relevant variables, which can scale as $s = o(n^{4/25})$.

REMARK 5.2. Comminges and Dalalyan [6] have shown that under traditional smoothness constraints, even with a product distribution, variable selection is achievable only if $n > O(e^{s_0})$. It is interesting to observe that because of additive faithfulness, the convexity assumption enables a much better scaling of $n = O(\text{poly}(s_0))$, demonstrating that geometric constraints can be quite different from the previously studied smoothness conditions.

6. Experiments. We perform both synthetic and real data experiments.

6.1. *Simulations*. We first illustrate our methods using a simulation of the model

$$Y_i = f_0(x_{iS}) + w_i \qquad (i = 1, 2, \ldots, n).$$

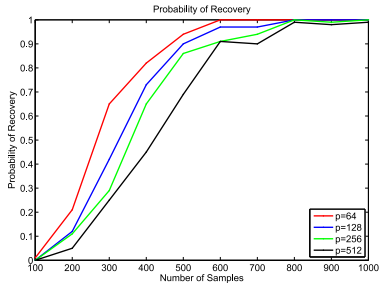
Here, x_i denotes data sample i drawn from some distribution P, and f_0 is the true regression function. The variables x_{iS} are a subset of x_i with dimension $|S| = s$, where S represents the set of relevant variables, and w_i is additive noise drawn from $\mathcal{N}(0, \sigma^2)$. For all simulations, we set σ^2 so that the signal-to-noise ratio [SNR, $\frac{\text{std}(Y)}{\sigma}$] is 5. Also, for all simulations except the sixth, we choose the set of relevant variables S uniformly at random among all variables $\{1, \ldots, p\}$.

We study both the independent case where $P = N(0, I_p)$ and the correlated case where P is a correlated Gaussian copula modified slightly to satisfy the boundary flatness condition. We measure the probability of exact selection in the independent case and the probability of screening in the correlated case. We also study both cases where the regression function is parametric (quadratic) and cases where the regression function is nonparametric (softmax of linear forms). In all our experiments, we mark a variable as selected if either the AC estimate $\|\widehat{f}_j\|_\infty$ or the DC estimate $\|\widehat{g}_k\|_\infty$ is larger than 10^{-6}. We set $\lambda = 0.5\sqrt{\frac{\log^2 np}{n}}$ for all the simulations.

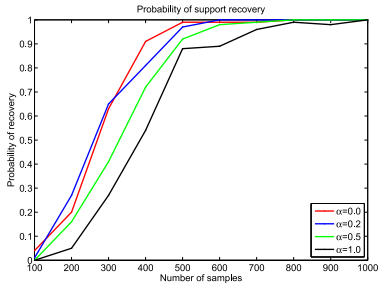
For the first three simulations, we use a quadratic form as the true regression function,

$$f_0(x_{iS}) = x_{iS}^\top Q x_{iS}.$$

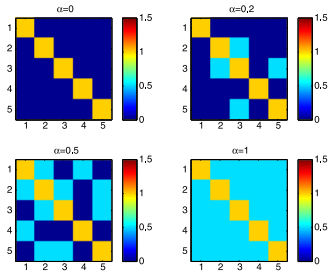
The matrix Q is a symmetric positive definite matrix of dimension $s \times s$. Note that if Q is diagonal, then the true function is convex additive; otherwise the true function is convex but not additive.



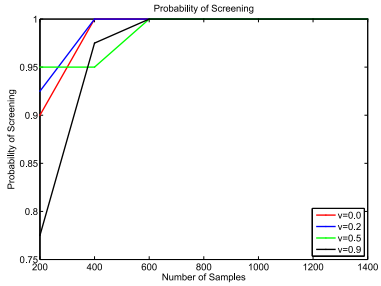
(a) quadratic f_0, independent X, varying p



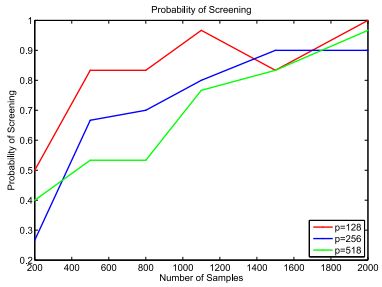
(b) quadratic f_0, independent X, varying Q



(c) four Q matrices used in (b)



(d) quadratic f_0, correlated X, varying correlation ν



(e) softmax f_0, correlated X, varying p

FIG. 4. *Support recovery results.*

6.1.1. *First simulation.* In the first simulation [Figure 4(a)], we vary the ambient dimension p. We set Q to be one on the diagonal and $1/2$ on the off-diagonal with 0.5 probability, set $s = 5$, and $p = 64$, 128, 256 and 512. We draw $X \sim N(0, I_p)$. For each (n, p) combination, we generate 100 independent trials. In Figure 4(a), we plot the probability of exact support recovery. We observe that the algorithm performs consistent variable selection even if the dimensionality is large. To give the reader a sense of the running speed, for a dataset with $n = 1000$ and $p = 512$, the code runs in roughly two minutes on a machine with a 2.3 GHz Intel Core i5 CPU and 4 GB memory.

6.1.2. *Second simulation.* In the second simulation [Figure 4(b), (c)], we vary the sparsity of the Q matrix, thus varying the extent to which the true function is

nonadditive. We generate four Q matrices plotted in Figure 4(c), where the diagonal elements are all one and the off-diagonal elements are $\frac{1}{2}$ with probability α ($\alpha = 0, 0.2, 0.5, 1$ for the four cases). We show the four Q matrices we used in Figure 4(c). We fix $p = 128$, $s = 5$, and $X \sim N(0, I_p)$. We again run the AC/DC optimization on 100 independent trials and plot the probability of exact recovery in Figure 4(b). The results demonstrate that AC/DC performs consistent variable selection even if the true function is not additive (but still convex).

In the third, fourth and fifth simulation, we use a correlated design. We generate X from a non-Gaussian boundary flat distribution with covariance Σ. The distribution we used is a mixture of a uniform distribution and a Gaussian copula,

$$X \sim \gamma U\big([-2, 2]^p\big) + (1 - \gamma)\,\text{Copula}(0, \Sigma, F).$$

The Gaussian copula is a way to customize the marginal distributions of a Gaussian random variable while maintaining the same covariance. Gaussian copula results when one applies a monotone transformation $F^{-1}\Phi$ onto each of the variables of a Gaussian random vector where Φ is the normal CDF and F is the CDF of the new marginal distribution. In all our experiments, we set $\gamma = 0.05$ and set the marginal CDF F so that marginal density of the copula is bimodal and supported on $[-1.8, 1.8]$. The resulting marginal density of the mixture is shown in Figure 5. Notice that boundary flatness holds because the distribution is uniform in the boundary area $[-2, 2]^p \setminus [-1.8, 1.8]^p$.

6.1.3. *Third simulation.* In the third simulation [Figure 4(d), Figure 6(a)], we use the non-Gaussian distribution described above and set the covariance to be $\Sigma_{ij} = \nu^{|i-j|}$ for $\nu = 0, 0.2, 0.5, 0.9$. We use the nonadditive Q, same as in the first experiment, with $\alpha = 0.5$ and fix $p = 128, s = 5$. In Figure 4(d), we say that a trial is successful if all relevant variables were recovered, that is, there are no false negatives. In Figure 6(a), we also show the total number of selected variables versus the sample size as boxplots; since the true sparsity level is 5, these

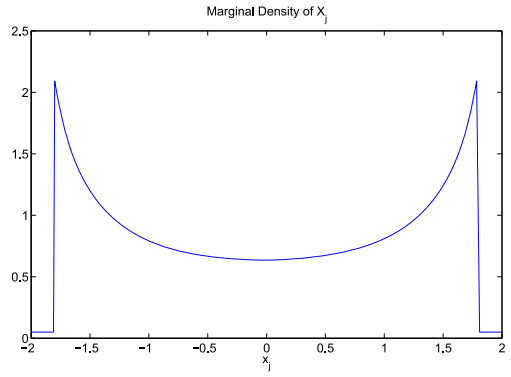
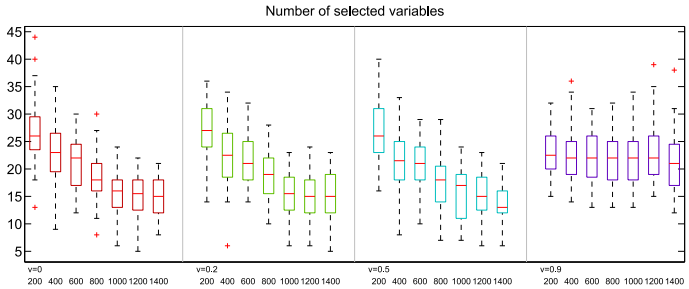
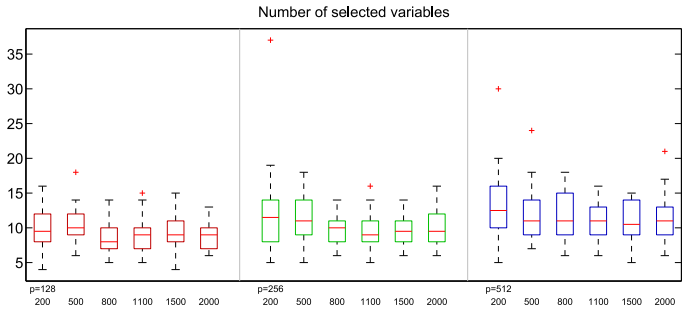


FIG. 5. *Marginal density of the Gaussian copula and uniform mixture.*



(a) recovered support sizes for figure 5(d)



(b) recovered support sizes for figure 5(e)

FIG. 6. *Recovered support size in correlated design cases.*

plots show the number of false positives. We use the same λ as before. The probabilities of success are computed from 40 independent trials and plotted against various values of ν in Figure 4(d). As seen, for all correlation levels, AC/DC can successfully recover the relevant variables with only a small number of false positives. For $\nu = 0, 0.2, 0.5$, the number of false positives steadily decrease with the sample size, but for $\nu = 0.9$, the number of false positives stays roughly constant. The latter phenomenon does not contradict our theory, which assumes $X_S \perp X_{S^c}$ for some set $S \supset S_0$, but it does demonstrate a weakness of our method when the design is highly correlated.

In the fourth and fifth simulation, we use a softmax function as the ground truth,

$$f_0(x_{iS}) = \log\left(\sum_{k=1}^{K} \exp(\beta_k^\top x_{iS})\right) - \mu. \tag{6.1}$$

We generate K random unit vectors $\{\beta_k \in \mathbb{R}^s\}_{k=1,\dots,K}$ and choose μ so that f_0 has mean-zero. We set $K = 7$ for all the experiments.

6.1.4. *Fourth simulation.* For the fourth simulation [Figure 4(e), Figure 6(b)], we let f_0 be the softmax function and let X be drawn from the boundary flat mixture distribution described earlier with the Toeplitz covariance $\Sigma_{ij} = \nu^{|i-j|}$ for

$\nu = 0.5$. We set $s = 5$ and vary $p = 128,256,512$. We use the same faithful recovery criterion as in the third simulation and plot the probability of faithful recovery against the number of samples in Figure 4(e). The probabilities are computed over 30 independent trials. Also, we show the number of selected variables versus the sample size as boxplots in Figure 6(b). The softmax function is more challenging to estimate than the quadratic function, regardless, we see that increasing the ambient dimension p does not significantly affect the recovery probability. We also see from Figure 6(b) that larger p does not lead to more number of false positives.

6.1.5. *Fifth simulation.* For the fifth simulation (Figure 7), we compare the variables selected via the AC stage and the variables selected via the DC stage. We use the softmax regression function and X drawn from the boundary flat mixture distribution with a Toeplitz covariance and correlation level $\nu = -0.7$. We set $s = 5$, $n = 500$ and $p = 128$. We perform 30 independent trials and plot the frequency of variable selection in Figure 7. The true variables are X_j for $j = 5, 6, 7, 8, 9, 10$. We plot the frequency of selection among only the first 20 variables, that is, X_j for $j = 1, \ldots, 20$. We do not plot selection frequencies for variables 21 to 128 because they are almost never selected by either the AC or DC stage. As can be seen, the DC stage is slightly helpful in recovering the true variables but its effect is not significant. We thus believe that the DC stage, though important in theory, is not as important in practice; it may be omitted without significant detriment to the overall result.

6.2. *Boston housing data.* We next use the Boston housing data rather than simulated data. This data set contains 13 covariates, 506 samples and one response

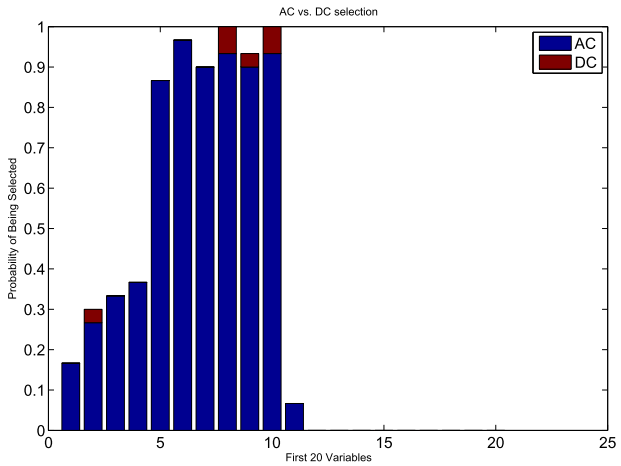
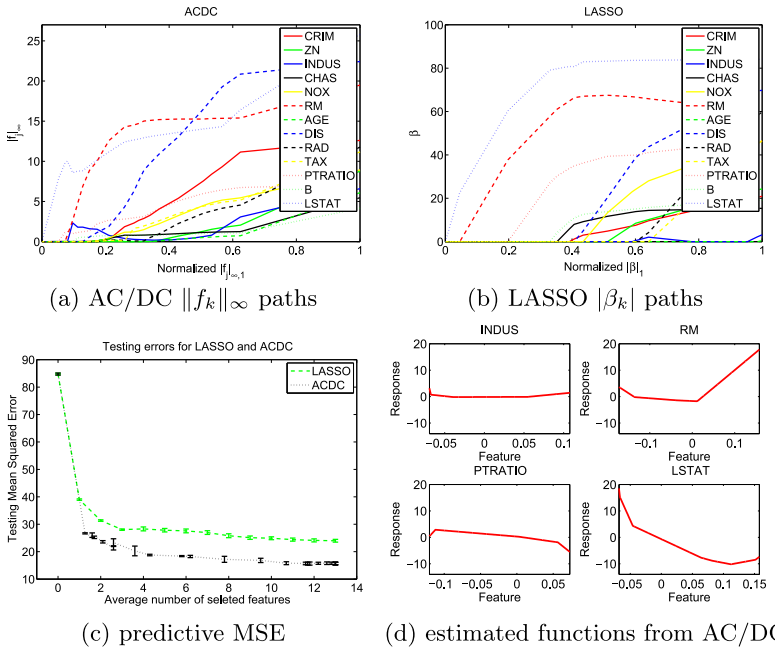


FIG. 7. *Frequency of variable selection among the first* 20 *variables* (X_j *for* $j = 1, \ldots, 20$) *in the AC stage vs. in the DC stage. The true variables are* [5, 6, 7, 8, 9, 10].

variable indicating housing values in suburbs of Boston. The data and a detailed description can be found on the UCI Machine Learning Repository website.[2]

We first use all $n = 506$ samples (with standardization) in the AC/DC algorithm, using a set of candidate regularization parameters $\{\lambda^{(t)}\}$ ranging from $\lambda^{(1)} = 0$ (no regularization) to 2. For each $\lambda^{(t)}$, we obtain a function value matrix $f^{(t)}$ with $p = 13$ columns and $n = 506$ rows. The nonzero columns in this matrix indicate the variables selected using $\lambda^{(t)}$.

In Figure 8(a), we plot on the y-axis the norm $\|f_j^{(t)}\|_\infty$ of every column j against the regularization strength $\lambda^{(t)}$. Instead of plotting the value of $\lambda^{(t)}$ on the x-axis, however, we plot the total norm at $\lambda^{(t)}$ normalized against the total norm at $\lambda^{(1)}$: $\frac{\sum_j \|f_j^{(t)}\|_\infty}{\sum_j \|f^{(1)}\|_\infty}$. Thus, as x moves from 0 to 1, the regularization goes from strong to weak. For comparison, we plot the LASSO/LARS result in a similar way in Figure 8(b). From the figures, we observe that the first three variables selected by AC/DC and LASSO are the same: `LSTAT`, `RM` and `PTRATIO`, consistent with previous findings [25]. The fourth variable selected by AC/DC is `INDUS` (with $\lambda^{(t)} = 0.7$). We then refit AC/DC with only these four variables without regularization, and plot the estimated additive functions in Figure 8(d). When refitting, we



(a) AC/DC $\|f_k\|_\infty$ paths (b) LASSO $|\beta_k|$ paths

(c) predictive MSE (d) estimated functions from AC/DC

FIG. 8. *Results on Boston housing data showing regularization paths, MSE and fitted functions.*

[2] http://archive.ics.uci.edu/ml/datasets/Housing.

constrain a component to be convex if it is nonzero in the AC stage and concave if it is nonzero in the DC stage. As can be seen, these functions contain clear nonlinear effects which cannot be captured by LASSO. The shapes of these functions, including the concave shape of the PTRATIO function, are in agreement with those obtained by SpAM [25].

Next, in order to quantitatively study the predictive performance, we run 5-fold cross validation three times, following the same procedure described above—training, variable selection and refitting. A plot of the mean and standard deviation of the predictive mean squared error (MSE) is shown in Figure 8(c). Since for AC/DC the same regularization level $\lambda^{(t)}$ may lead to a slightly different number of selected variables in different folds and runs, the values on the x-axis for AC/DC are not necessarily integers. The figure clearly shows that AC/DC has a lower predictive MSE than LASSO. We also compared the performance of AC/DC with that of Additive Forward Regression (AFR) presented in [19], and found that they are similar. The main advantages of AC/DC compared with AFR and SpAM are that there are no smoothing parameters required, and the optimization is formulated as a convex program, guaranteeing a global optimum.

7. Discussion. We have introduced a framework for estimating high dimensional but sparse convex functions. Because of the special properties of convexity, variable selection for convex functions enjoys additive faithfulness—it suffices to carry out variable selection over an additive model, in spite of the approximation error this introduces. Sparse convex additive models can be optimized using block coordinate quadratic programming, which we have found to be effective and scalable. We established variable selection consistency results, allowing exponential scaling in the ambient dimension. We expect that the technical assumptions we have used in these analyses can be weakened; this is one direction for future work. Another interesting direction for building on this work is to allow for additive models that are a combination of convex and concave components. If the convexity/concavity of each component function is known, this again yields a convex program. The challenge is to develop a method to automatically detect the concavity or convexity pattern of the variables.

Acknowledgments. The authors thank Ming Yuan for valuable discussions and suggestions on this research. The authors are also grateful to the anonymous reviewers whose careful proofreading and helpful comments significantly improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Faithful variable screening for high-dimensional convex regression" (DOI: 10.1214/15-AOS1425SUPP; .pdf). The supplement provides detailed proofs of certain technical results, together with further explanation of the Gaussian example and simplifications when the density is a product.

REFERENCES

[1] BERTIN, K. and LECUÉ, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.* **2** 1224–1241. MR2461900

[2] BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

[3] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

[4] CHEN, H. and YAO, D. D. (2001). *Fundamentals of Queueing Networks*: *Performance*, *Asymptotics*, *and Optimization*. *Applications of Mathematics* (*New York*) **46**. Springer, New York. MR1835969

[5] CHEN, Y. and SAMWORTH, R. J. (2014). Generalised additive and index models with shape constraints. Preprint. Available at arXiv:1404.2957.

[6] COMMINGES, L. and DALALYAN, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616

[7] CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237

[8] DEVORE, R., PETROVA, G. and WOJTASZCZYK, P. (2011). Approximation of functions of few variables in high dimensions. *Constr. Approx.* **33** 125–143. MR2747059

[9] GOLDENSHLUGER, A. and ZEEVI, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist.* **34** 1375–1394. MR2278361

[10] GUNTUBOYINA, A. and SEN, B. (2013). Global risk bounds and adaptation in univariate convex regression. Preprint. Available at arXiv:1305.1648.

[11] HANNAH, L. A. and DUNSON, D. B. (2012). Ensemble methods for convex regression with applications to geometric programming based circuit design. In *International Conference on Machine Learning* (*ICML*). Edinburgh.

[12] HORN, R. A. and JOHNSON, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

[13] HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

[14] KIM, A. K. and SAMWORTH, R. J. (2014). Global rates of convergence in log-concave density estimation. Preprint. Available at arXiv:1404.2298.

[15] KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695. MR2766864

[16] LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963

[17] LELE, A. S., KULKARNI, S. R. and WILLSKY, A. S. (1992). Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America*, *Series A* **9** 1693–1714.

[18] LIM, E. and GLYNN, P. W. (2012). Consistency of multidimensional convex regression. *Oper. Res.* **60** 196–208. MR2911667

[19] LIU, H. and CHEN, X. (2009). Nonparametric greedy algorithm for the sparse learning problems. In *Advances in Neural Information Processing Systems* **22**.

[20] MEYER, R. F. and PRATT, J. W. (1968). The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics* **4** 270–278.

[21] MOSSEL, E., O'DONNELL, R. and SERVEDIO, R. A. (2004). Learning functions of k relevant variables. *J. Comput. System Sci.* **69** 421–434. MR2087943

[22] PRINCE, J. L. and WILLSKY, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 377–389.

[23] PYA, N. and WOOD, S. N. (2015). Shape constrained additive models. *Stat. Comput.* **25** 543–559.

[24] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

[25] RAVIKUMAR, P., LIU, H., LAFFERTY, J. and WASSERMAN, L. (2007). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems* **20**.

[26] SEIJO, E. and SEN, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* **39** 1633–1657. MR2850215

[27] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[28] XU, M., CHEN, M. and LAFFERTY, J. (2016). Supplement to "Faithful variable screening for high-dimensional convex regression." DOI:10.1214/15-AOS1425SUPP.

M. XU
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: minx@wharton.upenn.edu

M. CHEN
AMAZON.COM
410 TERRY AVENUE NORTH
SEATTLE, WASHINGTON 98109
USA

J. LAFFERTY
DEPARTMENT OF STATISTICS
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CHICAGO
5734 SOUTH UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: lafferty@galton.uchicago.edu