Behaviour of the Monotone Single Index Model Under Repeated Measurements

Fadoua Balabdaoui
ETH Zürich, Zürich, Switzerland

Cécile Durot
Université Paris Nanterre, Nanterre, France

Hanna Jankowski
York University, Toronto, Canada

Abstract

The generalized linear model is an important method in the statistical toolkit. The isotonic single index model can be thought of as a further generalization whereby the link function is assumed to be monotone non-decreasing as opposed to known and fixed. Such a shape constraint is quite natural in many statistical problems, and is fulfilled by the usual generalized linear models. In this paper we consider inference in this model in the setting where repeated measurements of predictor values and associated responses are observed. This setting is encountered in medical studies and is very different from the one considered in the classical monotone single index model studied in the literature. Here, we use nonparametric maximum likelihood estimation to infer the unknown regression vector and link function. We present a detailed study of finite and asymptotic properties of this estimator and propose goodness-of-fit tests for the model. Through an extended simulation study, we show that the model has competitive predictive performance. We illustrate our estimation approach using a Leukemia data set.

AMS (2000) subject classification. Primary 62G08; Secondary 62P10.

Keywords and phrases. Goodness-of-fit, Isotonic regression, Maximum likelihood, Shape-constrained estimation, Single index model, Set identifiability.

1 Introduction

The class of generalized linear models was introduced as a unified, and hence more general, framework which encompasses many popular tools such

Supplementary Information The online version of this article (<https://doi.org/10.1007/s13171-021-00250-7>) contains supplementary material, which is available to authorized users.

as linear and logistic regression (Nelder and Wedderburn, 1972). Both these examples, which are widely used in practice, involve an increasing link function: the identity and logit function respectively. In fact, link functions in the usual generalized linear models are increasing. Thus, these generalized linear models can be themselves viewed as a sub-family of the even more general class of isotonic single index models.

Furthermore, isotonic single index models offer flexibility and robustness against picking the wrong model. As rightly pointed out in Härdle et al. (1997, page 213) “the advantage of [the single index] approach is that an interpretable linear single index, a weighted sum of the predictor variables, is still produced.”

Estimation in the isotonic single index model has recently received increasing attention, see for example Kakade et al. (2011), Chen and Samworth (2016), Foster et al. (2013) for the high-dimensional setting, and the recent work of Balabdaoui et al. (2019). The model is closely related to the semi-parametric single index model which was much popularized in econometrics (Li and Racine, 2007). Unlike the isotonic single index model, no shape assumptions on the link function are made. However, it is necessary to impose some smoothness conditions on the unknown link function, which is typically estimated via kernel-based nonparametric methods. The most popular estimation methods are due to Ichimura (1993) and Klein and Spady (1993), which are implemented in the R package `np`. Despite their popularity, one of the main downsides of such approaches is the fact that they require bandwidth selection, known to be a hard problem to solve. For a review of existing estimation algorithms we refer the reader to Li and Racine (Li and Racine, Chapter 8). Unlike kernel-based methods, isotonic estimation is fully automatic and no tuning parameter needs to be selected to achieve optimal performance. In addition, Foster et al. (2013) argue that, whenever appropriate, the assumption of monotonicity “noticeably improves performance”. The advantages of the monotone single index model have also been investigated in Qin et al. (2019), Wan et al. (2017) where applications to, respectively, causal inference and drug interaction modelling were considered. Qin et al. (2019) considered a similar estimator as we do here assuming continuous design, while Wan et al. (2017) studied a spline approach, which has the drawback of having to use cross-validation in order to select optimal knot positions.

To the best of our knowledge, theoretical developments for the isotonic single index model are restricted to the setting of random and continuous

design where it is assumed that the covariates are generated from a multivariate density. In the recent work of Balabdaoui et al. (2019) it is shown that the rate of convergence of the least squares estimator (LSE) of the isotonic link function in this setting is $n^{1/3}$ under some standard regularity assumptions. The methods used there cannot be adapted to datasets where repeated measurements of predictor values are observed. Such discrete design settings occur for example in drug-interaction studies, cf. Carter et al. (1986), Wan et al. (2017). We note that although Wan et al. (2017) considered an estimation method which is not fully automatic like the one presented here, they stressed the benefit to be gained from fitting a monotone single index model. Hence, it is crucial to understand the behaviour of this estimator as well as any potential technical challenges associated with the setting with repeated predictors.

Our main contributions in this paper are to give a precise description of the asymptotic behavior of the maximum likelihood estimator (MLE) in the monotone single index model under discrete design and develop an asymptotic goodness-of-fit test which could be used before applying our approach. Similarly to Qin et al. (2019), Chen and Samworth (2016), we study the case where the sample size n is allowed to grow while the dimension d of the predictors is held fixed. One of our main findings is that, as opposed to the classical monotone single index models where the nonparametric LSE or MLE of the link function converges at the $n^{1/3}$ -rate, the asymptotics of the MLE under the discrete design exhibit the faster rate of $n^{1/2}$. We also investigate the finite sample performance of the MLE using synthetic data, where the MLE is found to have competitive predictive performance when compared to the MLE in the true generalized linear model from which the data are drawn.

The manuscript is organized as follows. In Section 2 we give a description of the isotonic single index model under discrete design and prove some of the fundamental properties of the MLE such as existence and lack of uniqueness of the estimator. In Section 3 we prove consistency of the MLE of the link function and give sufficient conditions under which it is asymptotically normal. Furthermore, we propose two goodness-of-fit tests. In Section 4 we illustrate through simulations the performance of the proposed method. In Section 5, following Wan et al. (2017), we give an example involving drug interactions in a Leukemia dataset, where we compare the monotone single model approach with kernel-based nonparametric single-index and

fully parametric methods. To keep the reading more fluid, we defer all proofs and technical details to the [Supplementary Material](#).

2 The Model and Maximum Likelihood Estimation

2.1. The Model and Corresponding Log-Likelihood Let n be the total size of a sample of observed responses organized into m groups. Observations in the i -th group for $i = 1, \dots, m$ are assumed to have expectation of the form

$$\mu_{0i} = f_0(\alpha_0^T x_i) \quad (2.1)$$

where $x_i \in \mathbb{R}^d$ is a design point, $\alpha_0 \in \mathbb{R}^d \setminus \{0\}$ is an unknown regression parameter, also called the “index”, and f_0 is an unknown increasing link function. We use the term increasing to mean non-decreasing. All values of $x_i, i = 1, \dots, m$ are assumed to be distinct. We assume that the density function of the observed response in the i -th group, with respect to a given dominating measure λ which is either Lebesgue measure or counting measure, is given by

$$c(y, \phi) \exp \left\{ \frac{y \ell(\mu_{0i}) - B \circ \ell(\mu_{0i})}{\phi} \right\}, \quad (2.2)$$

where c is a normalizing function, $\phi > 0$ is the dispersion parameter, and ℓ is a given real valued function with first derivative $\ell' > 0$, and inverse $\ell^{-1} = B'$.

Balabdaoui et al. (2019, Proposition 5.1) study identifiability of the model when the distribution of the predictors has a density with respect to Lebesgue measure. The situation considered here is very different in that we observe $f_0(\alpha_0^T x)$ with noise only at the discrete design points $x = x_1, \dots, x_m$. Hence, the model is not necessarily identifiable in the classical sense. However, we can show that the index parameter satisfies a weaker form of identifiability. Let $\mathcal{S}^{d-1}$ denote the unit sphere in d -dimensional Euclidean space.

PROPOSITION 1. *Assume that there exist an increasing function f_0 and $\alpha_0 \in \mathcal{S}^{d-1}$ such that observations in the i -th group have expectation μ_{0i} of the form (2.1). Define the set*

$$\mathcal{S}_0 = \left\{ \alpha \in \mathcal{S}^{d-1} : \text{for all } i, j, \quad \alpha^T x_i \leq \alpha^T x_j \quad \text{implies} \quad \mu_{0i} \leq \mu_{0j} \right\}. \quad (2.3)$$

Consider any increasing function f and index $\alpha \in \mathcal{S}^{d-1}$ such that $\mu_{0i} = f(\alpha^T x_i), i = 1, \dots, m$. Then, the index α is set identifiable in the sense that $\alpha \in \mathcal{S}_0$.

Heuristically, the set $\mathcal{S}_0$ identifies all α which, via the projected predictors $\alpha^T x_i$, preserve the correct order in the components of μ_0 . This can be rephrased as “ $\mathcal{S}_0$ is the subset of $\mathcal{S}^{d-1}$ such that $\alpha^T x_i < \alpha^T x_j$ whenever $\mu_{0i} < \mu_{0j}$, for $i, j = 1, \dots, m$ ”. Hence, if μ_0 is constant, then $\mathcal{S}_0 = \mathcal{S}^{d-1}$. At the other extreme, if μ_0 satisfies $\mu_{01} < \dots < \mu_{0m}$, then $\mathcal{S}_0$ contains all $\alpha \in \mathcal{S}^{d-1}$ such that $\alpha^T x_1 < \dots < \alpha^T x_m$.

Assume that we have n_i independent observations Y_{ij} ($j = 1, \dots, n_i$) from group i ($i = 1, \dots, m$) with density as in Eq. 2.2 with respect to λ , and where $\mu_{0i} = f_0(\alpha_0^T x_i)$ for some unknown increasing function f_0 and index $\alpha_0 \in \mathcal{S}_{d-1}$. The overall sample size is $n = \sum_{i=1}^m n_i$, and the log-likelihood is

$$l_n(f, \alpha) = \sum_{i=1}^m n_i [Y_i \ell\{f(\alpha^T x_i)\} - B \circ \ell\{f(\alpha^T x_i)\}]. \quad (2.4)$$

Here, $Y_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$ is the average of the observations from group i .

2.2. The MLE in the Monotone Single Index Model The MLE of (f_0, α_0) maximizes $l_n(f, \alpha)$ in Eq. 2.4 over all increasing functions f and index vectors $\alpha \in \mathcal{S}^{d-1}$. To maximize l_n we use the following profile approach: we first hold α fixed and maximize in f ; then we maximize the resulting likelihood in α .

Consider first maximization in f for a fixed α . Let $\mathcal{M}$ denote the class of increasing functions $f : \mathbb{R} \rightarrow \mathbb{R}$. For $k \in \mathbb{N} \setminus \{0\}$, vectors $w \in \mathbb{R}_+^k$ and $W \in \mathbb{R}^k$, let $\text{iso}(w, W)$ denote the weighted isotonic regression of W , which is the unique solution to

$$\text{iso}(w, W) = \arg \min_{\beta \in \mathcal{M}_k} \sum_{i=1}^k w_i (W_i - \beta_i)^2, \quad (2.5)$$

where $\mathcal{M}_k$ denotes the class of vectors $\beta \in \mathbb{R}^k$ such that $\beta_1 \leq \dots \leq \beta_k$. Some well-known properties of isotonic regression, such as uniqueness, are reviewed in the Supplementary Material, Section A.

PROPOSITION 2. Fix $\alpha \in \mathcal{S}_{d-1}$. Let $z_1, \dots, z_k$ denote the unique values of $\alpha^T x_1, \dots, \alpha^T x_m$. The maximizer of $l_n(f, \alpha)$ over $f \in \mathcal{M}$ exists and is unique at the points $z_1, \dots, z_k$. Moreover, if π denotes the unique permutation on $\{1, \dots, k\}$ such that $z_{\pi(1)} < \dots < z_{\pi(k)}$, then the values of the maximizer at the points $z_{\pi(1)}, \dots, z_{\pi(k)}$ are the components of $\text{iso}(w, W)$, where

$$w_j = \sum_{i: \alpha^T x_i = z_{\pi(j)}} n_i, \quad \text{and} \quad W_j = (w_j)^{-1} \sum_{i: \alpha^T x_i = z_{\pi(j)}} n_i Y_i,$$

for $j = 1, \dots, k$.

Let $\widehat{f}_n|_\alpha$ denote the maximizer of $l_n(f, \alpha)$ over $f \in \mathcal{M}$ for fixed α . We now maximize $l_n(\widehat{f}_n|_\alpha, \alpha)$ in α to find the MLE. The MLE of (f_0, α_0) is any pair $(\widehat{f}_n, \widehat{\alpha}_n)$ where $\widehat{\alpha}_n \in \{\widehat{\alpha}_n\}$, where the set $\{\widehat{\alpha}_n\}$ is defined in the next proposition, and $\widehat{f}_n = \widehat{f}_n|_{\widehat{\alpha}_n}$, whereas the MLE $\widehat{\mu}_n$ of μ_0 has components $\widehat{\mu}_{ni} = \widehat{f}_n(\widehat{\alpha}_n^T x_i)$, $i = 1, \dots, m$. The MLE exists, but is not unique, as the following propositions shows.

PROPOSITION 3. *Let $\{\widehat{\alpha}_n\}$ denote the set of all maximizers of $l_n(\widehat{f}_n|_\alpha, \alpha)$ over $\alpha \in \mathcal{S}^{d-1}$. We can find $\widehat{\alpha}_n \in \{\widehat{\alpha}_n\}$ such that $\widehat{\alpha}_n^T x_i \neq \widehat{\alpha}_n^T x_j$ for all $i \neq j$. Furthermore, $\{\widehat{\alpha}_n\}$ has positive Lebesgue measure in $\mathcal{S}^{d-1}$.*

It is also important to note the ways in which the maximum likelihood estimator is not unique. First, the maximum likelihood estimator $\widehat{f}_n|_\alpha$ of the link function for a fixed α is uniquely defined only on the set of projections $\alpha^T x_i$ ($i = 1, \dots, m$), and any monotone interpolation of the values at these points gives a maximum likelihood estimator. More importantly, with $\widehat{\alpha}_n$ as above, it can be seen from the proof of Proposition 3 that any $\alpha \in \mathcal{S}_{d-1}$ such that $\alpha^T x_i < \alpha^T x_j$ for all i, j satisfying $\widehat{\alpha}_n^T x_i < \widehat{\alpha}_n^T x_j$, maximizes the likelihood.

3 Asymptotics of the Maximum Likelihood Estimator

3.1. Consistency Assume that the design points $x_1, \dots, x_m$ are fixed. Throughout, we make one of the two following assumptions about the design.

Fixed design: We assume that there are n_i ($i = 1, \dots, m$) observations collected from group i at each design point x_i ($i = 1, \dots, m$) with a non-random n_i . We consider what happens as n increases in such a way that $\lambda_{n,i} = n_i/n$ converges to some $\lambda_i \in (0, 1)$ for all $i = 1, \dots, m$. We call this assumption (A1).

Random design: We assume that at each design point x_i there is a random number n_i of observations Y_{ij} where $(n_1, \dots, n_m)$ follows a multinomial distribution with parameters n and λ_i ($i = 1, \dots, m$). This setting occurs if one observes independent copies of (X, Y) where X is a discrete variable such that $\text{pr}(X = x_i) = \lambda_i \in (0, 1)$ for $i = 1, \dots, m$ and Y is from the group i if and only if $X = x_i$. Again, we consider what happens as n increases, and call this assumption (A2).

Recall the definition of $\mathcal{S}_0$ given above in Eq. 2.3.

PROPOSITION 4. *Under either assumption (A1) or (A2), almost surely there exists a positive integer $N_0 = N_0(\omega)$ such that for all $n \geq N_0$, $\{\widehat{\alpha}_n\} \subset \mathcal{S}_0$.*

Proposition 4 shows that any monotone single index model MLE of $\hat{\alpha}_n$ is consistent for the set $\mathcal{S}_0$. The model is set identifiable for α_0 up to the set $\mathcal{S}_0$, see Proposition 1, and hence a stronger statement should not be expected. However, simulations in Section 4.1 indicate that the method estimates α_0 reasonably well for finite sample sizes, with efficiency similar to or only slightly worse than a fully parametric model.

We also note that, defining $d(\hat{\alpha}_n, \mathcal{S}_0) = \min_{\alpha_0 \in \mathcal{S}_0} \|\hat{\alpha}_n - \alpha_0\|$, Proposition 4 implies the following. For all $\varepsilon > 0$ and all rates r_n (i.e. r_n is a positive number which converges to infinity as $n \rightarrow \infty$) we have

$$P(r_n d(\hat{\alpha}_n, \mathcal{S}_0) > \varepsilon) \leq P(\hat{\alpha}_n \notin \mathcal{S}_0) \leq P(N_0 > n) \rightarrow 0,$$

as $n \rightarrow \infty$, where N_0 is taken from Proposition 4. Hence, $d(\hat{\alpha}_n, \mathcal{S}_0) = o_p(r_n^{-1})$ for any rate r_n .

Lastly, an interesting alternative to the approach above is to consider instead estimation of the entire set $\mathcal{S}_0$. Alas, algorithms to produce the entire set $\{\hat{\alpha}_n\}$ are still an open problem, and these are a crucial first step to such lines of inquiry.

One important consequence of Proposition 4 is the following.

PROPOSITION 5. *Let $\hat{\mu}_n$ denote the MLE of μ_0 , and $(f_0, \alpha_0) \in \mathcal{M} \times \mathcal{S}^{d-1}$ such that $\mu_{0j} = f_0(\alpha^T x_j)$ for $j = 1, \dots, m$. Then, under either assumption (A1) or (A2), almost surely, there exists a positive integer N_0 such that for all $n \geq N_0$*

$$\sum_{j=1}^m n_j \Psi(\hat{\mu}_{0j} - \mu_{0j}) \leq \sum_{j=1}^m n_j \Psi(Y_j - \mu_{0j}),$$

for any convex function Ψ . In particular, we have

$$\sum_{j=1}^m n_j (\hat{\mu}_{0j} - \mu_{0j})^2 \leq \sum_{j=1}^m n_j (Y_j - \mu_{0j})^2.$$

The preceding result shows the benefits of the proposed single index model over the naive model where nothing is assumed about the true means μ_{0i} in Eq. 2.2. In the naive model, the natural estimator for μ_{0i} is Y_i , and thus the proposition shows that our estimator has smaller risk than this naive estimator. Similar results were obtained for the discrete Grenander estimator (Jankowski and Wellner, 2009). Some simulations studying Proposition 5 are given in the Supplementary Material, Section C.3.

3.2. Limiting Distribution For a vector $\mu \in \mathbb{R}^m$ we write $\text{part}(\mu)$ to denote the partition of the set $\{1, \dots, m\}$ generated by the flat regions of μ . For example, suppose that $m = 5$ and $\mu_1 < \mu_2 = \mu_4 < \mu_3 = \mu_5$. Then, in this example, $\text{part}(\mu) = \{\{1\}, \{2, 4\}, \{3, 5\}\}$. To avoid ambiguity, the indices in each $J \in \text{part}(\mu_0)$ are arranged in increasing order. Let

$$Z = (Z_1, \dots, Z_m) \text{ where } Z_i \text{ are independent } N(0, 1/\lambda_i) \text{ random variables.} \quad (3.1)$$

Also, for any vector $\beta \in \mathbb{R}^m$ and $J \in \text{part}(\mu_0)$, let $\beta^J = \{\beta_i, i \in J\}$ be the restriction of β to J . With some abuse of notation we write μ_0^J for the constant value taken by μ_0 on a subset $J \in \text{part}(\mu_0)$. Our main theorem follows.

THEOREM 1. *For all $\alpha \in \mathcal{S}_0$, let $\widehat{g}|_\alpha$ be such that for all $J \in \text{part}(\mu_0)$, $(\widehat{g}|_\alpha)^J$ is the unique minimizer of $\sum_{i \in J} \lambda_i (g_i - Z_i)^2$ over the set of vectors g of length $|J|$ such that $g_i \leq g_j$ whenever $\alpha^T x_i \leq \alpha^T x_j$, for all $(i, j) \in J \times J$. Then, the function*

$$\alpha \mapsto \sum_{i=1}^m \lambda_i ((\widehat{g}|_\alpha)_i - Z_i)^2 \quad (3.2)$$

achieves its minimum over $\mathcal{S}_0$. Let $\widehat{\alpha}$ denote an arbitrary minimizer. Then, under either assumption (A1) or (A2),

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) \Rightarrow \widehat{h}|_{\widehat{\alpha}}$$

where $(\widehat{h}|_{\widehat{\alpha}})_i = (\phi/\ell'(\mu_{0i}))^{1/2}(\widehat{g}|_{\widehat{\alpha}})_i$ ($i = 1, \dots, m$).

Despite its complex appearance, it is simple to sample from the limiting distribution described in Theorem 1, assuming that μ_0 is known.

It follows from Theorem 1 that in the important setting where all components of μ_0 differ from each other, the limiting distribution of $\sqrt{n}(\widehat{\mu}_n - \mu_0)$ is Gaussian since $(\widehat{g}|_{\widehat{\alpha}})_i = Z_i$ ($i = 1, \dots, m$), with Z_i as in Eq. 3.1. This holds in particular if the true link function f_0 is strictly monotone. Practitioners comfortable with this assumption may therefore use the Gaussian limiting distribution to obtain confidence intervals for μ_0 , as the variance can be easily estimated as $\widehat{\phi}_n/\lambda_{n,i}\ell'(\widehat{\mu}_{n,i})$ (estimation of dispersion is reviewed in Section E.2 of the Supplementary Material, while ℓ comes from the exponential family assumption). We also note that the limiting distribution depends on the exponential family only through the normalization $(\phi/\ell'(\mu_{0i}))^{1/2}$.

The $\sqrt{n}$ convergence rate is a result of the size of the underlying estimation space. In our set-up, the underlying distribution is discrete, and

therefore the space of possible values of the estimators is small enough that, despite the nonparametric nature of the estimator, the convergence rate is still as fast as in the parametric setting. This is not the case, however, in the continuous set-up, where the large size of the underlying space decreases the convergence rate to $n^{1/3}$, see Balabdaoui et al. (2019).

3.3. Goodness-of-Fit Define the null hypothesis H_0 : “ $\mu_0(x_i) = f_0(\alpha_0^T x_i)$, $i = 1, \dots, m$ for some $(f_0, \alpha_0) \in \mathcal{M} \times \mathcal{S}^{d-1}$ ”. The alternative hypothesis H_1 is that H_0 does not hold.

Define the test statistic

$$D_n^2 = n \sum_{i=1}^m \lambda_{n,i} \ell'(Y_i) (Y_i - \hat{\mu}_{ni})^2. \quad (3.3)$$

We consider a weighted sum of squares since the variables Y_i have unequal variances. Under assumption (A1), the variance of Y_i is $\phi[n_i \ell'(\mu_{0i})]^{-1}$, see for instance Silvapulle and Sen (Silvapulle and Sen, page 50). Under assumption (A2) the previous expression gives the conditional variance of Y_i , (conditioning on n_i , $i = 1, \dots, m$).

PROPOSITION 6. *Suppose that assumption (A1) or (A2) holds. Then, under H_0 the statistic D_n^2 converges in distribution to*

$$D^2 = \phi \sum_{i=1}^m \lambda_i ((\hat{g}|\hat{\alpha})_i - Z_i)^2 \quad (3.4)$$

where $\hat{g}|\hat{\alpha}$ is taken from Theorem 1.

The distribution of the random variable D^2 , depends on several unknowns which require estimation. First, it involves the dispersion ϕ , though this can be estimated using standard techniques, see Section E.2 in the Supplementary Material. Second, it depends on the unknown true mean μ_0 in a nontrivial manner via $\mathcal{S}_0$ and $\text{part}(\mu_0)$. This dependence makes it difficult to estimate in practice. To circumvent this issue, we calibrate the test using quantiles of a simpler random variable $\widehat{D}_n^2$ which we conjecture to be asymptotically stochastically larger than D^2 .

To define $\widehat{D}_n^2$, consider independent mean zero Gaussian random variables $Z_{n,1}, \dots, Z_{n,m}$ with variances $1/\lambda_{n,i}$, $i = 1, \dots, m$, respectively. For all $\alpha \in \mathcal{S}^{d-1}$, let $\tilde{\mathcal{G}}_\alpha$ be the set of all vectors $g \in \mathbb{R}^m$ such that $g_i \leq g_j$ whenever $\alpha^T x_i \leq \alpha^T x_j$, for all $(i, j) \in \{1, \dots, m\}^2$, and $\check{g}|\alpha$ be the unique minimizer of $\sum_{i=1}^n \lambda_{n,i} (g_i - Z_{n,i})^2$ over $\tilde{\mathcal{G}}_\alpha$. Define

$$\widehat{D}_n^2 = \hat{\phi}_n \inf_{\alpha \in \mathcal{S}^{d-1}} \sum_{i=1}^m \lambda_{n,i} ((\check{g}|\alpha)_i - Z_{n,i})^2, \quad (3.5)$$

where $\hat{\phi}_n$ is any consistent estimator of the true dispersion parameter ϕ ; see Section E.2 of the Supplementary Material. Clearly, as $n \rightarrow \infty$, $\widehat{D}_n^2$ converges in distribution to

$$\widehat{D}^2 = \phi \inf_{\alpha \in \mathcal{S}^{d-1}} \inf_{g \in \widehat{\mathcal{G}}_\alpha} \sum_{i=1}^m \lambda_i (g_i - Z_i)^2. \quad (3.6)$$

If μ_0 is constant then $\text{part}(\mu_0)$ contains the single element $\{1, \dots, m\}$ and $\mathcal{S}_0 = \mathcal{S}^{d-1}$. In this case, we have the equality $D^2 = \widehat{D}^2$. Therefore, the asymptotic distribution of $\widehat{D}_n^2$ is the same as D^2 when μ_0 is constant. In tests for monotonicity in the context of shape constrained estimation, the constant setting has previously been shown to be least favorable, see for example Ghosal et al. (2000, Inequality (2.7)) or Durot (2003, Theorem 1). We conjecture that this holds here as well; i.e. we conjecture that $D^2 \leq_{st} \widehat{D}^2$. The conjecture clearly holds if μ_0 is constant. It also holds if μ_0 is such that $\mu_{0i} \neq \mu_{0j}$ for all $1 \leq i \neq j \leq m$ (which is true if the link function is strictly increasing) since in this case $D^2 = 0$. Therefore, our conjecture holds in two key scenarios. To prove that the constant case is indeed least favorable is out of the scope of the paper.

Let $\widehat{q}_{n,1-A}$ denote the $(1-A)$ -quantile of $\widehat{D}_n^2$, for some fixed probability $A \in (0, 1)$. The goodness-of-fit test rejects H_0 if $D_n^2 > \widehat{q}_{n,1-A}$. This guarantees a hypothesis test with asymptotic level no larger than A when μ_0 is constant and when $\mu_{0i} \neq \mu_{0j}$ for $i \neq j$ ($i, j = 1, \dots, m$). Moreover, the asymptotic level is exact when μ_0 is constant.

A second goodness-of-fit test is based on the following stochastic inequality.

PROPOSITION 7. *Let $Z = (Z_1, \dots, Z_m)^T$ such that $Z_1, \dots, Z_m$ are independent and $Z_i \sim \mathcal{N}(0, \lambda_i^{-1})$. Also, let $K^2 = Z^T Q Z$ denote a generalized chi-squared random variable, where $Q = A_m \Lambda A_m$, with Λ denoting the matrix with diagonal elements equal to $\lambda_i, i = 1, \dots, m$ and $\{A_m\}_{i,j} = \delta_{i,j} - 1/m$. Then,*

$$D^2 \leq_{st} \phi K^2,$$

where D^2 is the random variable defined in (3.4). When $\lambda_i = 1/m, i = 1, \dots, m$, then K^2 is a chi-squared random variable with $m-1$ degrees of freedom.

The proof is straightforward and is provided in the [Supplementary Material](#). Note that the same approach can also be used to show that $\widehat{D}^2 \leq_{st} \phi K^2$, which is consistent with the conjecture that $D^2 \leq_{st} \widehat{D}^2$. A histogram showing the extent of the difference between the distributions of $D^2, \widehat{D}_n^2$, and

ϕK^2 for a particular choice of design is given in Fig. 2 in the Supplementary Material. Section 4.2 studies the finite sample performance of the two proposed tests.

4 Finite Sample Performance

We next investigate the finite sample performance of the MLE in different scenarios. To compute the MLE, we use a modified stochastic search algorithm over α , and find $\hat{f}_n|_\alpha(\alpha^T x_j), j = 1, \dots, m$ using PAVA. We use linear interpolation to obtain $\hat{f}_n|_\alpha(\alpha^T x)$.

4.1. Estimation and Prediction

EXAMPLE 1. This example is taken from Wan et al. (2017), where it forms the basis of their simulation study. The true model is $Y_{ij} = \mu_0(x_i) + \varepsilon_{ij}$ where ε_{ij} are i.i.d. $\sim \mathcal{N}(0, \sigma_0^2)$. In this example the different combinations of the levels of two drugs, d_1 and d_2 say, are used for prediction. The interaction of these drugs is modeled using the Emax model (Wagner, 1968; Kong and Lee, 2006). The dose levels for both drugs were set to $(0, 0.1, 0.5, 1, 2, 4)$ and each dose level for drug 1 was combined with each dose level for drug 2, for a total number of interactions equal to $m = 36$. The true mean response curve is given by

$$\mu_0(x) = \frac{100\beta_0^T x}{1 + \beta_0^T x},$$

where

$$x^T = \left(d_1, d_2, \sqrt{d_1 d_2}, d_1 \sqrt{d_2}, d_2 \sqrt{d_1}, \sqrt{d_1^3 d_2}, \sqrt{d_1 d_2^3} \right), \quad (4.1)$$

where again d_i is the level of drug $i, i = 1, 2$. In this model, we can re-write $\mu_0(x) = f_0(\alpha_0^T x)$ with $\alpha_0 = \beta_0 / \|\beta_0\|$ and $f_0(t) = 100\|\beta_0\|t / (1 + \|\beta_0\|t)$, which is a monotone increasing function. As in Wan et al. (2017), we take

$$\beta_0^T = (1, 1, 0.9, 0, 0, -0.3, -0.3, 0).$$

In our simulation study, we considered several estimation methods, as follows.

- (a). Generalized linear model with inverse link function as in the true model. We call this the “correct GLM”.
- (b). Monotone single index model with $\mu_0(x) = f_0(\alpha_0^T x)$ where x is as given in Eq. 4.1.

- (c). Empirical approach where $\mu_0(x_i)$ is estimated by Y_i . This yields no estimates of α_0 and cannot be used in predicting the response for new predictor values.
- (d). A misspecified monotone single index model with $\mu_0(x) = f_0(\alpha_0^T x)$ where $x^T = (d_1, d_2)$.
- (e). A misspecified single index model with $\mu_0(x) = f_0(\alpha_0^T x)$ where $x^T = (d_1, d_2)$ and no assumptions are made on f_0 . We use a kernel-based estimator calculated with the R package **np** (Hayfield and Racine, 2008).
- (f). A misspecified generalized linear model with inverse link function f_0 as in the true model but with $\mu_0(x) = f_0(\alpha_0^T x)$ where $x^T = (d_1, d_2)$.
- (g). A misspecified generalized linear model with link function equal to the identity and the predictor x as given in Eq. 4.1.
- (h). A well-specified single index model with $\mu_0(x) = f_0(\alpha_0^T x)$ where the predictor x is as given in Eq. 4.1 but no assumptions are made on f_0 . We use a kernel-based estimator calculated with the R package **np** (Hayfield and Racine, 2008).
- (i). A misspecified generalized linear model with identity link function and $x^T = (d_1, d_2)$.

Table 1 gives the mean of the estimated α_0 values for approaches (a), (b), (f), and (h)—as these are the only approaches which enable its estimation. Although nothing beats the correct GLM approach, the monotone single index model (b) performs well. Figure 1 shows boxplots of the estimates for models (a) and (b), from which one can note the striking fact that the MLE in the monotone single index model does not exhibit a big loss efficiency when compared with the MLE in the true GLM model. Note that in Table 1 and Fig. 1 we only report the case where $\sigma_0^2 = 1$ and $n_i = 3$.

Table 1: Estimation of α_0 in Example 1 with $\sigma_0 = 1$ and $n_i = 3$: mean values of estimated α across $B = 200$ simulations

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
true α_0	0.659	0.659	0.593	0	0	-0.198	-0.918	0
(a)	0.575	0.573	0.514	-0.010	0.016	-0.168	-0.177	-0.001
(b)	0.568	0.567	0.305	0.105	0.130	-0.176	-0.183	-0.049
(f)	0.132	0.132	0.775	-0.407	-0.404	-0.007	-0.008	0.184
(h)	0.015	0.068	0.288	-0.192	-0.228	0.226	0.178	-0.216

Figure 1: The plot compares the estimates of α_0 for Example 1 for the correct GLM (a) and well-specified monotone single index method (b) over $B = 200$ simulations. Boxplots of $\hat{\alpha}_{n,j} - \alpha_{0j}, j = 1, \dots, d = 2$ are shown, with j appearing on the x -axis. Here, $\sigma_0^2 = 1$ and $n_i = 3$

The five other settings performed similarly and hence their results are not reported here. In Tables 2 and 3, we consider estimation and prediction of the response surface: To study the estimation performance, we report in Table 2 the average sum of squared deviations of the different estimators from the true mean response surface at the observed predictors. The prediction performances are compared in Table 3 where we report the average sum of squared errors evaluated at 100 new and randomly selected predictor values. As expected, the correct GLM approach is best in all instances. In terms of prediction and estimation, the second best model is (b), the well-specified monotone single index model—with the exception of model (c) which sometimes outperforms (b) in Table 2. Theoretically, we know from Proposition 5 that (b) will outperform (c) asymptotically. Recall also that (c) does not produce estimates of α_0 and cannot be used for prediction. We note that the kernel-based single-index approach (e) and (h) does not perform well throughout. The comparison here may not necessarily be fair, as one would not expect a nonparametric method perform well with small values of m . We include it here to highlight the importance of the

Table 2: Estimation of the response surface in Example 1: mean values of sum of squared errors at observed predictors across $B = 200$ simulations

σ_0^2	n_i	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
0.1	3	0.026	0.290	0.123	174.97	624.34	862.97	1826.27	3119.24	7860.14
	10	0.008	0.207	0.036	174.92	620.34	862.97	1826.25	3304.51	7860.13
	100	0.001	0.173	0.004	174.91	623.10	862.96	1826.24	3190.56	7860.13
1	3	2.585	10.70	11.94	181.35	641.30	863.97	1829.15	3549.25	7861.09
	10	0.765	3.34	3.62	176.82	619.86	863.25	1827.11	3663.78	7860.41
	100	0.081	0.40	0.35	175.09	621.65	863.00	1826.33	3252.70	7860.16

monotonicity assumption in method (b). As mentioned in the introduction, Foster et al. (2013) also find that the assumption of monotonicity “noticeably improves performance”. In fact, even the misspecified monotone model (d) outperforms methods (e)–(i) in Tables 2 and 3.

EXAMPLE 2. In our second example, we consider the setting of a logistic regression. We take Y_{ij} to be independent Bernoulli variables, with probability of success given by

$$\mu_0(x_i) = \text{logit}^{-1} (h_0(\alpha_0^T x_i))$$

and consider three different cases of h_0 . We will assume throughout that $x = (x_1, x_2)^T$, and take $\alpha_0^T = (1, 1)/\sqrt{2}$. For the first variable x_1 , the design levels are set to $(0, 0.01, 0.1, 0.25)$ and $(0, 0.01, 0.1, 1.0)$ for the second variable x_2 . Each level for x_1 was combined with each level for x_2 , for a total of $m = 16$. This is the same design as that in case study in Section 5, except that the levels have been rescaled. We take values $n_i = 3, 10, 100$ as in Example 1 and consider the following four estimation methods. Note that the samples n_i in the data set of Section 5 are close 100.

Table 3: Prediction of the response surface in Example 1: mean values of sum of squared errors at 100 randomly sampled new predictor values across $B = 200$ simulations

σ_0^2	n_i	(a)	(b)	(d)	(g)	(f)	(i)	(h)	(e)
0.1	3	0.066	20.62	268.30	498.65	3226.68	11998.56	40884.34	269698.1
	10	0.020	21.94	270.54	502.19	3214.53	12003.50	39089.48	273460.0
	100	0.002	18.33	267.26	499.00	3243.34	12062.79	36247.1	276533.0
1	3	6.40	56.60	274.03	517.20	3164.58	11995.58	36222.50	255228.5
	10	1.94	39.46	274.13	512.07	3248.82	12124.38	39956.67	253959.1
	100	0.21	20.18	260.82	505.66	3177.71	11905.38	39552.99	270996.2

- (A.) Generalized linear model with $\mu_0(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. This yields the estimate

$$\hat{\mu}_n(x) = \hat{\beta}_{n,0} + \hat{\beta}_{n,1}x_1 + \hat{\beta}_{n,2}x_2 = \hat{\beta}_{n,0} + \hat{\beta}_{n,1,2}^T x$$

which we rewrite as

$$\hat{\mu}_n(x) = \hat{\beta}_{n,0} + \|\hat{\beta}_{n,1,2}\| \tilde{\alpha}_n^T x$$

to obtain the GLM estimate of α_0 , $\tilde{\alpha}_n = (\hat{\beta}_{n,1}, \hat{\beta}_{n,2})^T / \|\hat{\beta}_{n,1,2}\|$, with $\|\hat{\beta}_{n,1,2}\|^2 = \hat{\beta}_{n,1}^2 + \hat{\beta}_{n,2}^2$.

- (B.) Monotone single index model with $\mu_0(x) = f_0(\alpha_0^T x)$.
- (C.) Empirical approach where $\mu_0(x_i)$ is estimated by Y_i . Again, this yields no estimates of α_0 and cannot be used for making prediction.
- (D.) A well-specified single index model with $\mu_0(x) = f_0(\alpha_0^T x)$ but with no assumptions on f_0 . We use a kernel-based estimator calculated with the R package `np` (Hayfield and Racine, 2008).

The three different cases for h_0 are chosen to consider a variety of scenarios.

- (i) $h_0(x) = 5x$. With this choice of h_0 , models (A)–(C) are well-specified.
- (ii) $h_0(x) = 7.5 - \max\{(5x - 3)^2, 6.25\}$. This function has both increasing and flat regions. With this choice of h_0 , models (B)–(C) are well-specified, but (A) is not.
- (iii) $h_0(x) = |5x - 1|$. With this choice of h_0 , only model (C) is well-specified.

Plots of $\mu_0(x_i)$ under the three scenarios are given in Fig. 2.

The results of our simulations follow. Table 4 gives the mean of the estimated α_0 values for approaches (A), (B), (D), and Fig. 3 shows boxplots of the estimates for $n_i = 100$ (this is the setting most similar to the real data set). Table 5 reports the mean sum of squared errors for the different estimators from the true mean response surface at the observed predictors, while Table 6 reports the same, except evaluated at 100 randomly selected predictor values, chosen to be within the range of the fixed design. In case (i), as expected, GLM performs best in all metrics, followed by the monotone SIM. For a semi-parametric approach, the monotone SIM method performs

Figure 2: Plots of $\mu_0(x_i), i = 1, \dots, m$ against $\alpha_0^T x_i$ for cases (i)–(iii), left to right, in Example 4. Since the plot shows the response level mean, the function in case (i) is not linear after the logit transformation

remarkably well when compared with the correct parametric method. In case (ii), where approach (A) is misspecified, the monotone SIM performs better than GLM. Both the GLM and monotone SIM methods also do not perform well for case (iii), as expected. As in Example 1, the kernel-based estimator of the single index model performs poorly for prediction of the response surface, despite being well-specified in all three cases.

We also repeated this example with a Poisson distribution instead of the binomial. The results are similar to those presented here, and are detailed in the [Supplementary Material](#).

4.2. Goodness-of-Fit Test We perform a simulation study of the proposed goodness-of-fit test for all examples of the previous section, with results given in Table 7. For Example 1, Example 2 case (i) and (ii), the conservative nature of the test is clearly visible. Furthermore, in Example 1, Example 2 case (i), the true mean vector is strictly increasing, in which case asymptotically $D^2 = 0$, contributing to the effect. In Example 2 case (ii), the true mean has a region where it is flat. However, for the particular choice of covariates, this region is quite small. For this reason, we added an artificial Example 2 case (iv) where $\mu_{0i} = E[Y_i|x_i] = 0.5, i = 1, \dots, m$. In this example, the mean function is flat in which case we know that the asymptotic distributions D^2 and $\hat{D}^2$ are the same. This fact that can be seen from the observed significance levels. Lastly, in Example 2 case (iii) the model is misspecified. Here, the power performs reasonably well, noting that when $n_i = 3$, the overall sample size is $n = 108$. Additional simulations

Table 4: Estimation of α_0 in Example 2: mean values of estimated α across $B = 200$ simulations. Recall that $\alpha_0^T = (1, 1)/\sqrt{2} = (0.707, 0.707)$

Case	n_i	method	α_1	α_2
(i)	3	(A)	0.511	0.717
		(B)	0.513	0.635
		(D)	0.674	0.570
	10	(A)	0.602	0.734
		(B)	0.557	0.734
		(D)	0.759	0.544
	100	(A)	0.690	0.717
		(B)	0.694	0.683
		(D)	0.794	0.566
	3	(A)	0.914	0.252
		(B)	0.711	0.482
		(D)	0.898	0.296
(ii)	10	(A)	0.972	0.213
		(B)	0.701	0.579
		(D)	0.797	0.515
	100	(A)	0.979	0.202
		(B)	0.754	0.550
		(D)	0.681	0.715
	3	(A)	-0.440	0.650
		(B)	-0.341	0.562
		(D)	0.629	0.493
	10	(A)	-0.606	0.673
		(B)	-0.473	0.644
		(C)	0.715	0.560
(iii)	10	(A)	-0.735	0.669
		(B)	-0.872	0.470
	100	(B)	-0.872	0.470
		(D)	0.811	0.532

exploring the goodness-of-fit test for a variety of ridge functions in more depth are given in the Appendix.

When the exponential family is binomial, it is possible for small sample sizes to observe Y_i equal to 0 or 1, which makes computation of D_n^2 in Eq. 3.3 difficult, as $\ell'(Y_i) = 1/0$ in this case. Therefore, for the binomial family we replaced the term $\ell'(Y_i)$ with $\ell'(\hat{\mu}_{ni})$, as $\hat{\mu}_{ni}$ is smoother and hence has fewer such instances. This alteration does not change the asymptotic behaviour of the test statistic. For cases where $\ell'(\hat{\mu}_{ni}) = 1/0$, we simply set the term

Figure 3: Estimates of α_0 for Example 2 for methods (A)–(C) over $B = 200$ simulations. Boxplots of $\hat{\alpha}_{n,i} - \alpha_{0i}$ are shown, with i appearing on the x -axis. From top to bottom, the plots show cases (i)–(iii) for $n_i = 100$

artificially to zero (i.e. $\ell'(\hat{\mu}_{ni}) = 0$). The effect of this only augments the conservative nature of the tests.

Table 5: Estimation of the response surface in Example 2: mean values of sum of squared errors at observed predictors across $B = 200$ simulations

Case	n_i	(A)	(B)	(C)	(D)
(i)	3	0.161	0.350	0.935	0.230
	10	0.055	0.120	0.305	0.091
	100	0.005	0.015	0.027	0.014
(ii)	3	0.496	0.397	1.026	0.540
	10	0.364	0.140	0.303	0.202
	100	0.320	0.017	0.029	0.018
(iii)	3	0.194	0.382	0.926	0.263
	10	0.091	0.141	0.308	0.111
	100	0.040	0.024	0.028	0.013

5 A Drug Interaction Case Study

As in Wan et al. (2017), we consider a drug interaction data set. Carter et al. (1986) present an analysis of cytotoxicity data, where the effect of treatment combinations from multiple drug agents on the binary response is studied. The response counts the number of surviving human promyelocytic Leukemia cells after being exposed to a combination of doses of methylmethanesulfonate and phorbol 12-myristate, 13-acetate. In the analysis, the dose-response surface is estimated using multivariate parametric logistic regression. The data consists of a total of 16 design points, four per agent, with anywhere from 83 – 98 observations per point and an overall sample size of $n = 1436$ (Carter et al. 1986, Table 1, page 125).

Table 6: Prediction of the response surface in Example 2: mean values of sum of squared errors at 100 randomly sampled new predictor values across $B = 200$ simulations

Case	n_i	(A)	(B)	(D)
(i)	3	0.655	1.508	4.046
	10	0.284	0.532	4.102
	100	0.022	0.221	5.493
(ii)	3	23.015	14.698	47.919
	10	16.502	5.172	59.482
	100	14.032	1.570	119.579
(iii)	3	1.845	2.553	4.899
	10	1.279	1.477	4.012
	100	0.938	1.167	4.065

Table 7: Significance level or power estimates for goodness-of-fit test for all examples in Section 4.1. Results are based on 200 simulations for a maximal standard error of 0.035 for each level estimate. Mean observed p -value is also reported

		$\widehat{D}^2$		Chi-square	
			Mean		Mean
	n_i	Level	p -value	Level	p -value
Example 1	$\sigma_0^2 = 1$	3	0.015	0.005	0.987
		10	0	0	1.000
		100	0	0	1.000
	$\sigma_0^2 = 0.1$	3	0	0	1.000
		10	0	0	1.000
		100	0	0	1.000
		3	0.01	0	0.897
	case (i)	10	0.025	0	0.899
		100	0	0	0.967
Example 2	case (i)	3	0.060	0	0.829
		10	0.055	0	0.872
		100	0.010	0	0.940
	case (ii)	3	0.060	0	0.809
		10	0.140	0.035	0.729
		100	0.570	0.170	0.396
	case (iii)	1000	1	1	0
		3	0.085	0	0.785
		10	0.060	0	0.832
	case (iv)	100	0.075	0.005	0.818

The goal of the data analysis in Carter et al. (1986) was to evaluate the dose-response surface for the combination of two drug agents. Following Carter et al. (1986), we similarly transform the methylmethanesulfonate variable. In their paper, the authors perform a second order polynomial logistic regression, resulting in the fitted parametric mean function $\widehat{\mu}_1(x) = -1.33 - 0.84x_1 + 0.159x_2 + 0.003883x_1^2 - 0.001308x_2^2$, after backward stepwise model selection. Here, $x = (x_1, x_2)$, where x_1 is the transformed methylmethanesulfonate variable, and x_2 is the phorbol 12-myristate, 13-acetate variable.

We compare their parametric model with a single index model where $\mu(x) = f_0(\alpha_0^T x)$ with $x = (x_1, x_2)^T$ as above, under two assumptions: f_0 is either entirely unknown or increasing. When no assumptions are

made on f_0 , we use the fitting method of Klein and Spady (1993) and let $\hat{\mu}_2(x) = \hat{f}_{n,2}(\hat{\alpha}_{n,2}^T x)$ denote the nonparametric single index estimate. When f_0 is assumed to be increasing, we use the maximum likelihood estimator proposed in this work, and let $\hat{\mu}_3(x) = \hat{f}_{n,3}(\hat{\alpha}_{n,3}^T x)$ denote the corresponding estimate. The nonparametric estimate was computed using the R package `np` (Hayfield and Racine, 2008). We also consider the monotone single index model with elliptical contours, that is, $\hat{\mu}_4(x) = \hat{f}_{n,4}(\hat{\alpha}_{n,4}^T y)$ where $y = (x_1, x_2, x_1^2, x_1 x_2, x_2^2)^T$, and the model $\hat{\mu}_5(x) = \hat{f}_n(\hat{\alpha}_{n,5}^T \tilde{y})$, with $\tilde{y} = (x_1, x_2, x_1^2, x_2^2)^T$, where the ellipses have been constrained to have axes parallel to the coordinate axes. We do not include the intercept term as it is not identifiable. We choose the convention that $\hat{f}_{n,3}$, $\hat{f}_{n,4}$ and $\hat{f}_{n,5}$ are piecewise linear.

Figure 4, from left to right, shows grey-scale heatmap plots of $\hat{\mu}_1(x)$, $\hat{\mu}_3(x)$, and $\hat{\mu}_5(x)$. The phorbol 12-myristate, 13-acetate variable appears on the vertical axis and the transformed methylmethanesulfonate variable appears on the horizontal axis. Black dots indicate the locations of the design points, and the size of the dot is proportional to n_i . While all three approaches

Figure 4: From left to right: grey-scale heatmap of the response surface for logistic regression, isotonic single index model with linear contours, and isotonic single index model with elliptical contours

appear to identify the phorbol 12-myristate, 13-acetate axis as the more important direction, the heatmaps appear quite different. We also implemented the proposed goodness of fit test for models $\hat{\mu}_3, \hat{\mu}_4, \hat{\mu}_5$. In all three cases, the test fails to reject at the 5% level the hypothesis that the monotone single index model is valid for the dataset. The residual sum of squares is 264.04, 264.39, 261.82, and 261.82 for $\hat{\mu}_1, \hat{\mu}_3, \hat{\mu}_4, \hat{\mu}_5$ respectively. The residual sum of squares is very similar for $\hat{\mu}_1$ and $\hat{\mu}_3$, however, the contour plots away from the design points are more reasonable for $\hat{\mu}_3$. The residual sum of squares for $\hat{\mu}_4$ and $\hat{\mu}_5$ are the same, as are the fitted values for both models. Additional plots of the models are given in Figs. 5 and 6.

An unusual characteristic of the parametric logistic regression is that it selects a polynomial model where a decrease in the response surface is seen for large values of phorbol 12-myristate, 13-acetate. Carter et al. (1986) discuss this and suggest that this is due to lysis; High concentrations of the drug render the cells uncountable. Comparing $\hat{\mu}_1$ with $\hat{\mu}_3$ explains that this is an artefact of the parametric model, and not a true phenomenon. Figure 7, left, plots the monotone estimate $\hat{f}_{n,3}(t)$, dotted line. For comparison, the solid line denoting $\hat{\mu}_1(x)$ projected, with smoothing, onto the line

Figure 5: From left to right: grey-scale heatmap of the response surface $\hat{\mu}_3, \hat{\mu}_4, \hat{\mu}_5$

Figure 6: Close up plots of the grey-scale heatmap of the response surface $\hat{\mu}_1$ (top left) $\hat{\mu}_3$ (top right), $\hat{\mu}_4$ (bottom left), and $\hat{\mu}_5$ (bottom right)

$t = \hat{\alpha}_{n,3}^T x$ has been added. The observed values Y_i are shown as grey dots. We immediately see that the parabolic effect of the parametric regression is artificial and not due to lysis. Figure 7, right, similarly compares the parametric regression with the nonparametric single index regression. Here, we see $\hat{f}_{n,2}(t)$, dashed line, plotted with $\hat{\mu}_1(x)$ in solid, projected, with smoothing, onto the line $t = \hat{\alpha}_{n,2}^T x$. The observed values Y_i are again shown as grey dots. The nonparametric method identifies the methylmethanesulfonate axis as the important direction, which contradicts both the parametric and isotonic results, as well as a three dimensional visual inspection of the data set.

Figure 7: From left to right: monotone and nonparametric estimates of the ridge function compared with a projected and smoothed parametric response surface onto the line given by the estimate of the index. In both cases the single index model has linear contours

We believe that this anomalous result is due to the small value of m . In fact, the comparison of the kernel-based is not necessarily a fair one, as this approach would not be expected to perform well with such a small design. We include it to, yet again, emphasize the benefits of the monotonicity assumption in nonparametric single index models. R scripts with the analysis and additional details can be found in the [Supplementary Material](#).

6 Conclusions

From the results obtained in the simulation study, one can observe that the main advantage of the approach proposed in this work is robustness and protection against model specification. The assumption of monotonicity of the link function is a natural one, and does not require the correct link function to be known which is a clear limitation with the parametric GLM. The only prior knowledge required from the user is the exponential family (e.g. Gaussian, binomial, Poisson) to which the distribution of the response belongs. Our simulation results show also competitiveness of the MLE. Indeed, of all estimation procedures considered, it is rivalled only by the correct parametric GLM approach, which in practice may not be known. Although the index vector in the monotone single index model with discrete design is only set-identifiable, our estimator shows remarkably small loss of efficiency against the MLE in the correct GLM. This is a very important

feature which gives much incentive to apply our method in case the user is not ready to impose some parametric form on the link function.

The assumption of monotonicity is mild, and natural in a plethora of practical situations. In our simulations, we have also considered cases when this assumption is violated. Notably the goodness-of-fit test developed here has reasonable performance in terms of its ability to detect these situations and signal to the practitioner that the monotone model is not appropriate, at least for large sample sizes. Studying further the theoretical behaviour of the model in the misspecified setting, in both the continuous and discrete cases, would be of interest though. We conjecture that the MLE will continue to converge at rate $\sqrt{n}$ to the monotone model closest to the misspecified truth in the discrete setting. Due to set identifiability in our model (see Proposition 1), these projections would require careful consideration.

Supplementary material

Supplementary material contains proofs and additional technical details.

Acknowledgements. We thank Georges Monettes and Konstantinos Ntentes from York University for helpful discussions. The second author is supported in part by MME-DII (ANR11-LBX-0023-01) and by the FP2M federation (CNRS FR 2036).

Funding. Open Access funding provided by ETH Zurich.

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- BALABDAOUI, F., DUROT, C. and JANKOWSKI, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli* **25**, 3276–3310.
- CARTER, W., CHINCHILLI, V., WILSON, J., CAMPBELL, E., KESSLER, F. and CARCHMAN, R. (1986). An asymptotic confidence region for the ed_{100p} from the logistic response surface for a combination of agents. *Am. Stat.* **40**, 124–128.

- CHEN, Y. and SAMWORTH, R. (2016). Generalised additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B* **78**, 729–754.
- DUROT, C. (2003). A Kolmogorov-type test for monotonicity of regression. *Stat. Probab. Lett.* **63**, 425–433.
- FOSTER, J.C., TAYLOR, J.M. and NAN, B. (2013). Variable selection in monotone single-index models via the adaptive lasso. *Stat. Med.* **32**, 3944–3954.
- GHOSAL, S., SEN, A. and VAN DER VAART, A.W. (2000). Testing monotonicity of regression. *Ann. Stat.* **28**, 1054–1082.
- HÄRDLE, W., SPOKOINY, V. and SPERLICH, S. (1997). Semiparametric single index versus fixed link function modelling. *Ann. Stat.* **25**, 212–243.
- HAYFIELD, T. and RACINE, J.S. (2008). Nonparametric econometrics: the np package. *J. Stat. Softw.* **27**, 1–32.
- ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econom.* **58**, 71–120.
- JANKOWSKI, H.K. and WELLNER, J.A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3**, 1567–1605.
- KAKADE, S.M., KANADE, V., SHAMIR, O. and KALAI, A. (2011). Efficient learning of generalized linear and single index models with isotonic regression, *Advances in Neural Information Processing Systems*.
- KLEIN, R.W. and SPADY, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econom.: J. Econom. Soc.* **61**, 387–421.
- KONG, M. and LEE, J.J. (2006). A generalized response surface model with varying relative potency for assessing drug interaction. *Biometrics* **62**, 986–995.
- LI, Q. and RACINE, J.S. (2007). *Nonparametric Econometrics*. Princeton University Press, Princeton.
- NELDER, J.A. and WEDDERBURN, R. (1972). Generalized linear models. *J. R. Stat. Soc. A.* **135**, 370–385.
- QIN, J., YU, T., LI, P., LIU, H. and CHEN, B. (2019). Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference. *Stat. Med.* **38**, 1442–1458.
- SILVAPULLE, M.J. and SEN, P.K. (2005). *Constrained Statistical Inference. Wiley Series in Probability and Statistics*. Wiley-Interscience (Wiley), Hoboken. Inequality, order, and shape restrictions.
- WAGNER, J. (1968). Kinetics of pharmacologic response i. proposed relationships between response and drug concentration in the intact animal and man. *J. Theoret. Biol.* **20**, 173–201.
- WAN, Y., DATTA, S., LEE, J.J. and KONG, M. (2017). Monotonic single-index models to assess drug interactions. *Stat. Med.* **36**, 655–670.

FADOUA BALABDAOUI
SEMINAR FOR STATISTICS, ETH ZÜRICH,
ZÜRICH, SWITZERLAND

HANNA JANKOWSKI
DEPARTMENT OF MATHEMATICS AND
STATISTICS, YORK UNIVERSITY, TORONTO,
CANADA
E-mail: hkj@yorku.ca

CÉCILE DUROT
MODAL'X, UNIVERSITÉ PARIS NANTERRE,
NANTERRE, FRANCE

Paper received: 26 October 2020; accepted 11 May 2021.