Supplement for the "Behavior of the monotone single index model under repeated measurements"

Fadoua Balabdaoui, Cécile Durot, and Hanna Jankowski

March 15, 2021

The supplement contains additional explanations, details, and all proofs appearing in the main manuscript. Topics are organized in the order in which they first appear in the paper.

Contents

A Isotonic Regression

For a review of isotonic regression see e.g. Barlow et al. (1972). We recall here a few well-known facts. Given $w \in (0,\infty)^k$ and $W \in \mathbb{R}^k$, the minimum of

$$\sum_{i=1}^{k} w_i(W_i - \beta_i)^2 \tag{A.1}$$

over the set $\mathcal{M}_k$ of increasing vectors $\{\beta \in \mathbb{R}^k : \beta_1 \leq \ldots \leq \beta_k\}$ is achieved at a unique point. Moreover, a vector $\widetilde{\beta} \in \mathcal{M}_k$ is the minimizer if and only if the values $\widetilde{\beta}_i$ $(i = 1,\ldots,k)$ are given by the left-hand slope of the greatest convex minorant of the cumulative sum diagram defined by the set of points

$$\left\{(0,0), \left(\sum_{l=1}^{j} w_l, \sum_{l=1}^{j} w_l W_l\right) \ (j = 1,\ldots,k)\right\}.$$

The unique minimizer can be found from the above graphical representation, or alternatively using the pool adjacent violators algorithm. The latter is implemented, for example, in the R package "Iso" (Turner, 2015).

The minimizer $\widetilde{\beta}$ is piecewise constant and is given by weighted local averages on random blocks. That is, there exist random indices $i_1 < \cdots < i_L$, where L also is random and $i_L = k$, such that for all $r \in \{1,\ldots,L\}$ and $i \in \{i_{r-1}+1,\ldots,i_r\}$,

$$\widetilde{\beta}_i = \sum_{l=i_{r-1}+1}^{i_r} w_l W_l \Big/ \sum_{l=i_{r-1}+1}^{i_r} w_l.$$

Here, $i_0 = 0$. This means that we can write $\widetilde{\beta} = AW$ where $W = (W_1,\ldots,W_k)^T$ and A is the block-wise diagonal $k \times k$ matrix such that

$$A_{ij} = \begin{cases} w_j(\sum_{l=i_{r-1}+1}^{i_r} w_l)^{-1}, & \text{if } (i,j) \in \{i_{r-1}+1,\ldots,i_r\}^2 \text{ for some } r \in \{1,\ldots,L\} \\ 0, & \text{otherwise.} \end{cases} \tag{A.2}$$

In particular, A satisfies

$$\begin{cases} \sum_{j=1}^{k} A_{ij} = 1, \quad A_{ii} > 0 & \text{for all } i \in \{1,\ldots,k\}, \\ A_{ij} = 0 \text{ if and only if } A_{ji} = 0, & \text{for all } (i,j) \in \{1,\ldots,k\}^2, \\ A_{ij} = A_{i'j} > 0 & \text{for all } (i,i',j) \in \{1,\ldots,k\}^3 \text{ such that } A_{ij} \neq 0 \text{ and } A_{i'j} \neq 0. \end{cases} \tag{A.3}$$

The last identity is a consequence of the fact that all non null entries in a given column of A are equal to each other since the rows of A corresponding to a given block are all identical.

Now, consider isotonic regression with equality constraints: instead of minimizing the criterion in (A.1) over the whole set $\mathcal{M}_k$, we minimize the criterion over the subset of vectors $\beta \in \mathbb{R}^k$ such that

$$\beta_1 = \ldots = \beta_{m_1} \leq \beta_{m_1+1} = \ldots = \beta_{m_2} \leq \ldots \leq \beta_{m_{K-1}+1} = \ldots = \beta_{m_K} \tag{A.4}$$

where $m_1 < \ldots < m_K = k$ are fixed integers. The minimizer is achieved at a unique $\widetilde{\beta}$ such that

$$\widetilde{\beta}_i = \widehat{d}_j \text{ for all } j \in \{1, \ldots, K\} \text{ and } i \in \{m_{j-1}+1, \ldots, m_j\}, \tag{A.5}$$

where $m_0 = 0$ and $(\widehat{d}_1, \ldots, \widehat{d}_K)$ is the unique minimizer of

$$\sum_{j=1}^{K} \overline{w}_j (\overline{W}_j - d_j)^2$$

over the set $\mathcal{M}_K = \{(d_1, \ldots, d_K) \in \mathbb{R}^K : d_1 \leq \ldots \leq d_K\}$. Here,

$$\overline{w}_j = \sum_{l=m_{j-1}+1}^{m_j} w_l \quad \text{and} \quad \overline{W}_j = \frac{1}{\overline{w}_j} \sum_{l=m_{j-1}+1}^{m_j} w_l W_l$$

for all $j \in \{1, \ldots, K\}$. This comes from the fact that with β satisfying (A.4) and $d_j = \beta_i$ for all $j \in \{1, \ldots, K\}$ and $i \in \{m_{j-1}+1, \ldots, m_j\}$, the difference

$$\sum_{j=1}^{K} \overline{w}_j (\overline{W}_j - d_j)^2 - \sum_{i=1}^{k} w_i (W_i - \beta_i)^2$$

does not depend on β. Since we can write $\widehat{d} = \overline{A}\,\overline{W}$ where $\overline{A}$ is defined in a similar way as A in (A.2) with w replaced by $\overline{w}$, and $\widetilde{\beta}$ statisfies (A.5), we can still write

$$\widetilde{\beta} = AW$$

where A is a block-wise diagonal matrix that satisfies (A.3). Next, we apply Barlow et al. (1972, Theorem 1.5) to conclude that

$$\sum_{i=1}^{k} w_i \widetilde{\beta}_i (\widetilde{\beta}_i - W_i) = \sum_{j=1}^{K} \overline{w}_j \widehat{d}_j (\widehat{d}_j - \overline{W}_j) = 0. \tag{A.6}$$

This property will be used in Section 3.2 of the main manuscript.

B Proofs for Section 2.1

Proof of Proposition 1. The proposition immediately follows from the assumed monotonicity of f. □

Proof of Proposition 2. The likelihood $l_n(f, \alpha)$ is equal to

$$\begin{aligned}\sum_{j=1}^{m} n_j \left[Y_j \ell\left\{f(\alpha^T x_j)\right\} - B \circ \ell\left\{f(\alpha^T x_j)\right\}\right] &= \sum_{i=1}^{k} \sum_{j:\alpha^T x_j = z_i} n_j \left[Y_j \ell\left\{f(z_i)\right\} - B \circ \ell\left\{f(z_i)\right\}\right] \\ &= \sum_{i=1}^{k} w_i \left[W_i \ell\left\{f(z_{\pi(i)})\right\} - B \circ \ell\left\{f(z_{\pi(i)})\right\}\right] \\ &= \sum_{i=1}^{k} w_i \left(W_i \ell(\beta_i) - B \circ \ell(\beta_i)\right)\end{aligned}$$

where $\beta_i = f(z_{\pi(i)})$. It follows from Silvapulle and Sen (2005, Proposition 2.4.3, page 51) that the previous sum achieves its maximum over $\beta = (\beta_1, \ldots, \beta_k)^T \in \mathcal{M}_k$ at the unique point $\mathrm{iso}(w, W)$. Moreover, $\beta \in \mathcal{M}_k$ for all $f \in \mathcal{M}$ where $\beta_i = f(z_{\pi(i)})$ so Proposition 2 follows. □

Proof of Proposition 3. First, since $\widehat{f}_n|_\alpha(\alpha^T x_i)$ $(i = 1, \ldots, m)$ depends on α only through the corresponding number $k \in \{1, \ldots, m\}$ of distinct projections and the permutation π, as described in Proposition 2, maximizing $l_n(\widehat{f}_n|_\alpha, \alpha)$ in α is equivalent to maximizing an appropriately reformulated likelihood over the possible values of k and permutations π. Clearly, this is a finite set, and therefore a maximizer exists. In other words, $\{\widehat{\alpha}_n\}$ is non-empty.

To prove the first claim of the proposition, we will prove that for arbitrary $\alpha \in \mathcal{S}_{d-1}$ such that $\alpha^T x_i = \alpha^T x_j$ for some $i \neq j$, we can find $\tilde{\alpha} \in \mathcal{S}_{d-1}$ such that $\tilde{\alpha}^T x_h \neq \tilde{\alpha}^T x_{h'}$ for all $h \neq h'$, and

$$l_n(\widehat{f}_n|_{\tilde{\alpha}}, \tilde{\alpha}) \geq l_n(\widehat{f}_n|_\alpha, \alpha). \tag{B.7}$$

Let $\alpha \in \mathcal{S}_{d-1}$ such that there exists $i < j$ with $\alpha^T x_i = \alpha^T x_j$. We can find a permutation π, and a sequence of vectors $(\alpha_t)_{t \in \mathbb{N}}$ in $\mathcal{S}^{d-1}$ such that

$$\alpha_t^T x_{\pi(1)} < \alpha_t^T x_{\pi(2)} < \cdots < \alpha_t^T x_{\pi(m)} \tag{B.8}$$

for all t, and α_t converges to α as $t \to \infty$. Let $i_1, i_2 \in \{1, \ldots, m\}$ such that $i = \pi(i_1)$ and $j = \pi(i_2)$. Letting $t \to \infty$ yields

$$\alpha^T x_{\pi(1)} \leq \cdots \leq \alpha^T x_{\pi(i_1)} = \cdots = \alpha^T x_{\pi(i_2)} \leq \cdots \leq \alpha^T x_{\pi(m)}. \tag{B.9}$$

Rearranging the terms in (4) yields

$$l_n(f, \alpha_t) = \sum_{h=1}^{m} n_{\pi(h)} \left[Y_{\pi(h)} \ell\{f(\alpha_t^T x_{\pi(h)})\} - B \circ \ell\{f(\alpha_t^T x_{\pi(h)})\} \right]$$

for all $f \in \mathcal{M}$ and all $t \in \mathbb{N}$. Maximizing over $f \in \mathcal{M}$ we get

$$l_n(\widehat{f}_n|_{\alpha_t}, \alpha_t) = \sup_{\eta_1 \leq \cdots \leq \eta_m} \sum_{h=1}^{m} n_{\pi(h)} \left\{ Y_{\pi(h)} \ell(\eta_h) - B \circ \ell(\eta_h) \right\} \tag{B.10}$$

for all t whereas because of (B.9),

$$l_n(\widehat{f}_n|_{\alpha}, \alpha) = \sup_{\eta_1 \leq \cdots \leq \eta_{i_1} = \cdots = \eta_{i_2} \leq \cdots \leq \eta_m} \sum_{h=1}^{m} n_{\pi(h)} \left\{ Y_{\pi(h)} \ell(\eta_h) - B \circ \ell(\eta_h) \right\}.$$

The above supremum is taken over a restricted set as compared to (B.10), so we conclude that (B.7) holds with $\tilde{\alpha}$ replace by α_t, for all t. Since $\alpha \mapsto l_n(\widehat{f}_n|_{\alpha}, \alpha)$ achieves its maximum and α_t satisfies (B.8), this implies that there exists a maximizer $\widehat{\alpha}_n$ such that $\widehat{\alpha}_n^T x_i \neq \widehat{\alpha}_n^T x_j$ for all $i \neq j$, and the first claim follows.

Now, let $\widehat{\alpha}_n \in \{\widehat{\alpha}_n\}$ be such that $\widehat{\alpha}_n^T x_i \neq \widehat{\alpha}_n^T x_j$ for all $i \neq j$, and let π be the permutation such that

$$\widehat{\alpha}_n^T x_{\pi(1)} < \widehat{\alpha}_n^T x_{\pi(2)} < \cdots < \widehat{\alpha}_n^T x_{\pi(m)}.$$

Any $\alpha_t \in \mathcal{S}_{d-1}$ that satisfies (B.8), also satisfies (B.10) and therefore, such an α_t also belongs to $\{\widehat{\alpha}_n\}$. This means that $\{\widehat{\alpha}_n\}$ contains the set of all α_t's that satisfy (B.8). The latter is the non-empty intersection of $\mathcal{S}_{d-1}$ and the intersection of the open half-spaces $\{\alpha \in \mathbb{R}^d : \alpha^T(x_{\pi(i)} - x_{\pi(i-1)}) > 0\}$ for all $i \in \{2, \ldots, m\}$ and therefore, it has a strictly positive Lebesgue measure in $\mathcal{S}_{d-1}$. The last claim follows. □

C Proofs and additional results for Section 3.1

We give the details of the proofs under assumption (A1). To extend the result to also hold under assumption (A2), we argue conditionally on $X_1, \ldots, X_n$. This suffices since once the conditional result is proved, the unconditional result follows by application of the dominated convergence theorem. Because n_i is measurable with respect to $X_1, \ldots, X_n$, it can be considered as fixed conditionally on $X_1, \ldots, X_n$. It follows from the strong law of large numbers that under the assumption (A2), n_i/n converges to λ_i almost surely for all i. Hence, (A2) implies that (A1) holds under the conditional probability. Therefore, since all results proved here hold under (A1), they also hold under (A2).

C.1 Proof of Proposition 4

Proof. Under assumption (A1), by the strong law of large numbers, $Y_i \to \mu_{0i}$ $(i = 1, \ldots, m)$, almost surely. Fix $\eta > 0$. Therefore, almost surely, there exists an integer $N_1 > 0$ such that for all $n \geq N_1$,

$$\max_{i=1,\ldots,m} |Y_i - \mu_{0i}| \leq \eta, \tag{C.11}$$

implying that the variables Y_i $(i = 1, \ldots, m)$ are uniformly bounded for $n \geq N_1$. For α fixed, the values of $\widehat{f}_n|_\alpha$ at the points $\alpha^T x_i$ $(i = 1, \ldots, m)$ are the components of the maximizer $\widehat{\mu}_n|_\alpha$ of $\bar{l}_n(\beta)$, where

$$\bar{l}_n(\beta) = \sum_{i=1}^{m} \frac{n_i}{n} \left\{ Y_i \ell(\beta_i) - B \circ \ell(\beta_i) \right\}, \tag{C.12}$$

over the set $\mathcal{M}_\alpha$ of vectors $\beta \in \mathbb{R}^m$ satisfying $\beta_i \leq \beta_j$ whenever $\alpha^T x_i \leq \alpha^T x_j$, for all i, j. For instance, if $\alpha^T x_1 = \alpha^T x_2 < \cdots < \alpha^T x_m$ then $\mathcal{M}_\alpha = \{\beta : \beta_1 = \beta_2 \leq \cdots \leq \beta_m\}$. It follows from Silvapulle and Sen (2005, Proposition 2.4.3, page 51) that the maximizer is unique and equal to the minimizer of

$$\beta \mapsto \sum_{i=1}^{m} \frac{n_i}{n} (Y_i - \beta_i)^2 \tag{C.13}$$

over $\mathcal{M}_\alpha$. All components of the minimizer clearly belong to $[\min_{i=1,\ldots,m} Y_i, \max_{i=1,\ldots,m} Y_i]$. From (C.11), we conclude that $\widehat{\mu}_n|_\alpha$ is the unique minimizer of the criterion in (C.13) over the set $\overline{\mathcal{M}}_\alpha$ of vectors in $\mathcal{M}_\alpha$ that are bounded below by $\min_{i=1,\ldots,m} \mu_{0i} - \eta$ and above by $\max_{i=1,\ldots,m} \mu_{0i} + \eta$, say. We also have

$$\sum_{i=1}^{m} \frac{n_i}{n} (Y_i - \beta_i)^2 \;\; \longrightarrow \;\; \sum_{i=1}^{m} \lambda_i (\mu_{0i} - \beta_i)^2 \tag{C.14}$$

almost surely as $n \to \infty$. Let $\mu_0|_\alpha$ denote the unique minimizer over $\mathcal{M}_\alpha$ of the criterion function given on the right-hand side in the display above. Furthermore, it follows from Silvapulle and Sen (2005, Proposition 2.4.3, page 51) that $\mu_0|_\alpha$ is also the unique maximizer of $\bar{l}(\beta)$ over $\mathcal{M}_\alpha$, where

$$\bar{l}(\beta) = \sum_{i=1}^{m} \lambda_i \left\{ \mu_{0i} \ell(\beta_i) - B \circ \ell(\beta_i) \right\}. \tag{C.15}$$

In addition, $\mu_0|_\alpha$ clearly belongs to $\overline{\mathcal{M}}_\alpha$. Since the space $\overline{\mathcal{M}}_\alpha$ is bounded and closed, hence compact, we conclude that with probablity one, $\bar{l}_n$ converges to $\bar{l}$ uniformly over the set $\overline{\mathcal{M}}_\alpha$, which contains both $\widehat{\mu}_n|_\alpha$ and $\mu_0|_\alpha$. Since $\widehat{\mu}_n|_\alpha$ maximizes $\bar{l}_n$ over $\overline{\mathcal{M}}_\alpha$ we have $\bar{l}_n(\widehat{\mu}_n|_\alpha) \geq \bar{l}_n(\mu_0|_\alpha) = \bar{l}(\mu_0|_\alpha) + o_{a.s.}(1)$ and therefore,

$$\begin{aligned} 0 &\leq \bar{l}(\mu_0|_\alpha) - \bar{l}(\widehat{\mu}_n|_\alpha) \\ &\leq \bar{l}_n(\widehat{\mu}_n|_\alpha) - \bar{l}(\widehat{\mu}_n|_\alpha) + o_{a.s.}(1) \end{aligned}$$

where from the above uniform convergence, the right hand term converges to zero almost surely. Hence, $\bar{l}(\widehat{\mu}_n|_\alpha)$ converges to $\bar{l}(\mu_0|_\alpha)$ almost surely. Furthermore, the function $\beta \mapsto \bar{l}(\beta)$ is continuous on the compact set $\overline{\mathcal{M}}_\alpha$, implying by unicity of the maximizer that for all $\epsilon > 0$,

$$\sup_{\beta \in \overline{\mathcal{M}}_\alpha \ s.t. \ \|\beta - \mu_0|_\alpha\| \geq \epsilon} \bar{l}(\beta) < \bar{l}(\mu_0|_\alpha). \tag{C.16}$$

Hence, the almost sure convergence of $\bar{l}(\widehat{\mu}_n|_\alpha)$ to $\bar{l}(\mu_0|_\alpha)$ implies that for fixed α,

$$\widehat{\mu}_n|_\alpha \to \mu_0|_\alpha, \tag{C.17}$$

almost surely.

Note that the set of possible orders of the projections $\alpha^T x_i, i = 1, \ldots, m$ as α spans $\mathcal{S}^{d-1}$ is finite. Thus, $\mathcal{S}^{d-1}$ can be written as the finite union of sets $S_1, \ldots, S_k$ say, where each of these sets determines a unique ordering of the projections. Also, the functions $\alpha \mapsto \bar{l}_n(\widehat{\mu}_n|_\alpha)$ and its almost sure pointwise limit $\alpha \mapsto \bar{l}(\mu_0|_\alpha)$ take a constant value on $S_j, j = 1, \ldots, k$. It follows that the convergence occurs uniformly on $\mathcal{S}^{d-1}$. This implies in particular that

$$\lim_{n \to \infty} \Big| \sup_{\alpha \in \mathcal{S}^{d-1}} \bar{l}_n(\widehat{\mu}_n|_\alpha) - \sup_{\alpha \in \mathcal{S}^{d-1}} \bar{l}(\mu_0|_\alpha) \Big| = 0$$

almost surely. Now, let $\widehat{\alpha}_n$ be an arbitrary vector in the unit sphere that maximizes $\alpha \mapsto \bar{l}_n(\widehat{\mu}_n|_\alpha)$. Existence is given in Proposition 3. Since by definition $\bar{l}_n(\widehat{\mu}_n|_{\widehat{\alpha}_n}) = \sup_{\alpha \in \mathcal{S}^{d-1}} \bar{l}_n(\widehat{\mu}_n|_\alpha)$ it follows from the triangle inequality that

$$\lim_{n \to \infty} \Big| \bar{l}(\mu_0|_{\widehat{\alpha}_n}) - \sup_{\alpha \in \mathcal{S}^{d-1}} \bar{l}(\mu_0|_\alpha) \Big| = 0$$

almost surely. Using again the fact that $\alpha \mapsto \bar{l}(\mu_0|_\alpha)$ takes constant values, we conclude that we must have

$$\bar{l}(\mu_0|_{\widehat{\alpha}_n}) = \sup_{\alpha \in \mathcal{S}^{d-1}} \bar{l}(\mu_0|_\alpha)$$

almost surely for sufficiently large n, for any maximizer $\widehat{\alpha}_n$. In other words, with probability one and for n sufficiently large, any maximizer of the log-likelihood belongs to the set $\mathcal{S}_0$ of vectors in $\mathcal{S}^{d-1}$ that maximize the limiting criterion $\alpha \mapsto \bar{l}(\mu_0|_\alpha)$.

It remains to describe this set. Since ℓ is strictly increasing with $\ell^{-1} = B'$, it follows that $\beta \mapsto \mu_{0i}\ell(\beta) - B \circ \ell(\beta)$ achieves its maximum at the unique point $\beta = \mu_{0i}$. Hence, for all α we have

$$\bar{l}(\mu_0|_\alpha) \ \leq \ \sum_{i=1}^m \lambda_i \left\{ \mu_{0i}\ell(\mu_{0i}) - B \circ \ell(\mu_{0i}) \right\} = \bar{l}(\mu_0).$$

Equality occurs if and only if $\mu_0 \in \mathcal{M}_\alpha$, hence $\mathcal{S}_0$ is the set of all $\alpha \in \mathcal{S}^{d-1}$ such that $\mu_0 \in \mathcal{M}_\alpha$. □

C.2 Proof of Proposition 5

Proof. By Proposition 4, a.s. there exists an N_0 such that $\{\widehat{\alpha}_n\} \subset \mathcal{S}_0$ for all $n \geq N_0$. Assume then that $n \geq N_0$ and consider $\alpha \in \{\widehat{\alpha}_n\}$. Furthermore, without loss of generality, we can assume that the index $j = 1, \ldots, m$ is such that $z_j = \alpha^T x_j$ is nondecreasing in j. Since $\alpha \in \{\widehat{\alpha}_n\} \subset \mathcal{S}_0$, it follows that μ_{0j} is also increasing in j. It follows that

$$\widehat{\mu}_n \;=\; \mathrm{argmin}_{\beta \in \mathcal{M}_m} \sum_{j=1}^{m} n_j (Y_j - \beta_j)^2.$$

Now, from Robertson et al. (1988, Theorem 1.6.1) follows that

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

since μ_0 is increasing, and the result follows. □

C.3 Simulations studying Proposition 5

Let us consider the following example with only 6 design points in $\mathbb{R}^2$: $x_1 = (1,2)^T$, $x_2 = (-2,2)^T$, $x_3 = (1,-3)^T$, $x_4 = (0,-1)^T$, $x_5 = (1,1)^T$ and $x_6 = (0,0)^T$. Also, assume that the response Y_{ij} for $1 \leq i \leq 6$ and $1 \leq j \leq 100$ is Gaussian with mean $f_0(\alpha_0^T x_i)$ and variance equal to one. We consider two different settings for f_0, α_0 and compare the total relative risk of the MLE of the means under the isotonic single index model with the MLE of the means under the naive model described after Proposition 5. In the latter case, the estimator of the means is simply the naive estimator $Y_i, i = 1, \ldots, 6$ given by the empirical means.

In the first setting, $\alpha_0 = (-1,1)^T/\sqrt{2}$ and $f_0(z) = 0$ if $z < 0.7$ and $f_0(z) = 3$ otherwise. Thus, the true mean in this setting takes only two distinct values: 0 (for $i \in \{3,4,5,6\}$) and 3 (for $\{1,2\}$). In the second setting, $\alpha_0 = (1,4)^T/\sqrt{17}$ and $f_0(z)$ is strictly increasing so that the true mean, after ordering, is equal to $0.1, 0.2, \ldots, 0.6$. For each model, we consider three different values for $n_i = 10, 100, 1000$ and each boxplot is computed based on 200 replications. The total sample sizes considered are therefore $n = 60, 600$ and 6000.

The plots of the boxplots of the quadratic risks are shown on the right in Figure 1. The boxplots clearly indicate that estimation in the monotone single index model yields a smaller risk than the one based on the empirical means for the first setting. In that example, we have considered the case where the underlying link function has flat regions. In the situation where this function is strictly increasing, we know that the MLE coincides with the means Y_i almost surely provided that n is large enough. This is the case for the second setting, where for $n_i = 1000$, the boxplots of the total risk of the single index and naive estimators are found to be almost identical. Interestingly, for smaller sample sizes ($n = 60, 600$), we continue to observe improved performance for the single index model, even

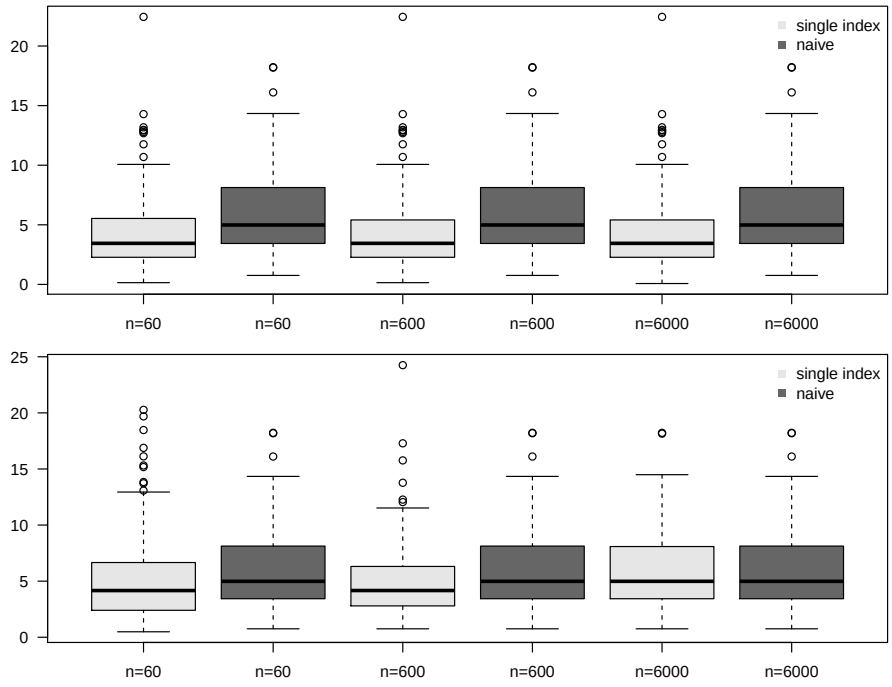

Figure 1: Boxplots comparing the quadratic risk of the single index model estimator with the naive estimator for the first setting (top) and second setting (bottom). The boxplots were produced on the basis of 200 replications.

in this setting. Similar behaviour was already observed for the Grenander estimator in Jankowski and Wellner (2009).

D Proofs for Section 3.2

The proof of our asymptotic results is the most complicated part of the manuscript. For ease of reading we we first state two technical lemmas, and then prove Theorem 1. The technical lemmas are proved afterwards. As for the proofs given in Section C, we prove the results under assumption (A1). The result under the assumption (A2) follows by application of the dominated convergence theorem, as before, noting also that the limiting distribution under (A1) does not depend on $X_1, \ldots, X_n$.

Throughout, we use the following notation. For fixed $\alpha \in \mathcal{S}^{d-1}$, let

$$\widehat{\mu}_n|_\alpha = (\widehat{f}_n|_\alpha(\alpha^T x_1), \ldots, \widehat{f}_n|_\alpha(\alpha^T x_m))^T, \tag{D.18}$$

where $\widehat{f}_n|_\alpha$ is a maximizer of $l_n(f, \alpha)$ over $f \in \mathcal{M}$. Let $\widehat{\mu}_n$ be defined similarly with α replaced by an arbitrary MLE $\widehat{\alpha}_n$ of α_n. Hence, $\widehat{\mu}_n$ is the MLE of μ_0.

D.1 Technical lemmas

Lemma 1. *Under either assumption (A1) or (A2), the sequence of random variables* $\sqrt{n}(\widehat{\mu}_n|_\alpha - \mu_0)$ *is bounded in probability for all fixed* $\alpha \in \mathcal{S}_0$. *The same holds also for* $\sqrt{n}(\widehat{\mu}_n - \mu_0)$.

Lemma 2. *Let* M *be the process defined in* (D.24) *below. Then, the maximizer of* M *over the set* $\mathcal{M}(\mu_0)$ *is almost surely unique.*

D.2 Proof of Theorem 1

Proof. With fixed $\lambda_1, \ldots, \lambda_m$, for all $\alpha \in \mathcal{S}_0$, $\widehat{g}|_\alpha$ depends only on the way the projections $\alpha^T x_i, i = 1, \ldots, m$ are ordered (with ties allowed). As the number of ways is finite, this implies that the function

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

takes a finite number of values on $\mathcal{S}_0$ and therefore, it achieves its maximum over $\alpha \in \mathcal{S}_0$. This proves the first claim.

Below, we denote by $\alpha_1, \ldots, \alpha_r$, with some integer $r \geq 1$, some representative points in $\mathcal{S}_0$ of the finitely many ways the projections $\alpha^T x_i, i = 1, \ldots, m$ can be ordered (with ties allowed). Fix $\epsilon > 0$. By Lemma 1, we can find $K_\epsilon > 0$ such that

$$P\left(\sqrt{n}\left\|\widehat{\mu}_n - \mu_0\right\|_\infty \leq K_\epsilon\right) \geq 1 - \epsilon \tag{D.19}$$

for all $n \geq 1$, where $\|\cdot\|_\infty$ denotes the supremum norm. Next, we recall that a subset $J \in \mathrm{part}(\mu_0)$ is a set of indices in $\{1, \ldots, m\}$ such that for all $i \in J$ we have that $\mu_{0i} = \mu_0^J$, the common value that μ_0 takes on that set. Now, $\mu_{0i} \neq \mu_{0j}$ if and only if $i \in J$ and $j \in J'$ for some J and J' such that $J \neq J'$, or equivalently $J \cap J' = \emptyset$. In the sequel, we use the same notation $\bar{l}_n$ as in (C.12) and the notation $\widehat{\mu}_n|_\alpha$ in (D.18). Hence, $\sqrt{n}(\widehat{\mu}_n|_\alpha - \mu_0)$ is the maximizer of $\bar{l}_n(\mu_0 + n^{-1/2}h)$ over the set of all $h = (h_1, \ldots, h_m)^T \in \mathbb{R}^m$ such that $\mu_{0i} + n^{-1/2}h_i \leq \mu_{0j} + n^{-1/2}h_j$ whenever $\alpha^T x_i \leq \alpha^T x_j$, for all i, j. Moreover, Proposition 4 ensures that

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

where for sufficiently large n, $\widehat{\alpha}_n$ maximizes $\bar{l}_n(\widehat{\mu}_n|_\alpha)$ over $\alpha \in \mathcal{S}_0$. Hence, (D.19) implies that with probability larger than $1 - \epsilon$, $\sqrt{n}(\widehat{\mu}_n - \mu_0)$ maximizes $h \mapsto \bar{l}_n(\mu_0 + n^{-1/2}h)$ over the set $\mathcal{M}_{n\epsilon}(\mu_0)$ of all $h = (h_1, \ldots, h_m)^T \in \mathbb{R}^m$ such that $\|h\|_\infty \leq K_\epsilon$ and for some $k \in \{1, \ldots, r\}$ we have that $\mu_{0i} + n^{-1/2}h_i \leq \mu_{0j} + n^{-1/2}h_j$ whenever $\alpha_k^T x_i \leq \alpha_k^T x_j$, for all i, j. Fix $k \in \{1, \ldots, r\}$ and let $i, j \in \{1, \ldots, m\}$. Modulo a possible exchange of the roles of i and j let us assume that $\alpha_k^T x_i \leq \alpha_k^T x_j$. Since $\alpha_k \in \mathcal{S}_0$ we must have $\mu_{0i} \leq \mu_{0j}$. If $(i, j) \in J^2$ for some $J \in \mathrm{part}(\mu_0)$, then $\mu_{0i} = \mu_{0j}$ so we have the equivalence $\mu_{0i} + n^{-1/2}h_i \leq \mu_{0j} + n^{-1/2}h_j$ if and only if $h_i \leq h_j$. On the other hand if $(i, j) \in J \times J'$ such that $J \neq J'$ then $\mu_{0i} < \mu_{0j}$, implying that $\mu_{0i} + n^{-1/2}h_i \leq \mu_{0j} + n^{-1/2}h_j$ for all

$h \in \mathbb{R}^m$ such that $\|h\|_\infty \leq K_\epsilon$ and n is taken sufficiently large. For all $J \in \mathrm{part}(\mu_0)$ and $\alpha \in \{\alpha_1, \ldots, \alpha_r\}$ define

$$\mathcal{M}_\alpha^J = \left\{h \in \mathbb{R}^m : h_i \leq h_j \text{ whenever } \alpha^T x_i \leq \alpha^T x_j \text{ for } (i,j) \in J^2\right\}.$$

It follows from the reasoning above that for sufficiently large n, the set $\mathcal{M}_{n\epsilon}(\mu_0) = [-K_\epsilon, K_\epsilon]^m \cap \mathcal{M}(\mu_0)$ where

$$\mathcal{M}(\mu_0) = \cup_{k=1}^r \left\{\cap_{J \in \mathrm{part}(\mu_0)} \mathcal{M}_{\alpha_k}^J\right\}.$$

We show now that $\sqrt{n}(\widehat{\mu}_n - \mu_0)$ is equal to the maximizer of a certain process with probability tending to one. To this aim, let us define the process

$$M_n(h) = n\left(\bar{l}_n(\mu_0 + n^{-1/2}h) - \bar{l}_n(\mu_0)\right) \tag{D.20}$$

for $h \in \mathcal{M}(\mu_0)$ defined above. It follows from Silvapulle and Sen (2005, Proposition 2.4.3, page 51), combined with results on isotonic regression recalled in Section A, that M_n admits a maximizer, $\widehat{h}_n$ over $\mathcal{M}(\mu_0)$, which can be shown to be bounded in probability. At the cost of enlarging K_ϵ we have that $P(\|\widehat{h}_n\|_\infty \leq K_\epsilon) \geq 1 - \epsilon$ and we conclude that with probability larger than $1 - 2\epsilon$ and sufficiently large n we have $\widehat{h}_n = \sqrt{n}(\widehat{\mu}_n - \mu_0)$. Letting $\epsilon \to 0$ we can conclude that

$$\lim_{n\to\infty} P\left(\widehat{h}_n = \sqrt{n}(\widehat{\mu}_n - \mu_0)\right) = 1. \tag{D.21}$$

In the sequel we study the limiting behavior of $\widehat{h}_n$.

Now, fix $K > 0$. Since the function ℓ is three times differentiable it follows from the Taylor expansion that for all $h \in [-K, K]^m$, there exist θ_i and θ_i' both lying between μ_{0i} and $\mu_{0i} + n^{-1/2}h_i$ such that

$$\begin{aligned} M_n(h) &= \sum_{i=1}^m n_i Y_i \left\{\ell'(\mu_{0i})n^{-1/2}h_i + \frac{1}{2}\ell''(\mu_{0i})n^{-1}h_i^2 + \frac{1}{6}\ell'''(\theta_i)n^{-3/2}h_i^3\right\} \\ &\quad - \sum_{i=1}^m n_i \left\{(B\circ\ell)'(\mu_{0i})n^{-1/2}h_i + \frac{1}{2}(B\circ\ell)''(\mu_{0i})n^{-1}h_i^2 + \frac{1}{6}(B\circ\ell)'''(\theta_i')n^{-3/2}h_i^3\right\}. \end{aligned}$$

Using the identities $(B\circ\ell)'(x) = x\ell'(x)$ and $(B\circ\ell)''(x) = \ell'(x) + x\ell''(x)$ for all x and

setting $Z_{ni} = n_i^{1/2}(Y_i - \mu_{0i})$, we obtain

$$\begin{aligned} M_n(h) &= \sum_{i=1}^{m} \left(\frac{n_i}{n}\right)^{1/2} \ell'(\mu_{0i}) Z_{ni} h_i + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{n_i}{n}\right)^{1/2} \ell''(\mu_{0i}) Z_{ni} h_i^2 n^{-1/2} \\ &\quad -\frac{1}{2} \sum_{i=1}^{m} \frac{n_i}{n} \ell'(\mu_{0i}) h_i^2 + \sum_{i=1}^{m} \frac{n_i}{6n} Y_i \ell'''(\theta_i) n^{-1/2} h_i^3 - \sum_{i=1}^{m} \frac{n_i}{6n} (B \circ \ell)'''(\theta_i') n^{-1/2} h_i^3 \\ &= -\frac{1}{2} \sum_{i=1}^{m} \ell'(\mu_{0i}) \frac{n_i}{n} \left(h_i - (n_i/n)^{-1/2} Z_{ni}\right)^2 + \frac{1}{2} \sum_{i=1}^{m} \ell'(\mu_{0i}) Z_{ni}^2 + O_p(n^{-1/2}) \\ &= -\frac{1}{2} \sum_{J \in \mathrm{part}(\mu_0)} \ell'(\mu_0^J) \sum_{i \in J} \frac{n_i}{n} \left(h_i - (n_i/n)^{-1/2} Z_{ni}\right)^2 + \frac{1}{2} \sum_{i=1}^{m} \ell'(\mu_{0i}) Z_{ni}^2 + O_p(n^{-1/2}) \end{aligned} \tag{D.22}$$

where $O_p(n^{-1/2})$ does not depend on h (instead, only on K). By the multivariate central limit theorem we have that $Z_{ni}, i = 1, \ldots, m$ converge jointly in distribution to $\widetilde{Z}_i, i = 1, \ldots, m$. Here, $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)^T$ is a Gaussian random vector with mean zero and diagonal covariance matrix whose i-th diagonal element is $\Sigma_{ii} = \phi B'' \circ \ell(\mu_{0i})$ $(i = 1, \ldots, m)$. See for instance (Silvapulle and Sen, 2005, Page 50) for the computation of the variance. Combining this with Assumption (A) we conclude that

$$M_n \Rightarrow M \tag{D.23}$$

in $\ell^\infty([-K, K]^m)$ for all $K > 0$, where

$$M(h) = -\frac{1}{2} \sum_{J \in \mathrm{part}(\mu_0)} \ell'(\mu_0^J) \sum_{i \in J} \lambda_i \left(h_i - \lambda_i^{-1/2} \widetilde{Z}_i\right)^2 + \frac{1}{2} \sum_{i=1}^{m} \ell'(\mu_{0i}) \widetilde{Z}_i^2. \tag{D.24}$$

We are now interested in maximizing M over $\mathcal{M}(\mu_0)$. As the second sum in the display above does not depend on h, maximizing M amounts to minimizing

$$\widetilde{M}(h) = \sum_{J \in \mathrm{part}(\mu_0)} \ell'\left(\mu_0^J\right) \sum_{i \in J} \lambda_i \left(h_i - \lambda_i^{-1/2} \widetilde{Z}_i\right)^2 \tag{D.25}$$

over $\mathcal{M}(\mu_0)$. This can be done via a profiling approach as in Proposition 3 and Proposition 4 . First, we fix $\alpha \in \{\alpha_1, \ldots, \alpha_r\}$. If we write $h^J = \{h_j, j \in J\}$, then it is clear that a vector $\widehat{h}|_\alpha$ minimizes $\widetilde{M}$ over $\cap_{J \in \mathrm{part}(\mu_0)} \mathcal{M}_\alpha^J$ if and only if for all $J \in \mathrm{part}(\mu_0)$, $(\widehat{h}|_\alpha)^J$ minimizes

$$h^J \mapsto \sum_{i \in J} \lambda_i (h_i - \lambda_i^{-1/2} \widetilde{Z}_i)^2 \tag{D.26}$$

over the set of vectors $h^J = \{h_i, i \in J\}$ such that $h_i \leq h_j$ whenever $\alpha^T x_i \leq \alpha^T x_j$, for all $(i, j) \in J^2$. Existence and uniqueness of these minimizers follow from Barlow et al.

(1972, Theorem 2.1). This proves existence and uniqueness of the minimizer $\widehat{h}|_\alpha$ of $\widetilde{M}$ over $\cap_J \mathcal{M}_\alpha^J$ and hence the maximizer of M over the same set.

Now, using the identity $\ell^{-1} = B'$ it holds that

$$\mathrm{var}(\widetilde{Z}_i) = \frac{\phi}{\ell'(\mu_{0i})}$$

for $i = 1, \ldots, m$. Put $Z_i = \phi^{-1/2}(\ell'(\mu_{0i}))^{1/2}\lambda_i^{-1/2}\widetilde{Z}_i \sim \mathcal{N}(0, \lambda_i^{-1})$, $i = 1, \ldots, m$. Thus, for $J \in \mathrm{part}(\mu_0)$, we have

$$\begin{aligned}\sum_{i\in J} \lambda_i((\widehat{h}|_\alpha)_i - \lambda_i^{-1/2}\widetilde{Z}_i)^2 &= \sum_{i\in J} \lambda_i \left((\widehat{h}|_\alpha)_i - \phi^{1/2}(\ell'(\mu_{0i}))^{-1/2} Z_i\right)^2 \\ &= \sum_{i\in J} \lambda_i \phi(\ell'(\mu_{0i}))^{-1} \left(\phi^{-1/2}(\ell'(\mu_{0i}))^{1/2}(\widehat{h}|_\alpha)_i - Z_i\right)^2.\end{aligned}$$

Recalling that $\widehat{h}|_\alpha$ is such that $(\widehat{h}|_\alpha)^J$ is the unique minimizer of the function in (D.26) over the set of all $h^J = \{h_i, i \in J\}$ such that $h_i \le h_j$ whenever $\alpha^T x_i \le \alpha^T x_j$, for all $(i, j) \in J^2$, and noting that $\ell'(\mu_{0i}) = \ell'(\mu_0^J)$ takes the same (positive) value for all $i \in J$, it follows that

$$\sum_{i\in J} \lambda_i((\widehat{h}|_\alpha)_i - \lambda_i^{-1/2}\widetilde{Z}_i)^2 = \phi(\ell'(\mu_0^J))^{-1} \sum_{i\in J} \lambda_i \left((\widehat{g}|_\alpha)_i - Z_i\right)^2$$

where $\widehat{g}|_\alpha$ is such that for $i \in J$,

$$(\widehat{g}|_\alpha)_i = \phi^{-1/2}(\ell'(\mu_0^J))^{1/2}(\widehat{h}|_\alpha)_i, \tag{D.27}$$

which means that $(\widehat{g}|_\alpha)^J$ is the unique minimizer of $\sum_{i\in J} \lambda_i(g_i - Z_i)^2$ over the set of all $g^J = \{g_i, i \in J\}$ such that $g_i \le g_j$ whenever $\alpha^T x_i \le \alpha^T x_j$, for all $(i, j) \in J^2$. Now we can write

$$\begin{aligned}\widetilde{M}(\widehat{h}|_\alpha) &= \phi \sum_{J\in \mathrm{part}(\mu_0)} \sum_{i\in J} \lambda_i \left((\widehat{g}|_\alpha)_i - Z_i\right)^2 \\ &= \phi \sum_{i=1}^m \lambda_i \left((\widehat{g}|_\alpha)_i - Z_i\right)^2\end{aligned}$$

and therefore, using the same notation as in the first claim of the theorem, $\widehat{\alpha}$ maximizes $\alpha \mapsto M(\widehat{h}|_\alpha)$ over $\mathcal{S}_0$. This imples that $\widehat{h}|_{\widehat{\alpha}}$ is a global maximizer of M over $\mathcal{M}(\mu_0)$. By Lemma 2, we know that this maximizer is almost surely unique.

Noting that M is continuous on $\mathbb{R}^m$ and hence on $\mathcal{M}(\mu_0)$ we see that all assumptions of van der Vaart (1998, Corollary 5.58) are met. This allows us to conclude that

$$\widehat{h}_n \Rightarrow \widehat{h}|_{\widehat{\alpha}}.$$

Note that the asymptotic equivalence in (D.21) implies now that $\sqrt{n}(\widehat{\mu}_n - \mu_0) \to_d \widehat{h}|_{\widehat{\alpha}}$. This finishes the proof, using the connection between $\widehat{h}|_{\widehat{\alpha}}$ and $\widehat{g}|_{\widehat{\alpha}}$ given in (D.27) for $i \in J$, with $J \in \mathrm{part}(\mu_0)$. □

D.3 Proofs of the technical lemmas

Proof of Lemma 1. For all $\alpha \in \mathcal{S}^{d-1}$, let

$$\mathcal{M}_\alpha = \big\{h \in \mathbb{R}^m : h_i \le h_j \text{ whenever } \alpha^T x_i \le \alpha^T x_j \text{ for } (i,j) \in \{1,\dots,m\}\big\}.$$

and note that there are finitely many such sets. We can find $r \ge 1$ representative points $\alpha_j \in \mathcal{S}_0$ for $j = 1,\dots,r$ so that

$$\widehat{\mu}_n \in \{\widehat{\mu}_n|_\alpha, \alpha \in \{\alpha_1, \cdots, \alpha_r\}\}. \tag{D.28}$$

It follows from its definition that $\widehat{\mu}_n|_\alpha$ is the unique minimizer of the criterion in (C.13) over $\mathcal{M}_\alpha$. Hence, $\sqrt{n}(\widehat{\mu}_n|_\alpha - \mu_0)$ minimizes

$$h \mapsto \sum_{i=1}^m n_i \left(Y_i - (\mu_{0i} + n^{-1/2}h_i)\right)^2 = \sum_{i=1}^m \left(\sqrt{n_i}(Y_i - \mu_{0i}) - \sqrt{\frac{n_i}{n}}h_i\right)^2$$

over the set of all $h = (h_1,\dots,h_m)^T$ such that $\mu_0 + n^{-1/2}h \in \mathcal{M}_\alpha$. Under assumption (A1), the variance of Y_i is $B'' \circ \ell(\mu_{0i})\phi/n_i$, see e.g. Silvapulle and Sen (2005, Page 50), so $\sqrt{n_i}(Y_i - \mu_{0i})$ has mean zero and finite variance whence it is bounded in probability. This proves that the term on the right hand side of the display above is bounded in probability if and only if $(n_i/n)^{1/2}h_i$ is bounded for all i. Under assumption (A1), this happens if and only if h_i are bounded for all i. If for some i, $h_i \to \infty$ along a subsequence, then the term on the right hand side of the display above goes to infinity along that subsequence with arbitrarily large probability. Hence, the minimum cannot be achieved for such a subsequence, which proves that $\sqrt{n}(\widehat{\mu}_n|_\alpha - \mu_0)$ is bounded in probability for all fixed α. By (D.28), $\widehat{\mu}_n$ takes the form $\widehat{\mu}_n|_\alpha$ for some α in a finite set $\{\alpha_1,\dots,\alpha_r\}$ so we conclude that $\sqrt{n}(\widehat{\mu}_n - \mu_0)$ is bounded in probability. □

Proof of Lemma 2. Let $\tilde{Z} = (\tilde{Z}_1,\dots,\tilde{Z}_m)$ with $\tilde{Z}_i = \lambda_i^{-1/2}Z_i$. As the second sum in (D.24) does not depend on h, maximizing M amounts to minimizing $\widetilde{M}$ in (D.25), so it suffices to show that the minimizer of $\widetilde{M}$ over $\mathcal{M}(\mu_0)$ is almost surely unique. From the proof of Theorem 1, there exists a minimizer $\widehat{h}$ and any minimizer takes the form $\widehat{h} = \widehat{h}|_\alpha$ for some random $\alpha \in \{\alpha_1,\dots,\alpha_r\}$. Hence, there exists $\alpha \in \{\alpha_1,\dots,\alpha_r\}$, where $\alpha_k, k \in \{1,\dots,r\}$, are defined in proof of Theorem 1, such that for all $J \in \mathrm{part}(\mu_0)$, $\widehat{h}^J = \{\widehat{h}_i, i \in J\}$ minimizes

$$h^J \mapsto \sum_{i \in J} \lambda_i (h_i - \tilde{Z}_i)^2$$

over the set of vectors $h^J = \{h_i, i \in J\}$ such that $h_i \le h_j$ whenever $\alpha^T x_i \le \alpha^T x_j$, for all $(i,j) \in J^2$. Let π be any permutation of $\{1,\dots,m\}$ which arranges the values $\alpha^T x_i, i \in J$ in an ascending order for each subset $J \in \mathrm{part}(\mu_0)$. The permutation is not unique if there are $i \ne j$ in a given $J \in \mathrm{part}(\mu_0)$ such that $\alpha^T x_i = \alpha^T x_j$. Consider an arbitrary

$J \in \mathrm{part}(\mu_0)$ and denote $J = \{j_1, \ldots, j_{|J|}\}$, where $j_1 < \cdots < j_{|J|}$. By definition of π, there are indices $m_1 < \cdots < m_K = |J|$ such that

$$\alpha^T x_{\pi(j_1)} = \ldots = \alpha^T x_{\pi(j_{m_1})} < \alpha^T x_{\pi(j_{m_1}+1)} = \ldots = \alpha^T x_{\pi(j_{m_2})} < \ldots < \alpha^T x_{\pi(j_{m_{K-1}+1})} = \ldots = \alpha^T x_{\pi(j_{m_K})},$$

and the vector $(\widehat{h}_{\pi(j_1)}, \ldots, \widehat{h}_{\pi(j_{|J|})})^T$ minimizes the rearranged sum

$$(\beta_{j_1}, \ldots, \beta_{j_{|J|}})^T \mapsto \sum_{i=1}^{|J|} \lambda_{\pi(j_i)} (\beta_{j_i} - \tilde{Z}_{\pi(j_i)})^2$$

over the set of all vectors $(\beta_{j_1}, \ldots, \beta_{j_{|J|}})^T$ such that

$$\beta_{j_1} = \ldots = \beta_{j_{m_1}} \leq \beta_{j_{m_1}+1} = \ldots = \beta_{j_{m_2}} \leq \ldots \leq \beta_{j_{m_{K-1}}} = \ldots = \beta_{j_{m_K}}.$$

Hence, $(\widehat{h}_{\pi(j_1)}, \ldots, \widehat{h}_{\pi(j_{|J|})})^T$ is the solution of an isotonic regression problem with equality constraints, see Section A. This implies that it can be written in the form $A_J(\tilde{Z}_{\pi(j_1)}, \ldots, \tilde{Z}_{\pi(j_{|J|})})^T$ where A_J is a $|J| \times |J|$-matrix that satisfies the conditions in (A.3) with A and k are replaced by A_J and $|J|$ respectively. Putting together all possible $J \in \mathrm{part}(\mu_0)$, and using that the permutation π only changes the order of indices in each subset $J \in \mathrm{part}(\mu_0)$, we conclude that there exists an $m \times m$ matrix $\widetilde{A}$ such that

$$(\widehat{h}_{\pi(1)}, \ldots, \widehat{h}_{\pi(k)})^T = \widetilde{A}(\tilde{Z}_{\pi(1)}, \ldots, \tilde{Z}_{\pi(k)})^T$$

where $\widetilde{A}$ satisfies the conditions in (A.3) with A and k replaced by $\widetilde{A}$ and m respectively. Now, let P be the permutation matrix such that $P_{ij} = 1$ if $\pi(i) = j$ and $P_{ij} = 0$ otherwise, for $(i, j) \in \{1, \ldots, m\}$. Note that we can have $P_{ij} = 1$ only if $(i, j) \in J^2$ for some $J \in \mathrm{part}(\mu_0)$. For all $i \in \{1, \ldots, m\}$, we have $\tilde{Z}_{\pi(i)} = P_i \tilde{Z}$ and $\widehat{h}_{\pi(i)} = P_i \widehat{h}$, with P_i the i-th row of P, whence $P\widehat{h} = \widetilde{A} P \widetilde{Z}$, implying that

$$\widehat{h} = \widehat{A}\widetilde{Z} \tag{D.29}$$

where $\widehat{A} = P^T \widetilde{A} P$. Since P is a permutation matrix, $\widehat{A}$ is obtained from $\widetilde{A}$ by permuting the rows and the columns, so the conditions in (A.3) still hold with A and k replaced by $\widehat{A}$ and m respectively.

On the other hand, it follows from the above characterisation of $\widehat{h}^J$ together with (A.6) that for all $J \in \mathrm{part}(\mu_0)$,

$$\sum_{i \in J} \lambda_i \widehat{h}_i (\widehat{h}_i - \tilde{Z}_i) = 0,$$

implying that

$$\sum_{i \in J} \lambda_i (\widehat{h}_i - \tilde{Z}_i)^2 = \sum_{i \in J} \lambda_i \tilde{Z}_i (\tilde{Z}_i - \widehat{h}_i).$$

Combining with (D.29) yields

$$\begin{aligned}\widetilde{M}(\widehat{h}) &= \sum_{i=1}^{m} \ell'(\mu_{0i})\,\lambda_i \tilde{Z}_i(\tilde{Z}_i - \widehat{h}_i) \\ &= \widetilde{Z}^T \Lambda(I - \widehat{A})\widetilde{Z} \end{aligned} \tag{D.30}$$

where Λ is the diagonal matrix with diagonal terms $\ell'(\mu_{0i})\lambda_i, i = 1, \ldots, m$. Suppose that there exist a minimizer $\widetilde{h} \neq \widehat{h}$ of $\widetilde{M}$ over $\mathcal{M}(\mu_0)$. Similar to (D.29) and (D.30) above, there exists a random $m \times m$-matrix $\widehat{B}$ that satisfies the conditions in (A.3) with k and A replaced by m and $\widehat{B}$ respectively, such that

$$\widetilde{h} = \widehat{B}\widetilde{Z} \quad \text{and} \quad \widetilde{M}(\widetilde{h}) = \widetilde{Z}^T \Lambda(I - \widehat{B})\widetilde{Z}. \tag{D.31}$$

Since $\widehat{h} \neq \widetilde{h}$, we have $\widehat{A} \neq \widehat{B}$. Since both $\widehat{h}$ and $\widetilde{h}$ minimize $\widetilde{M}$, it follows from (D.30) and (D.31) that

$$\widetilde{Z}^T \Lambda(\widehat{B} - \widehat{A})\widetilde{Z} = 0.$$

Let $\mathcal{C}$ denote the finite collection of possible matrices $C \neq 0$ which can be written in the form $C = \Lambda(A - B)$, where A and B are fixed $m \times m$ matrices that satisfy the conditions in (A.3). From what precedes we have

$$\begin{aligned}\mathrm{pr}\left(\textit{The minimizer of } \widetilde{M} \textit{is not unique}\right) &\leq \mathrm{pr}\left(\widetilde{Z}^T \widehat{C} \widetilde{Z} = 0 \textit{ for some } \widehat{C} \in \mathcal{C}\right) \\ &\leq \sum_{C \in \mathcal{C}} \mathrm{pr}(\widetilde{Z}^T C \widetilde{Z} = 0). \end{aligned} \tag{D.32}$$

Let G be a standard Gaussian vector in $\mathbb{R}^m$ and Σ be the variance matrix of $\widetilde{Z}$, that is the diagonal matrix with diagonal terms $\lambda_i^{-1}\phi B'' \circ \ell(\mu_{0i}) > 0$ $(i = 1, \ldots, m)$. Hence, the Gaussian vector $\widetilde{Z}$ has the same distribution as $\Sigma^{1/2}G$. Consider an arbitrary $C = \Lambda(A - B) \in \mathcal{C}$ and let $S = (C + C^T)/2$. Since $(\widetilde{Z}^T C \widetilde{Z})^T = \widetilde{Z}^T C^T \widetilde{Z}$ we have $\widetilde{Z}^T C \widetilde{Z} = \widetilde{Z}^T S \widetilde{Z}$, and hence

$$\mathrm{pr}(\widetilde{Z}^T C \widetilde{Z} = 0) = \mathrm{pr}(G^T \Sigma^{1/2} S \Sigma^{1/2} G = 0).$$

Furthermore, $\Sigma^{1/2} S \Sigma^{1/2}$ is symmetric and therefore we can find an orthogonal matrix U and a diagonal matrix Δ such that $\Sigma^{1/2} S \Sigma^{1/2} = U \Delta U^T$. Since $U^T G$ has the same distribution as G we can write

$$\begin{aligned}\mathrm{pr}(\widetilde{Z}^T C \widetilde{Z} = 0) &= \mathrm{pr}(G^T U \Delta U^T G = 0) = \mathrm{pr}(G^T \Delta G = 0) \\ &= \mathrm{pr}\left(\sum_{i:\delta_i > 0} \delta_i G_i^2 = \sum_{i:\delta_i < 0} |\delta_i| G_i^2\right), \end{aligned}$$

where $\delta_i\,(i = 1, \ldots, m)$ are the eigenvalues of Δ. We next show that $S \neq 0$. This will imply that at least one of the above eigenvalues is non-zero, from which it follows that the

above probability is zero. Since $C \in \mathcal{C}$ was arbitrarily chosen, combining this with (D.32) completes the proof of the lemma.

It remains to show that $S \neq 0$. Arguing by contradiction, suppose that $S = 0$. Then, $C^T = -C$; i.e.

$$\Lambda B + B^T \Lambda = \Lambda A + A^T \Lambda. \tag{D.33}$$

Hence,

$$\lambda_i B_{ij} + \lambda_j B_{ji} = \lambda_i A_{ij} + \lambda_j A_{ji}, \text{for all } (i,j) \in \{1,\dots,m\}^2.$$

For a given $i \in \{1,\dots,m\}$, let J_i be the set of indices $j \in \{1,\dots,m\}$ such that $A_{ji} \neq 0$. Since A and B satisfy the conditions in (A.3), for all $j \notin J_i$ we have $A_{ij} = A_{ji} = 0$ so it follows from the equality in the above display that $B_{ij} = B_{ji} = 0$. Exchanging the roles of A and B we conclude that J_i is also the set of indices $j \in \{1,\dots,m\}$ such that $B_{ji} \neq 0$. Denoting by a_i and b_i the common value of the non null entries A_{ji} and B_{ji} for $j \in \{1,\dots,m\}$, respectively, and summing up the equality in the above display over $j \in \{1,\dots,m\}$, we obtain

$$\lambda_i + b_i \sum_{j \in J_i} \lambda_j = \lambda_i + a_i \sum_{j \in J_i} \lambda_j.$$

But J_i is non empty set since $i \in J_i$ and therefore, $b_i = a_i$. As i was arbitrarily chosen, we conclude that $A = B$ and also $C = 0$. As this is not true by assumption, this completes the proof. □

E Additional notes and proofs for Section 3.3

E.1 Proofs

Proof of Proposition 6. We use the same notations $\widehat{g}|_\alpha$, $\widehat{\alpha}$ and $\widehat{h}|_{\widehat{\alpha}}$ as in Theorem 1. Let $\widetilde{Z} = (\widetilde{Z}_1,\dots,\widetilde{Z}_m)^T$ be a Gaussian random vector with mean zero and diagonal covariance matrix whose i-th diagonal element is $\Sigma_{ii} = \phi B'' \circ \ell(\mu_{0i}) = \phi/\ell'(\mu_{0i})$ $(i = 1,\dots,m)$. In the course of the proof of Theorem 1, it is proved that $\sqrt{n_i}(Y_i - \mu_{i0}) \to_d \widetilde{Z}_i$ for all $i = 1,\dots,m$ and $\sqrt{n}(\widehat{\mu}_n - \mu_0) \to_d \widehat{h}|_{\widehat{\alpha}}$. It can be seen from that proof that the weak convergences happen jointly and therefore,

$$D_n^2 \to_d \sum_{i=1}^m \lambda_i \ell'(\mu_{0i})(\lambda_i^{-1/2}\widetilde{Z}_i - (\widehat{h}|_{\widehat{\alpha}})_i)^2.$$

Setting $Z_i = \phi^{-1/2}(\ell'(\mu_{0i}))^{1/2}\lambda_i^{-1/2}\widetilde{Z}_i \sim \mathcal{N}(0,\lambda_i^{-1})$ for $i = 1,\dots,m$, we see that

$$\sum_{i=1}^m \lambda_i \ell'(\mu_{0i})(\lambda_i^{-1/2}\widetilde{Z}_i - (\widehat{h}|_{\widehat{\alpha}})_i)^2 \;=\; \phi \sum_{i=1}^m \lambda_i (Z_i - (\widehat{g}|_{\widehat{\alpha}})_i)^2$$

which completes the proof. □

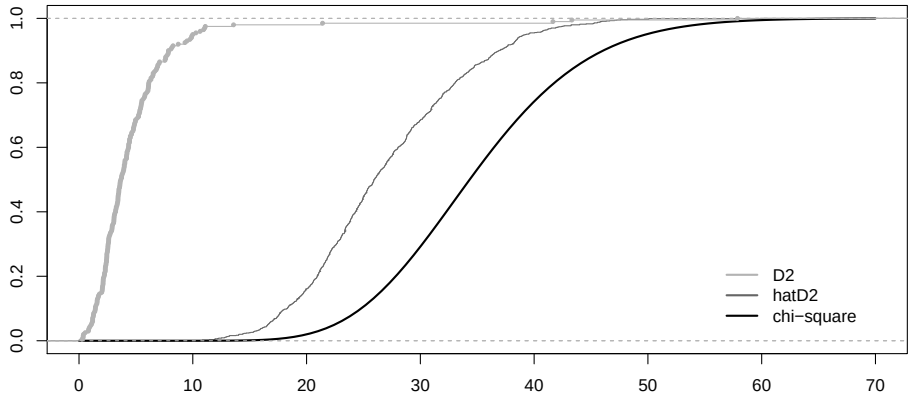

Figure 2: Plot showing the stochastic order of the distribution of $\widehat{D}^2$ for Example 1 in the main paper and that of D^2 and the chi-square with $m-1=35$ degrees of freedom.

Proof of Proposition 7. Recall that D^2 is the result of minimizing

$$\sum_{i=1}^{m}\lambda_i(g_i-Z_i)^2=\sum_{J\in\mathrm{part}(\mu_0)}\sum_{i\in J}\lambda_i(g_i-Z_i)^2.$$

This is done by minimizing each term $\sum_{i\in J}\lambda_i(g_i-Z_i)^2$ over the set of vectors g_J of length $|J|$ such that $g_{i,J}\leq g_{j,J}$ whenever $\alpha^T x_i\leq\alpha^T x_j$, for all $(i,j)\in J\times J$, and then minimizing over $\alpha\in\mathcal{S}_0$. Define $\overline{g}$ such that $\overline{g}_i=m^{-1}\sum_{j=1}^{m}Z_j=\overline{Z}$ for all $i=1,\ldots,m$. Since $\overline{g}$ is a constant vector in i, any subvector $\overline{g}_J$ satisfies the condition that $g_{i,J}\leq g_{j,J}$ whenever $\alpha^T x_i\leq\alpha^T x_j$, for all $(i,j)\in J\times J$. Most importantly, this is true for any choice of α. We therefore have that $D^2\leq\sum_{i=1}^{m}\lambda_i(\overline{g}-Z_i)^2$. Lastly, note that

$$\sum_{i=1}^{m}\lambda_i(\overline{g}-Z_i)^2\;\;=\;\;Z^TA_m\Lambda A_mZ,$$

whereΛ is the matrix with diagonal elements equal to λ_i, and $\{A_m\}_{i,j}=\delta_{i,j}-1/m$. If λ_i is constant (and hence equals $\lambda_i=1/m$), we have that $Z^TA_m\Lambda A_mZ=(Z/\sqrt{m})^TA_m(Z/\sqrt{m})$, which is chi-squared with $(m-1)$ degrees of freedom. □

E.2 Estimating the dispersion parameter

There are several well-known options available to estimate the dispersion parameter, see for example McCullagh and Nelder (1989). One option is by maximizing the full likelihood function from (2) in the saturated model

$$\widetilde{l}_n(\phi,\mu)\;\;=\;\;\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log c(Y_{ij},\phi)+\phi^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{Y_{ij}\ell(\mu_i)-B\circ\ell(\mu_i)\right\},$$

where for all $i = 1, \ldots, m$, the variables Y_{ij}, $j = 1, \ldots, n_i$ correspond to the observations from group i and $\mu = (\mu_1, \ldots, \mu_m)^T \in \mathbb{R}^m$. This results in a maximum likelihood estimator of ϕ given by $\widehat{\phi}_n$ as the solution to the score equation

$$\widehat{\phi}_n^2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\partial_\phi c(Y_{ij}, \phi)|_{\phi=\widehat{\phi}_n}}{c(Y_{ij}, \widehat{\phi}_n)} \quad = \quad \sum_{i=1}^{m} n_i \left\{ Y_i \ell(Y_i) - B \circ \ell(Y_i) \right\}$$

where Y_i is the average of the variables Y_{ij}, $j = 1, \ldots, n_i$. In the Gaussian setting, this results in the standard estimate

$$\widehat{\phi}_n = n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_i)^2$$

which converges in probability to $\phi = \mathrm{var}(Y_{ij})$ in that case. This is just one option, and we do not advocate for one method over another, and the user is free to use their favourite consistent estimator. For the binomial and Poisson families, the dispersion parameter is known and need not be estimated.

E.3 Histogram of distribution of $\widehat{D}^2$.

See Figure 2.

F Additional simulations

F.1 For Section 4.1

Example 3

Our third example is similar to Example 2 in the main mansucript, except that now we take Y_{ij} to be independent Poisson variables, with rate

$$\mu_0(x_i) \quad = \quad \log\left(h_0(\alpha_0^T x_i)\right)$$

and consider three different cases of h_0. As in Example 2, 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 to 100.

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

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

Table 1: Estimation of α_0 in Example 3: 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
		(A)	0.618	0.692
	3	(B)	0.494	0.722
		(D)	0.864	0.071
		(A)	0.647	0.731
(i)	10	(B)	0.544	0.757
		(D)	0.816	0.027
		(A)	0.703	0.709
	100	(B)	0.697	0.683
		(D)	0.846	-0.019
		(A)	0.914	0.252
	3	(B)	0.718	0.576
		(D)	0.919	0.100
		(A)	0.977	0.207
(ii)	10	(B)	0.740	0.573
		(D)	n/a	n/a
		(A)	0.980	0.200
	100	(B)	0.722	0.648
		(D)	n/a	n/a
		(A)	-0.476	-0.340
	3	(B)	-0.282	-0.436
		(D)	0.810	-0.054
		(A)	-0.794	-0.400
(iii)	10	(B)	-0.556	-0.609
		(C)	0.797	-0.028
		(A)	-0.924	-0.368
	100	(B)	-0.709	-0.660
		(D)	0.841	-0.011

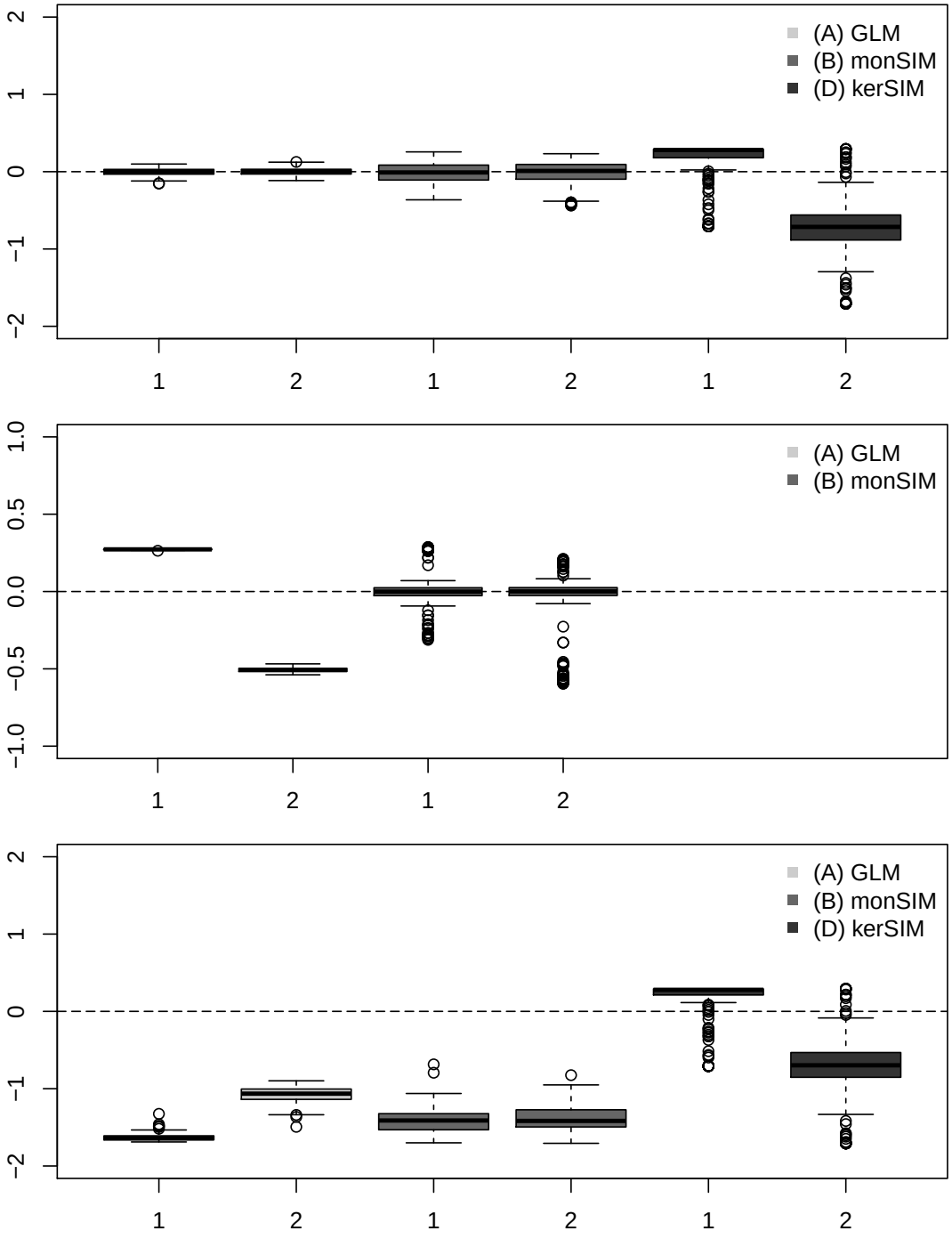

Figure 3: Estimates of α_0 for Example 3 for methods (A)-(C) over $B = 200$ simulations. Boxplots of $\widehat{\alpha}_{n,i} - \alpha_{0i}$ are shown, with i appearing on the x-axis. From top to bottom, the plots show cases (i)-(iii) for $n_i = 100$.

Table 2: Estimation of the response surface in Example 3 mean values of sum of squared errors at observed predictors across $B = 200$ simulations.

case	n_i	(A)	(B)	(C)	(D)
	3	2.984	6.005	10.98	48.934
(i)	10	1.009	1.886	3.346	59.297
	100	0.1	0.258	0.346	62.336
	3	17.845	5.755	12.223	48.453
(ii)	10	15.602	1.765	3.622	n/a
	100	15.046	0.185	0.366	n/a
	3	3.148	5.429	11.554	9.971
(iii)	10	1.57	1.667	3.378	15.921
	100	1.033	0.446	0.345	14.969

which we rewrite as $\widehat{\mu}_n(x) = \widehat{\beta}_{n,0} + \|\widehat{\beta}_{n,1,2}\|\ \widetilde{\alpha}_n^T x$ to obtain the GLM estimate of α_0, $\widetilde{\alpha}_n = (\widehat{\beta}_{n,1}, \widehat{\beta}_{n,2})^T / \|\widehat{\beta}_{n,1,2}\|$, with $\|\widehat{\beta}_{n,1,2}\|^2 = \widehat{\beta}_{n,1}^2 + \widehat{\beta}_{n,2}^2$.

(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

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

case	n_i	(A)	(B)	(D)
	3	15.224	55.861	281.336
(i)	10	4.521	27.006	394.649
	100	0.485	11.969	420.416
	3	127.356	43.574	511.517
(ii)	10	114.2	20.744	n/a
	100	115.157	7.512	n/a
	3	42.663	47.589	135.109
(iii)	10	33.696	25.801	213.349
	100	32.046	17.548	193.921

and Racine, 2008).

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

(i). $h_0(x) = 2x$. 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). $h_0(x) = |2x-1|$. With this choice of h_0, only model (C) is well-specified.

The functions were modified slightly from those used in Example 2 so that the mean responses are better suited for a simulation under the Poisson model. The results of our simulations follow, and are very similar to the results for Example 3, however, their shape has been preserved. Notably, the package `np` did not return results in several of the iterations for case (ii), so these have been omitted. Figure 3 and Table 1 summarize the results for estimation of α_0. As before, the monotone SIM does a good job of estimating α_0 when well-specified; and shows a remarkably small loss of efficiency against the correctly specified linear model in case (i). Table 2 considers estimation of the response and Table 3 considers prediction of the response surface. Again, we observe the monotone SIM performing well when correctly specified.

F.2 For Section 4.2

F.2.1 Example 3

As in the main manuscript, we repeat the goodness-of-fit test for all settings of Example 3. Results are given in Table 4 and show good performance of the test. Again, we added an additional setting, case (iv), with flat response to show how the test behaves when in a setting where it is not conservative, and the results are consistent with a level of 0.05.

F.2.2 More on goodness-of-fit

For our more detailed study of the effect of the true link function in the behavior of the goodness-of-fit test, we focus on the Gaussian setting under fixed design with variance $\sigma^2 = 1$. In this case, $\phi = 1$ need not to be estimated hence $\widehat{\phi}_n = 1$ in (10). Different scenarios were considered for the true link function f_0 : (a) strictly monotone with $f_{0,1}(x) = x^3$, (b) constant with $f_{0,2}(x) = 0$, (c) monotone piecewise constant $f_{0,3}$ and (d) non-monotone $f_{0,4}(x) = x^2$, and $f_{0,5}$, where

$$f_{0,3}(x) = \begin{cases} -1, & \text{if } x \le -1/2 \\ 0, & \text{if } -1/2 < x < 0 \\ 1, & \text{if } x \ge 0, \end{cases} \qquad f_{0,5}(x) = \begin{cases} 0, & \text{if } x \le -1/2 \\ -1, & \text{if } -1/2 < x < 0 \\ 1, & \text{if } x \ge 0. \end{cases}$$

We consider three different sample sizes $n \in \{100, 1000, 10000\}$ with $m = 10$ unique data points and $n_i = n/m$ $(i = 1, \ldots, m)$. For $d = 2$ and $d = 5$, the fixed design

Table 4: Significance level or power estimates for goodness-of-fit test for Example 3. 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	
		n_i	level	mean p-value	level	mean p-value
Example 3	case (i)	3	0.05	0.594	0.005	0.875
		10	0.015	0.643	0	0.904
		100	0.05	0.861	0	0.977
	case (ii)	3	0.015	0.669	0.005	0.913
		10	0.025	0.669	0	0.908
		100	0.010	0.709	0.005	0.927
	case (iii)	3	0.065	0.49	0	0.823
		10	0.105	0.443	0	0.776
		100	0.750	0.079	0.375	0.254
		1000	1	0	1	0
	case (iv)	3	0.045	0.511	0	0.843
		10	0.060	0.490	0	0.832
		100	0.025	0.486	0	0.836

matrices are given at the end of this section. Although we consider only the setting of fixed design, in order to ensure that the design points are in general position, the values were generated from the uniform distribution on $[-1, 1]^d$. The design matrices were then fixed throughout the remaining simulations. The responses Y_{ij} were drawn independently from $\mathcal{N}(f_0(\alpha_0^T x_i), 1)$ $(j = 1, \ldots, n_i; i = 1, \ldots, m)$. We also consider two different values for the true index in each dimension. For $d = 2$, we take $\alpha_0^{(1)} = (1, \sqrt{2})/\sqrt{3}$ and $\alpha_0^{(2)} = (1, 0)$; $\alpha_0^{(1)} = (-2, 1, -1, 3, -4)/\sqrt{31}$ and $\alpha_0^{(2)} = (-1, 0, 1, 0, 0)/\sqrt{2}$ for $d = 5$. The estimated probabilities of rejecting the null hypothesis were obtained using 100 replications. Throughout, the level is $A = 0.05$. The quantile $\widehat{q}_{n,1-A}$, with $\widehat{\phi}_n = 1$, was estimated based on 200 replications.

Table 5 gives simulation results for the estimated Type I error for $f_{0,i}$ $(i = 1, 2, 3)$ and power of the hypothesis test for $f_{0,i}$ $(i = 4, 5)$. For the first two link functions $f_{0,1}$ and $f_{0,2}$, we know that asymptotically the test has level at most $A = 0.05$, and this is consistent with the simulation results. The results are similar for $f_{0,3}$, not contradicting our conjecture that the constant case is least favourable. When the true link function is constant, the test is asymptotically exact, which we also see in the simulation results since the estimates for the

Table 5: Significance level and power estimates for Goodness-of-Fit test. The estimated standard error is no larger than 0.037 for link functions $f_{0,1}, f_{0,2}$ and $f_{0,3}$, and no larger than 0.05 otherwise.

		$d=2$		$d=5$	
True link runction	Sample size n	$\alpha_0^{(1)}$	$\alpha_0^{(2)}$	$\alpha_0^{(1)}$	$\alpha_0^{(2)}$
$f_{0,1}$	100	0.03	0.04	0.16	0.05
	1000	0.01	0.01	0.09	0.02
	10000	0.00	0.00	0.02	0.04
$f_{0,2}$	100	0.09	0.03	0.11	0.07
	1000	0.04	0.06	0.03	0.06
	10000	0.04	0.05	0.07	0.06
$f_{0,3}$	100	0.07	0.09	0.06	0.03
	1000	0.10	0.06	0.09	0.00
	10000	0.03	0.07	0.05	0.00
$f_{0,4}$	100	0.71	0.04	0.11	0.62
	1000	1.00	0.34	0.65	1.00
	10000	1.00	1.00	0.99	1.00
$f_{0,5}$	100	0.94	0.51	0.32	0.88
	1000	1.00	1.00	0.29	1.00
	10000	1.00	1.00	0.37	1.00

Type I error are reasonably close to the nominal level A. The link functions $f_{0,4}$ and $f_{0,5}$ are not increasing, allowing us to study the power of the test. Indeed, the test appears to be consistent, except for the setting $f_{0,5}, d = 5, \alpha_0^{(1)}$. This happens since even though $f_{0,5}$ is not increasing, there exist many $\widetilde{\alpha}_0$ such that $f_{0,5}(z_{(i)})$ is increasing, if $z_{(i)}$ denotes the ordered values of $\widetilde{\alpha}_0^T x_i \, (i = 1, \ldots, m)$.

The design points $x_i \, (i = 1, \ldots, 10)$ used in the simulations correspond to the rows of

X_2 for $d = 2$ and X_5 for $d = 5$, where

$$X_2 = \begin{pmatrix} 0.358 & 0.950 \\ -0.394 & 0.972 \\ 0.666 & -0.966 \\ -0.733 & 0.103 \\ -0.650 & 0.472 \\ -0.706 & 0.223 \\ -0.534 & -0.597 \\ -0.698 & -0.794 \\ 0.083 & 0.906 \\ -0.809 & -0.270 \end{pmatrix},$$

$$X_5 = \begin{pmatrix} 0.732 & -0.804 & 0.570 & -0.668 & -0.595 \\ -0.395 & 0.641 & -0.356 & -0.754 & 0.861 \\ 0.775 & -0.780 & 0.430 & 0.306 & -0.016 \\ 0.140 & 0.348 & -0.396 & -0.877 & 0.056 \\ -0.728 & -0.719 & 0.870 & 0.514 & -0.220 \\ -0.849 & 0.672 & 0.218 & -0.851 & -0.313 \\ -0.834 & 0.776 & 0.038 & 0.792 & 0.847 \\ -0.528 & -0.631 & 0.683 & -0.160 & 0.342 \\ 0.762 & -0.364 & -0.718 & -0.085 & -0.697 \\ -0.724 & 0.325 & -0.734 & -0.293 & -0.571 \end{pmatrix}.$$

G Advice on picking design

Proposition 3 in the main manuscript has important practical implications. It states that, for a fixed sample size, only information up to the set $\{\widehat{\alpha}_n\}$ can possibly be learned about the true index.

Naturally, the information we can learn is additionally brought by the response, and asymptotically, this is the same as μ_0. However, if μ_0 is strictly increasing, the only limitation on information comes from the design for larger sample sizes.

The most visible way through which the design reveals information about the index would be the number of possible regions for $\{\widehat{\alpha}_n\}$. Roughly speaking, the more possible regions the more detailed the information one can gain, however, the shape and size of the regions are also factors. Figure 4 gives an example with $m = 5$ unique design points, which yields 20 possible regions for $\{\widehat{\alpha}_n\}$ to choose from. On the other hand, the leukemia example from Section 5, has $m = 16$ unique design points, which could yield up to $m(m-1) = 240$ possible regions for $\{\widehat{\alpha}_n\}$. However, the design points in this data set are linearly dependent, and only 128 possible regions are provided by the design. This limits the possible information about the index; see Figure 5. Similar examples with a linearly dependent components to the design are discussed in Wan et al. (2017). Although popular in practice, designs such as these with many linearly dependent points may not be optimal

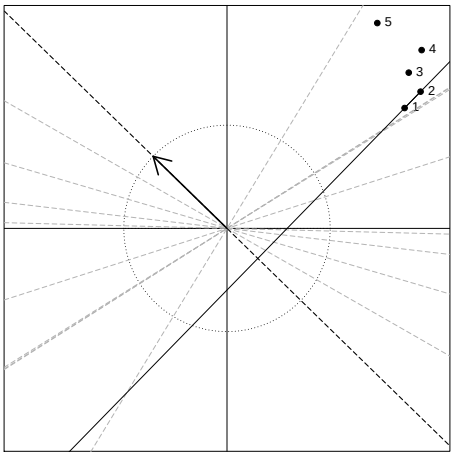

Figure 4: A graph showing $m = 5$ design points $x_1, \ldots, x_5$ (solid dots). The index α (solid arrow) is perpendicular to the vector $x_1 - x_2$. The $m(m-1) = 20$ regions of all linearly inducible permutations are separated by dashed lines, although only 16 regions are easily distinguishable.

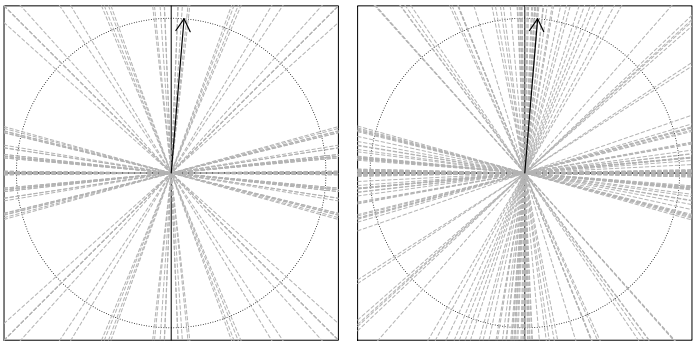

Figure 5: Left plot shows the 86 possible regions for $\{\widehat{\alpha}_n\}$ for the original $m = 16$ data points, and the arrow shows an element of the fitted $\{\widehat{\alpha}_n\}$. The right plot shows the 240 possible regions when the original $m = 16$ data points are randomly perturbed, so that they are no longer linearly dependent.

for the single index model. Additionally, the design informs not only about the number of possible sets, but also their shape and size. These can be visualized apriori to conduct the study as in Figure 4 or Figure 5. Alas, this is easiest when $d = 2$.

Our theoretical developments focus on discrete design, however, the relationship between the design and index estimates applies to the isotonic single index model MLE under any design (such as the continuous setting considered in Chen and Samworth (2016)), as this is a finite sample property.

References

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

CHEN, Y. and SAMWORTH, R. (2016). Generalised additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B* **78** 729–754.

HAYFIELD, T. and RACINE, J. S. (2008). Nonparametric econometrics: The np package. *J. Stat. Softw.* **27**.

JANKOWSKI, H. K. and WELLNER, J. A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman & Hall.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester.

SILVAPULLE, M. J. and SEN, P. K. (2005). *Constrained statistical inference*. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. Inequality, order, and shape restrictions.

TURNER, R. (2015). *Iso: Functions to Perform Isotonic Regression*. R package version 0.0-17.
URL `https://CRAN.R-project.org/package=Iso`

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

WAN, Y., DATTA, S., LEE, J. J. and KONG, M. (2017). Monotonic single-index models to assess drug interactions. *Stat. Med.* **36** 655–670.