Supplementary material

This supplementary material is organised as follows. In Section S1 we include further details of our algorithm and the proofs of results in Sections 2 & 3. The proofs of Theorems 5 and 6 along with a number of lemmas they require can be found in Section S2. Section S3 contains information regarding simulation settings and additional results for the experiments in Section 6 of the main paper.

S1 Additional algorithmic details

S1.1 Remarks on constrained and unconstrained formulations of the univariate objective

It is clear why the identifiability constraint (8) is important when we consider the multivariate problem in Section 3.2. However, for the univariate problem, both constrained and unconstrained formulations of the objective can be clearly defined:

$$\hat{\boldsymbol{\theta}}^{\mathrm{c}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \frac{1}{2}\sum_{k=1}^{K} w_k \left(\bar{Y}_k - \hat{\mu} - \theta_k\right)^2 + \sum_{k=1}^{K-1} \rho(\theta_{(k+1)} - \theta_{(k)}), \tag{30}$$

$$\hat{\boldsymbol{\theta}}^{\mathrm{u}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^K} \frac{1}{2}\sum_{k=1}^{K} w_k \left(\bar{Y}_k - \theta_k\right)^2 + \sum_{k=1}^{K-1} \rho(\theta_{(k+1)} - \theta_{(k)}). \tag{31}$$

As discussed in Section 3.1.1, we can enlarge the feasible set in (30) to be all of $\mathbb{R}^K$: similarly to the observation that $\sum_k w_k \hat{\theta}^{\mathrm{u}}_k = \hat{\mu} = \sum_k w_k \bar{Y}_k$, the minimiser of (30) over all of $\mathbb{R}^K$ will always be in Θ. This can be shown by following the argument at the beginning of the proof of Lemma 10. Therefore the algorithm defined in Section 3.1 can also be applied to the unconstrained formulation of the objective.

It is clear that these problems are essentially identical, as $\hat{\boldsymbol{\theta}}^{\mathrm{u}}$ is a minimiser of the unconstrained objective if and only if $\hat{\boldsymbol{\theta}}^{\mathrm{u}} - \hat{\mu}\mathbf{1}$ is a minimiser of the constrained objective. Observe that while $\hat{\boldsymbol{\theta}}^{\mathrm{u}} \in \mathbb{R}^K$, the solution to the constrained objective is in fact $(\hat{\mu}, \hat{\boldsymbol{\theta}}^{\mathrm{c}}) \in \mathbb{R} \times \Theta$, which is the same K-dimensional space only with a different parametersation. In particular, $\hat{\boldsymbol{\theta}}^{\mathrm{c}}$ is non-unique if and only if $\hat{\boldsymbol{\theta}}^{\mathrm{u}}$ is non-unique.

Since one can obtain the solution to the constrained objective by solving the unconstrained one and then reparameterising (and vice versa), we are free to assume without loss of generality that $w^T\bar{Y} = 0$, so $\hat{\mu} = 0$, when solving the univariate problem, and will remark where we do this.

S1.2 Proofs of results in Sections 2 & 3

Proof of Proposition 1. Assume, without loss of generality, that $\hat{\mu} = 0$. Suppose that there exists $l \neq k$ such that $\hat{\theta}_k = \hat{\theta}_l$. Without loss of generality we have that $\bar{Y}_k \neq \hat{\theta}_k$ (if $\bar{Y}_k = \hat{\theta}_k$ then $\bar{Y}_l \neq \hat{\theta}_l$ and it can be seen that $\hat{\theta}_{(1)} < \bar{Y}_l < \hat{\theta}_K$, in which case swap labels).

Now we construct $\tilde{\boldsymbol{\theta}}$ by setting $\tilde{\theta}_r = \hat{\theta}_r \wedge \bar{Y}_k$ for $r = 1, \ldots, k$, and $\tilde{\theta}_r = \hat{\theta}_r$ otherwise. We have $\ell(\hat{\mu}, \tilde{\boldsymbol{\theta}}) < \ell(\hat{\mu}, \hat{\boldsymbol{\theta}})$ and, by convexity of ρ, it follows that

$$\sum_{r=1}^{K-1} \rho(\tilde{\theta}_{(r+1)} - \tilde{\theta}_{(r)}) \leq \sum_{r=1}^{K-1} \rho(\hat{\theta}_{(r+1)} - \hat{\theta}_{(r)}).$$

This gives the conclusion $Q(\tilde{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\theta}$. □

Proof of Proposition 2. Suppose, for a contradiction, that $\hat{\theta}_k < \hat{\theta}_l$. Then at least one of the following must be true:

$$\left|\hat{\mu} + \hat{\theta}_k - \bar{Y}_k\right| > \left|\hat{\mu} + \hat{\theta}_l - \bar{Y}_k\right| \tag{32}$$

$$\left|\hat{\mu} + \hat{\theta}_l - \bar{Y}_l\right| > \left|\hat{\mu} + \hat{\theta}_k - \bar{Y}_l\right|. \tag{33}$$

Let $\tilde{\boldsymbol{\theta}}$ be defined as follows. Set $\tilde{\theta}_r = \hat{\theta}_r$ for all $r \neq k, l$. If (32) holds set $\tilde{\theta}_k = \hat{\theta}_l$ and if (33) holds set $\tilde{\theta}_l = \hat{\theta}_k$. Observe that

$$\sum_{r=1}^{n} \rho(\hat{\theta}_{(r+1)} - \hat{\theta}_{(r)}) \geq \sum_{r=1}^{n} \rho(\tilde{\theta}_{(r+1)} - \tilde{\theta}_{(r)})$$

and that the squared loss of $\tilde{\boldsymbol{\theta}}$ is strictly smaller than the squared loss of $\hat{\boldsymbol{\theta}}$, thus contradicting optimality of $\hat{\boldsymbol{\theta}}$. □

Proof of Proposition 3. In this proof we consider the unconstrained formulation of the objective (31) discussed in Section S1.1. Suppose that $(\bar{Y}_k)_{k=1}^K$ is such that there are two distinct solutions to (12), $\hat{\boldsymbol{\theta}}^{(1)} \neq \hat{\boldsymbol{\theta}}^{(2)}$. Let us assume that the levels are indexed such that $\bar{Y}_1 \leq \cdots \leq \bar{Y}_K$. Define $k^* = \max\{k : \hat{\theta}_k^{(1)} \neq \hat{\theta}_k^{(2)}\}$ to be the largest index at which the two solutions take different values and note that we must have $\hat{\theta}_1^{(r)} \leq \cdots \leq \hat{\theta}_K^{(r)}$.

First consider the case where $k^* < K$. Then

$$S_r := \{k : \hat{\theta}_k^{(r)} = \hat{\theta}_{k^*+1}^{(r)}\} \subseteq \{k^* + 1, k^* + 2, \ldots, K\},$$

for $r = 1, 2$. We now argue that we must have $\hat{\theta}_{k^*+1}^{(1)} = \hat{\theta}_{k^*+1}^{(2)} =: t^* \geq (\hat{\theta}_{k^*}^{(1)} \vee \hat{\theta}_{k^*}^{(2)}) + \gamma\lambda$. Indeed, suppose not, and suppose that without loss of generality $\hat{\theta}_{k^*}^{(2)} > \hat{\theta}_{k^*}^{(1)}$. Fix $r \in \{1, 2\}$. The directional derivative of the objective in the direction of the binary vector with ones at the indices given by S_r and zeroes elsewhere evaluated at $\hat{\boldsymbol{\theta}}^{(r)}$ must be 0. But comparing these for $r = 1, 2$, we see they are identical except for the term $\rho'(\theta_{k^*+1} - \hat{\theta}_{k^*}^{(r)})$, which will be strictly larger for $r = 2$, giving a contradiction. This then implies that both $\hat{\theta}_{k^*}^{(1)}$ and $\hat{\theta}_{k^*}^{(2)}$ must minimise f_{k^*} over $\theta \leq t^* - \gamma\lambda$ since the full objective value is

$$Q(\hat{\boldsymbol{\theta}}^{(r)}) = f_{k^*}(\hat{\theta}_{k^*}^{(r)}) + \frac{1}{2}\gamma\lambda^2 + (\text{terms featuring only index } k^* + 1 \text{ or higher})$$

for $r = 1, 2$. We also have that when $k^* = K$, both $\hat{\theta}_{k^*}^{(1)}$ and $\hat{\theta}_{k^*}^{(2)}$ must minimise f_{k^*}.

Using the functions g_{k-1} as defined in (15), we have the simple relationship that $g_{k-1}(\theta_k) = f_k(\theta_k) - \frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$. In particular, properties (i) and (iii) of Lemma 4 hold with f_k replaced by g_{k-1}. These can be characterised as $g_{k-1}(\theta_k) = \check{q}_{k,r}(\theta_k)$ for $\theta_k \in I_{k,r}$, where $I_{k,r}$ are the intervals associated with f_k and $\check{q}_{k,r}(\theta_k) := q_{k,r}(\theta_k) - \frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$. Note that for each r, $\check{q}_{k,r}$ depends on the values of $\bar{Y}_1, \ldots, \bar{Y}_{k-1}$ but not that of $\bar{Y}_k$ (observe that $q_{k,r}(\theta_k)$ includes a term $\frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$; see (13)).

Now as $\hat{\theta}_{k^*}^{(1)} \leq \hat{\theta}_{k^*+1}^{(1)} - \gamma\lambda$ and $\hat{\theta}_{k^*}^{(2)} \leq \hat{\theta}_{k^*+1}^{(2)} - \gamma\lambda$ (if $k^* < K$), by Lemma 4 (iii) both must be local minima of f_{k^*}, and we have that there must exist distinct $r_1 \neq r_2$ such that $\hat{\theta}_{k^*}^{(1)} \in I_{k^*,r_1}$ and $\hat{\theta}_{k^*}^{(2)} \in I_{k^*,r_2}$. Let

$$\check{q}_{k^*,r_1}(x) = a_1x^2 + b_1x + c_1,$$
$$\check{q}_{k^*,r_2}(x) = a_2x^2 + b_2x + c_2.$$

Since $\hat{\theta}^{(1)}_{k^*}$ must be the minimum of $\check{q}_{k^*,r_1}(\theta_{k^*}) + \frac{1}{2}w_{k^*}(\bar{Y}_{k^*} - \theta_{k^*})^2$ (and similarly for $\hat{\theta}^{(2)}_{k^*}$), we must have that

$$\min_x \left\{ a_1 x^2 + b_1 x + c_1 + \frac{1}{2} w_{k^*} (\bar{Y}_{k^*} - x)^2 \right\} = \min_x \left\{ a_2 x^2 + b_2 x + c_2 + \frac{1}{2} w_{k^*} (\bar{Y}_{k^*} - x)^2 \right\}$$
$$\implies c_1 - \frac{(b_1 - w_{k^*}\bar{Y}_{k^*})^2}{4a_1 + 2w_{k^*}} = c_2 - \frac{(b_2 - w_{k^*}\bar{Y}_{k^*})^2}{4a_2 + 2w_{k^*}}. \tag{34}$$

This is a quadratic equation in $\bar{Y}_{k^*}$, so there are at most two values for which (34) holds. Considering all pairs r_1, r_2, we see that in order for there to exist two solutions $\hat{\boldsymbol{\theta}}^{(1)} \neq \hat{\boldsymbol{\theta}}^{(2)}$, $\bar{Y}_{k^*}$ must take values in a set of size at most $c(K)$, for some function $c : \mathbb{N} \to \mathbb{N}$.

Now let

$$\mathcal{S} := \{(\bar{Y}_k)_{k=1}^K : \text{ the minimiser of the objective is not unique}\} \subseteq \mathbb{R}^K.$$

What we have shown, is that associated with each element $(\bar{Y}_k)_{k=1}^K \in \mathcal{S}$, there is at least one k^* such that

$$|\{(\bar{Y}'_k)_{k=1}^K \in \mathcal{S} : \bar{Y}'_k = \bar{Y}_k \text{ for all } k \neq k^*\}|$$

is bounded above by $c(K)$. Now for each $j = 1, \ldots, K$, let $\mathcal{S}_j$ be the set of $(\bar{Y}_k)_{k=1}^K \in \mathcal{S}$ for which the there exists a k^* with the property above and $k^* = j$. Note that $\cup_j \mathcal{S}_j = \mathcal{S}$. Now $\mathcal{S}_j \subset \mathbb{R}^K$ has Lebesgue measure zero as a finite union of graphs of measurable functions $f : \mathbb{R}^{K-1} \to \mathbb{R}$. Thus $\mathcal{S}$ has Lebesgue measure zero. □

Proof of Lemma 4. Assume, without loss of generality, that $\hat{\mu} = 0$. We proceed inductively, assuming that the properties (i) and (iii) hold for f_k, and (ii) holds for b_{k+1}. Additionally we include in our inductive hypothesis that for all x, $f'_k(x_-) \geq f'_k(x_+)$, where we define $f'_k(x_-)$ and $f'_k(x_+)$ to be the left-derivative and right-derivative of f_k at x, respectively. We note that these trivially hold for the base case f_1, and the case b_2 can be checked by direct calculation.

We first prove (i), that f_{k+1} is continuous, coercive, and piecewise quadratic and with finitely many pieces. We then show that $f'_{k+1}(x_-) \geq f'_{k+1}(x_+)$ for all x, which allows us to show that (iii) holds for f_{k+1}. Finally, we use these results to show that (ii) holds for b_{k+2}.

We now show that f_{k+1} is coercive and continuous. Clearly $g_k(x) \geq \min_{y \leq x} f_k(y)$, so it follows that $g_k(x) \to \infty$ as $x \to -\infty$ as f_k is coercive. Furthermore g_k is bounded from below as f_k is coercive and continuous. Thus since $f_{k+1}(x) = g_k(x) + \frac{1}{2}w_{k+1}(\bar{Y}_{k+1} - x)^2$, it follows that f_{k+1} is coercive. Next as $g_k(x) = \min_{y \leq x} f_k(y) + \rho(y - x)$, and f_k and ρ are continuous, it follows that g_k is continuous and therefore that f_{k+1} is continuous.

To see why f_{k+1} is piecewise quadratic with finitely many pieces, we observe that it can be written $f_{k+1}(x) = f_k(b_{k+1}(x)) + \rho(x - b_{k+1}(x)) + \frac{1}{2}w_{k+1}(\bar{Y}_{k+1} - x)^2$. We have by our inductive hypothesis that f_k is piecewise quadratic and $b_{k+1}(x)$ is piecewise linear, both with finitely many pieces. Since the composition of a piecewise linear function inside a piecewise quadratic function is piecewise quadratic, the remainder of (i) is shown.

We now turn our attention to (iii), and define for $x \in \mathbb{R}$:

$$y_*(x) = \operatorname*{sarg\,min}_{y \leq x} f_k(y) + \rho(x - y),$$
$$y^*(x) = \operatorname*{sarg\,min}_{y \leq x} f_{k+1}(y) + \rho(x - y).$$

We will first show that $f'_{k+1}(x_+) \leq f'_{k+1}(x_-)$ for all $x \in \mathbb{R}$. Suppose that we are increasing x and we have reached a point where $g_k(x)$ is not differentiable (that is, the left-derivative and

the right-derivative do not match). By assumption (ii) for b_{k+1}, we can assume that there is some window $\delta > 0$ such that $y_*(t)$ is linear for $t \in (x-\delta, x)$, say $y_*(t) = \alpha + \beta t$.

In order to proceed with the following argument, we must show that for sufficiently small $\epsilon > 0$, we have $\alpha + \beta(x+\epsilon) \leq x + \epsilon$. If $\alpha + \beta x < x$, this is immediate. Therefore it remains to consider the case $\alpha + \beta x = x$, for which we show that we must have $\alpha = 0$ and $\beta = 1$, i.e $y_*(t) = t$ for $t \in (x-\delta, x)$. This follows from the observation that if $y_*(t) < t$, then for all $t_1 > t$ we have $y_*(t_1) \notin (y_*(t), t]$. Indeed, suppose not, then

$$\begin{aligned} & f_k(y_*(t_1)) + \rho(t_1 - y_*(t_1)) < f_k(y_*(t)) + \rho(t_1 - y_*(t)) \\ \implies \quad & f_k(y_*(t_1)) + \rho(t - y_*(t_1)) < f_k(y_*(t)) + \rho(t_1 - y_*(t)) + \rho(t - y_*(t_1)) - \rho(t_1 - y_*(t_1)) \\ & \leq f_k(y_*(t)) + \rho(t - y_*(t)), \end{aligned}$$

contradicting the definition of $y_*(t)$. The last line uses $\rho(t_1 - y_*(t)) - \rho(t_1 - y_*(t_1)) \leq \rho(t - y_*(t)) - \rho(t - y_*(t_1))$, which follows from concavity of ρ and $y_*(t) < y_*(t_1) \leq t < t_1$.

With this established, we have that:

$$\begin{aligned} g_k(x-\epsilon) &= f_k(\alpha + \beta(x-\epsilon)) + \rho(x - \epsilon - (\alpha + \beta(x-\epsilon))) \\ g_k(x+\epsilon) &= f_k(y_*(x+\epsilon)) + \rho(x + \epsilon - y_*(x+\epsilon)) \\ &\leq f_k(\alpha + \beta(x+\epsilon)) + \rho(x + \epsilon - (\alpha + \beta(x+\epsilon))). \end{aligned}$$

Note that f_k has both left-derivatives and right-derivatives at every point in $\mathbb{R}$. Suppose first that $\beta \geq 0$, and we observe that

$$g_k'(x_-) = \beta f_k'(y_*(x)_-) + (1-\beta)\rho'(x - y_*(x))$$

Then by the basic definition of the right-derivative,

$$\begin{aligned} g_k'(x_+) &= \lim_{\epsilon \to 0^+} \frac{f_k(y_*(x+\epsilon)) + \rho(x + \epsilon - y_*(x+\epsilon)) - f_k(y_*(x)) - \rho(x - y_*(x))}{\epsilon} \\ &\leq \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \Big[f_k(\alpha + \beta(x+\epsilon)) + \rho(x + \epsilon - (\alpha + \beta(x+\epsilon))) \\ &\qquad - f_k(\alpha + \beta x) - \rho(x - (\alpha + \beta x)) \Big] \\ &= \beta f_k'(y_*(x)_+) + (1-\beta)\rho'(x - y_*(x)) \\ &= g_k'(x_-) + \beta(f_k'(y_*(x)_+) - f_k'(y_*(x)_-)) \\ &\leq g_k'(x_-), \end{aligned}$$

where the last inequality follows from our inductive hypothesis that $f_k'(y_+) \leq f_k'(y_-)$ for all $y \in \mathbb{R}$. An analogous argument shows that the same conclusion holds when $\beta < 0$.

Now we use this to prove the claim. Because there are no points of f_{k+1} at which the left-derivative is less than the right-derivative, without loss of generality we claim that f_{k+1} is differentiable at $y^*(x)$ for all x, unless $y^*(x) = x$. Indeed, suppose not, then we have that $f_{k+1}'(y^*(x)_-) > f_{k+1}'(y^*(x)_+)$ and necessarily that defining $h(y) := f_{k+1}(y) + \rho(x-y)$, we have $0 \in \partial h(y^*(x))$. But since $h(y^*(x)_+) < h(y^*(x)_-)$, we contradict the optimality of $y^*(x)$ as this point is in fact a local maximum.

We finally consider claim (ii). By (iii), we have that for every point x, $y^*(x)$ is either x or at the minimum of one of the quadratic pieces of $f_{k+1}(\cdot) + \rho(x - \cdot)$. In either case, we have that $y^*(x)$ is linear in x and thus $f_{k+1}(y^*(x)) + \rho(x - y^*(x))$ is quadratic in x. We can define $g_{k+1}(x)$ pointwise as the minimum of this finite set of quadratic functions of x, whose expressions are

given in Appendix A.1. Importantly, the coefficients in the linear expression $y^*(x)$ of x depend only on which of these functions is the minimum at x. As the number of intersections between elements in this set of quadratic functions is bounded above by twice the square of the size of the set, we can conclude that $b_{k+2}(x)$ is piecewise linear and with a finite number of pieces, thus concluding the proof. □

S1.3 Computation time experiments

A small experiment was performed to demonstrate the runtimes one can expect in practice for the univariate problem. Note that this clustering is applied iteratively in the block coordinate descent procedure we propose to use in multivariate settings. We considered 3 settings: one with no signal, one with 2 true clusters and one with 5 true clusters. Independent and identically distributed Gaussian noise was added to each of the subaverages. As in Section 6.3 the number of categories was increased by random splitting of the levels. Each of these tests were repeated 25 times, on a computer with a 3.2GHz processor. The results are shown in Figure 8.

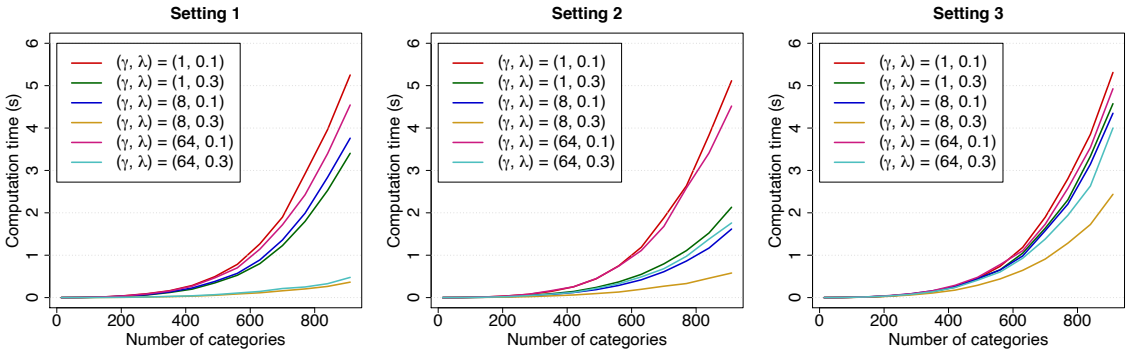

Figure 8: Computation times for solving the univariate problem.

S1.4 Discretised algorithm

For very large-scale problems, speed can be improved if we only allow coefficients to take values in some fixed finite grid, rather than any real value. Below we describe how such an algorithm would approximately solve the univariate objective (12). We will use the unconstrained objective as discussed in Section S1.1. We would first fix L grid points $\vartheta_1 < \cdots < \vartheta_L$, and then proceed as described in Algorithm 3.

This algorithm has the same basic structure to the approach we use in Section 3.1 for computing the exact global optimum. The difference is that now, instead of as in (14), we define f_k in the following way:

$$f_k(\theta_k) := \min_{\substack{(\theta_1,\ldots,\theta_{k-1})^T \in \{\vartheta_1,\ldots,\vartheta_L\}^{k-1} \\ \theta_1 \leq \cdots \leq \theta_{k-1} \leq \theta_k}} \left\{ \frac{1}{2} \sum_{l=1}^{k} w_l (\bar{Y}_l - \theta_l)^2 + \sum_{l=1}^{k-1} \rho(\theta_{l+1} - \theta_l) \right\}.$$

The objects F and B play analogous roles to f_k and b_k in Section 3.1. Since we restrict $\theta_k \in \{\vartheta_1, \ldots, \vartheta_L\}$, we only need to store the values that f_k takes at these L values; this is the purpose of the vector F in Algorithm 3. Similarly, the rows $B(k, \cdot)$ serve the same purpose as the functions b_k where, again, we only need to store L values corresponding to the different options for θ_k.

Algorithm 3 Discrete algorithm for computing approximate solution to (12)

for $l = 1, \ldots, L$ **do**
  Set $F_{\text{new}}(l) = \frac{1}{2} w_1 (\bar{Y}_1 - \vartheta_l)^2$
  Set $B(1, l) = l$
end for
for $k = 2, \ldots, K$ **do**
  Set $F_{\text{old}} = F_{\text{new}}$
  for $l = 1, \ldots, L$ **do**
    Set $B(k, l) = \arg\min_{l' \in \{1, \ldots, l\}} F_{\text{old}}(l') + \rho(\vartheta_l - \vartheta_{l'}) + \frac{1}{2} w_k (\bar{Y}_k - \vartheta_l)^2$
    Set $F_{\text{new}}(l) = F_{\text{old}}(B(k, l)) + \rho(\vartheta_l - \vartheta_{B(k,l)}) + \frac{1}{2} w_k (\bar{Y}_k - \vartheta_l)^2$
  end for
end for
Set $B^*(K) = \arg\min F_{\text{new}}$, and $\hat{\theta}_K = \vartheta_{B^*(K)}$
for $k = K - 1, \ldots, 1$ **do**
  Set $B^*(k) = B(k + 1, B^*(k + 1))$, and $\hat{\theta}_k = \vartheta_{B^*(k)}$
end for

This algorithm returns the optimal solution $\hat{\boldsymbol{\theta}}$ to the objective where each of the coefficients are restricted to take values only in $\{\vartheta_1, \ldots, \vartheta_L\}$. We must ensure that the grid of values has fine enough resolution that interesting answers can be obtained, which requires L being sufficiently large. The number of clusters obtained by this approximate algorithm is bounded above by L, so this must not be chosen too small.

One can see that the computational complexity of this algorithm is linear in K, with a total of $O(KL^2)$ operations required. This is of course in addition to the $O(n)$ operations needed to compute $w_1, \ldots, w_K$ and $\bar{Y}_1, \ldots, \bar{Y}_K$ beforehand. In particular, choosing $L \lesssim \sqrt{K}$ guarantees that the complexity of this algorithm is at worst quadratic in K.

S2 Proofs of results in Section 4

S2.1 Proof of Theorem 5

The proof of Theorem 5 requires a number of auxiliary lemmas, which can be found in Section S2.1.1.

Let us define $R_i = Y_i - \hat{\mu}$ for $i = 1, \ldots, n$, and $\bar{R}_k = \frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i = k\}} R_i$ for $k = 1, \ldots, K$. Note that

$$R_i = \sum_{k=1}^K \mathbb{1}_{\{X_i = k\}} \theta_k^0 + (P\varepsilon)_i$$

where $P = I - \mathbf{1}\mathbf{1}^T/n$.

For each $k = 1, \ldots, K$, we define the event

$$\Lambda_k = \left\{ \left| \frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i = k\}} (P\varepsilon)_i \right| < \frac{1}{2} \sqrt{\eta \gamma_* s} \lambda \right\}.$$

By a union bound, we have that $\mathbb{P}(\cap_{k=1}^K \Lambda_k) \geq 1 - \sum_{k=1}^K \mathbb{P}(\Lambda_k^c)$. Now observe we can write

$$\frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i = k\}} (P\varepsilon)_i = v^{(k)T} P\varepsilon,$$

where we define $v^{(k)} \in \mathbb{R}^n$ by $v_i^{(k)} = \frac{1}{n_k}\mathbb{1}_{\{X_i=k\}}$. Since P is an orthogonal projection matrix, we have that $\|Pv^{(k)}\|_2 \leq \|v^{(k)}\|_2 = \frac{1}{\sqrt{n_k}}$. It follows that ${v^{(k)}}^T P\varepsilon$ is sub-Gaussian with parameter $\sigma/\sqrt{n_k}$. Applying the standard sub-Gaussian tail bound, we obtain

$$\begin{aligned}\mathbb{P}(\Lambda_k^c) &= \mathbb{P}\left(\left|\frac{1}{n_k}\sum_{i=1}^n \mathbb{1}_{\{X_i=k\}}(P\varepsilon)_i\right| \geq \frac{1}{2}\sqrt{\eta\gamma_* s}\lambda\right) \\ &\leq 2\exp\left(-\frac{nw_k\eta\gamma_* s\lambda^2}{8\sigma^2}\right),\end{aligned}$$

where recall that $w_k = n_k/n$. Therefore, we have that

$$\mathbb{P}\left(\cap_{k=1}^K \Lambda_k\right) \geq 1 - 2\sum_{k=1}^K \exp\left(-\frac{nw_k\eta\gamma_* s\lambda^2}{8\sigma^2}\right) \geq 1 - 2\exp\left(-\frac{nw_{\min}\eta\gamma_* s\lambda^2}{8\sigma^2} + \log(K)\right). \quad (35)$$

In the following we work on the intersection $\Lambda := \cap_{k=1}^K \Lambda_k$. This entails that for each k, $|\bar{R}_k - \theta_k^0| < \sqrt{\eta\gamma_* s}\lambda/2$. We now relabel indices such that $\bar{R}_1 \leq \cdots \leq \bar{R}_K$, and so from Proposition 2 that $\hat{\theta}_1 \leq \cdots \leq \hat{\theta}_K$. Since our assumption (24) implies $\Delta(\boldsymbol{\theta}^0) \geq \sqrt{\eta\gamma_* s}\lambda$, it follows that on Λ the observed ordering is consistent with the ordering of the true coefficients, i.e. there exist $0 = k_0 < k_1 < \cdots < k_s = K$ such that

$$\theta_1^0 = \cdots = \theta_{k_1}^0 < \theta_{k_1+1}^0 = \cdots = \theta_{k_2}^0 < \cdots < \theta_{k_{s-1}+1}^0 = \cdots = \theta_{k_s}^0. \quad (36)$$

Indeed, observe that for $j = 1, \ldots, s-1$, we have by the triangle inequality and (24), the stronger property that

$$\begin{aligned}\bar{R}_{k_j+1} - \bar{R}_{k_j} &> 3\left(1 + \frac{\sqrt{2}}{\eta}\right)\sqrt{\gamma\gamma^*}\lambda - \sqrt{\eta\gamma_* s}\lambda \\ &> \gamma\lambda + 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + 2\sqrt{\eta\gamma_* s}\lambda. \quad (37)\end{aligned}$$

Our optimisation objective is therefore

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \frac{1}{2}\sum_{k=1}^K w_k(\bar{R}_k - \theta_k)^2 + \sum_{k=1}^K \rho(\theta_{k+1} - \theta_k). \quad (38)$$

Since $\bar{R}_{k_j} - \bar{R}_{k_{j-1}+1} < \sqrt{\eta\gamma_* s}\lambda$ for $j = 1, \ldots, s$, it follows from Lemma 8 that $\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j} \geq \gamma\lambda$ for $j = 1, \ldots, s-1$, so

$$\begin{aligned}Q(\hat{\boldsymbol{\theta}}) &= \frac{1}{2}\sum_{k=1}^K w_k(\bar{R}_k - \hat{\theta}_k)^2 + \sum_{k=1}^{K-1}\rho(\hat{\theta}_{k+1} - \hat{\theta}_k) \\ &= \frac{1}{2}\sum_{k=1}^K w_k(\bar{R}_k - \hat{\theta}_k)^2 + \sum_{j=1}^s \sum_{k=k_{j-1}+1}^{k_j-1}\rho(\hat{\theta}_{k+1} - \hat{\theta}_k) + \frac{s-1}{2}\gamma\lambda^2 \quad (39) \\ &= \min_{\boldsymbol{\theta}\in\mathbb{R}^K}\left(\frac{1}{2}\sum_{k=1}^K w_k(\bar{R}_k - \theta_k)^2 + \sum_{j=1}^s \sum_{k=k_{j-1}+1}^{k_j-1}\rho(\theta_{k+1} - \theta_k)\right) + \frac{s-1}{2}\gamma\lambda^2. \quad (40)\end{aligned}$$

Observe that we can have $k_{j-1}+1 > k_j - 1$ for some j, in which case we take the sum over that range to be zero. Note that (40) can be optimised over $(\theta_{k_{j-1}+1}, \ldots, \theta_{k_j})$ separately for each

$j = 1, \ldots, s$. If $s = 1$, i.e. the true signal is zero, then the result follows from Lemma 10. Now we see what happens when $s > 1$.

Without loss of generality, consider $j = 1$ and note that if $k_1 = 1$ it is immediate that $\hat{\theta}_1 = \hat{\theta}_1^0$. Hence, we can assume that $k_1 > 1$. We note that $\hat{\theta}_1^0 = \sum_{k=1}^{k_1} w_k \bar{R}_k / w_1^0$, where we define $w_k^0 = n_k^0/n$. We see that our goal is to compute

$$\begin{aligned} &\operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{k_1}} \frac{1}{2} \sum_{k=1}^{k_1} w_k (\bar{R}_k - \theta_k)^2 + \sum_{k=1}^{k_1 - 1} \rho(\theta_{k+1} - \theta_k) \\ = \hat{\theta}_1^0 \mathbf{1} + &\operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{k_1}} \frac{1}{2} \sum_{k=1}^{k_1} w_k (\tilde{R}_k - \theta_k)^2 - \sum_{k=1}^{k_1 - 1} \rho(\theta_{k+1} - \theta_k), \end{aligned} \tag{41}$$

where $\mathbf{1} \in \mathbb{R}^{k_1}$ is a vector of ones and $\tilde{R}_k := \bar{R}_k - \hat{\theta}_1^0$ for $k = 1, \ldots, k_1$. Note that we subtract $\hat{\theta}_1^0$ to ensure that

$$\sum_{k=1}^{k_1} w_k \tilde{R}_k = 0,$$

as required for application of Lemma 10. We have by assumption that for $k \in 1, \ldots, k_1$, $|\tilde{R}_k| \leq \sqrt{\eta \gamma_* s} \lambda / 2 \leq (2 \wedge \sqrt{w_1^0 \gamma}) \lambda / w_1^0$. Thus, Lemma 10 can be applied with $\check{w} = w_1^0$ and it follows that $\hat{\theta}_k = \hat{\theta}_1^0$ for $k = 1, \ldots, k_1$. □

S2.1.1 Auxiliary lemmas

Here we prove a number of results required to obtain conditions for recovering the oracle least squares estimate in the univariate case. Lemma 10 gives conditions for recovery of the true solution, in the case where there is zero signal. Lemmas 8 and 9 ensure that the true levels are far enough apart that they can be separated. Once we have this separation, we apply Lemma 10 on each of the levels to obtain the solution.

Lemma 7. *Consider the optimisation problem*

$$x^* = \operatorname*{arg\,min}_{x \geq 0} \frac{\kappa}{2} (2\tau - x)^2 + \rho(x),$$

where $\tau > 0$ and $\kappa \in (0, 1]$. Suppose further that $\tau < (1 \wedge \sqrt{\kappa\gamma}) \lambda / 2\kappa$. Then $x^* = 0$ is the unique optimum.

Proof. We first observe that

$$x^* = \operatorname*{arg\,min}_{x \geq 0} \frac{\kappa}{2} (2\tau - x)^2 + \rho_{\gamma,\lambda}(x) = \operatorname*{arg\,min}_{x \geq 0} \frac{1}{2} (2\tau - x)^2 + \rho_{\kappa\gamma, \lambda/\kappa}(x).$$

For convenience, we define $F(x) := (2\tau - x)^2/2 + \rho_{\kappa\gamma,\lambda/\kappa}(x)$. It now suffices to show that F is uniquely minimised at 0 provided $\tau < (1 \wedge \sqrt{\kappa\gamma}) \lambda / 2\kappa$. We can clearly see that $x^* \in [0, 2\tau]$. Equation (2.3) of Breheny and Huang [2011] gives the result when $\kappa\gamma \geq 1$.

When $\kappa\gamma < 1$, we see that any stationary point of F in $[0, \gamma\lambda \wedge 2\tau]$ must be a maximum, since on this interval $F(x)$ is a quadratic function with a negative coefficient of x^2. Therefore its minimum over $[0, \gamma\lambda]$ is attained at either $x = 0$ or $x = \gamma\lambda \wedge 2\tau$. If $2\tau \leq \gamma\lambda$, then it suffices to check that $F(0) < F(2\tau)$. This holds if and only if $\tau < \gamma\lambda / (\gamma\kappa + 1)$, but since we are assuming $\tau \leq \gamma\lambda/2$ and $\kappa\gamma < 1$, this is always satisfied.

If $\gamma\lambda < 2\tau$, then we can see that the minimum of F over $[\gamma\lambda, 2\tau]$ will be attained at exactly 2τ. Thus, here it also suffices to check $F(0) < F(2\tau)$, which holds if and only if $\tau < \sqrt{\gamma/\kappa}\lambda/2$. The final bound $\tau < (1 \wedge \sqrt{\kappa\gamma}) \lambda / 2\kappa$ follows from combining the results for these cases. □

The following is a deterministic result to establish separation between groups of coefficients.

Lemma 8. *Consider the setup of Theorem 5, and assume that $\hat{\mu} = 0$. Suppose that $\bar{Y}_1 \leq \cdots \leq \bar{Y}_K$, and that for $j = 1, \ldots, s$ we have*

$$\bar{Y}_{k_j} - \bar{Y}_{k_{j-1}+1} < \sqrt{\eta\gamma_* s}\lambda, \tag{42}$$

where k_j and k_{j-1} are as defined in (36). Suppose further that for $j = 1, \ldots, s-1$,

$$\bar{Y}_{k_j+1} - \bar{Y}_{k_j} \geq \gamma\lambda + 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + 2\sqrt{\eta\gamma_* s}\lambda. \tag{43}$$

Then for $j = 1, \ldots, s$, we have $\bar{Y}_{k_{j-1}+1} \leq \hat{\theta}_{k_{j-1}+1} \leq \hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$.

Proof. For convenience, within this lemma we define $\zeta := \sqrt{\eta\gamma_* s}\lambda$. Recall that the objective function which $\hat{\boldsymbol{\theta}}$ optimises takes the form

$$Q(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \theta_k)^2 + \sum_{k=1}^{K-1} \rho(\theta_{k+1} - \theta_k).$$

We first claim that $\hat{\theta}_k \in [\bar{Y}_1, \bar{Y}_K]$ for $k = 1, \ldots, K$. To see this, suppose that this is not the case and define $\check{\boldsymbol{\theta}}$ by projecting $\hat{\boldsymbol{\theta}}$ onto $[\bar{Y}_1, \bar{Y}_K]^K$ (i.e. $\check{\theta}_k = \bar{Y}_K \wedge (\bar{Y}_1 \vee \hat{\theta}_k)$ for $k = 1, \ldots, K$). The penalty contribution from $\check{\boldsymbol{\theta}}$ is no larger than that of $\hat{\boldsymbol{\theta}}$, and the loss contribution is strictly smaller, so we obtain the contradiction $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$.

We now proceed to show that for $j = 1, \ldots, s-1$, we have $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$ and $\hat{\theta}_{k_j+1} \geq \bar{Y}_{k_j+1}$. We prove the first of these sets of inequalities, since the second follows similarly by considering the problem with $-\hat{\boldsymbol{\theta}}$, $-\bar{Y}$ and reversing the indices. Suppose, for contradiction, that there exists some j in $\{1, \ldots, s-1\}$ with $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$. Let this j be minimal, such that for all $l < j$ we have $\hat{\theta}_{k_l} \leq \bar{Y}_{k_l}$.

Next define l_1 to be the maximal element of $\{k_{j-1}+1, \ldots, k_j - 1\}$ such that $\hat{\theta}_{l_1} \leq \bar{Y}_{k_j}$. Similarly, we define $l_2 \in \{k_j + 1, \ldots, k_{j+1}\}$ to be minimal such that $\hat{\theta}_{l_2} \geq \bar{Y}_{k_j+1}$. The existence of l_1 and l_2 is guaranteed by Lemma 9.

We note that for $l = l_1 + 1, \ldots, k_j$, $\hat{\theta}_l = \hat{\theta}_{k_j}$ and hence $(\bar{Y}_l - \hat{\theta}_l)^2 \geq (\bar{Y}_{k_j} - \hat{\theta}_l)^2 = (\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2$. This can be shown by contradiction, as in (55). For such l, we have from optimality of $\hat{\theta}$ that $\bar{Y}_l - \hat{\theta}_{l_1} \geq \hat{\theta}_{k_j} - \bar{Y}_l$ (otherwise one could improve the objective by setting $\hat{\theta}_{l_1} = \hat{\theta}_l$) which implies that $\hat{\theta}_{l_1} < \bar{Y}_l$. From this it follows that $(\bar{Y}_l - \hat{\theta}_{l_1})^2 \leq (\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2$, since $\hat{\theta}_{l_1} < \bar{Y}_l \leq \bar{Y}_{k_j}$.

Similarly, if $l_2 > k_j + 1$, then for $l = k_j + 1, \ldots, l_2 - 1$ we have $\hat{\theta}_l = \hat{\theta}_{k_j+1}$ and hence $(\bar{Y}_l - \hat{\theta}_l)^2 \geq (\bar{Y}_{k_j+1} - \hat{\theta}_l)^2 = (\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1})^2$. For such l, it follows that $\hat{\theta}_{l_2} > \bar{Y}_l$ and therefore that $(\bar{Y}_l - \hat{\theta}_{l_2})^2 \leq (\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2$.

Now, we define

$$\tilde{w}_{k_j} := \sum_{l \leq k_j \colon \hat{\theta}_l = \hat{\theta}_{k_j}} w_l$$

$$\text{and, if } l_2 > k_j + 1, \qquad \tilde{w}_{k_j+1} := \sum_{l \geq k_j+1 \colon \hat{\theta}_l = \hat{\theta}_{k_j+1}} w_l.$$

We also define $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^K$ according to

$$\tilde{\theta}_l = \begin{cases} \hat{\theta}_l \wedge \hat{\theta}_{l_1} & \text{for } l \leq k_j \\ \hat{\theta}_l \vee \hat{\theta}_{l_2} & \text{for } l > k_j. \end{cases}$$

We note that by assumption, both $\tilde{w}_{k_j} < 1/\eta s$ and $\tilde{w}_{k_j+1} < 1/\eta s$. We now consider two cases: (A) where $l_2 = k_j + 1$, so $\hat{\theta}_{k_j+1} \geq \bar{Y}_{k_j+1}$, and (B) where $l_2 > k_j + 1$, so $\hat{\theta}_{k_j+1} < \bar{Y}_{k_j+1}$.

We first consider case (A), where the penalty terms between l_1 and l_2 in $Q(\hat{\boldsymbol{\theta}})$ are

$$\sum_{l=l_1}^{l_2-1} \rho(\hat{\theta}_{l+1} - \hat{\theta}_l) = \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}).$$

Thus,

$$\begin{aligned}
Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) = &\sum_{l \leq k_j \,:\, \hat{\theta}_l = \hat{\theta}_{k_j}} \frac{w_l}{2}(\bar{Y}_l - \hat{\theta}_l)^2 - \sum_{l \leq k_j \,:\, \hat{\theta}_l = \hat{\theta}_{k_j}} \frac{w_l}{2}(\bar{Y}_l - \hat{\theta}_{l_1})^2 \\
&+ \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \\
\geq &\frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 \\
&+ \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \qquad (44) \\
\geq &\inf_{\bar{Y}_{k_j} < a \leq \hat{\theta}_{l_2}} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 \\
&+ \rho(\hat{\theta}_{l_2} - a) + \rho(a - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2. \qquad (45)
\end{aligned}$$

We specify the infimum in (47) because $(\bar{Y}_{k_j}, \hat{\theta}_{l_2}]$ is not closed, and let (a_m) be a convergent sequence in $(\bar{Y}_{k_j}, \hat{\theta}_{l_2}]$ whose limit attains this infimum. We define $a^* = \lim_{m\to\infty} a_m$.

By assumption (43), at least one of $(a^* - \hat{\theta}_{l_1})$ and $(\hat{\theta}_{l_2} - a^*)$ is greater than or equal to $\gamma\lambda$. Here, we use that the separation (43) $\geq 2\gamma\lambda$. If $\hat{\theta}_{l_2} - a^* \geq \gamma\lambda$ then we denote this case (A1) and (45) becomes

$$\begin{aligned}
Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) &\geq \inf_{\bar{Y}_{k_j} < a \leq \hat{\theta}_{l_2} - \gamma\lambda} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(a - \hat{\theta}_{l_1}) \qquad (46) \\
&\geq \min_{\hat{\theta}_{l_1} \leq \tilde{a} \leq \hat{\theta}_{l_2} - \gamma\lambda} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}). \qquad (47)
\end{aligned}$$

We define $\tilde{a}^*$ to be the minimiser over $\tilde{a}$ of (47). We can observe that since $\bar{Y}_{k_j} - \hat{\theta}_{l_1} < \zeta$ and $\zeta < (1 \wedge \sqrt{\gamma\tilde{w}_{k_j}})\lambda/\tilde{w}_{k_j}$, we have $\bar{Y}_{k_j} - \hat{\theta}_{l_1} < (1 \wedge \sqrt{\gamma\tilde{w}_{k_j}})\lambda/\tilde{w}_{k_j}$. Thus, we have by Lemma 7 that the uniquely optimal $\tilde{a}^* = \hat{\theta}_{l_1}$. This gives that the value of (47) is zero.

It is straightforward to see from (46) that $a^* = \bar{Y}_{k_j}$ must be the unique limit of (a_m). As we have assumed that $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$ and the infimum is not attained in $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$, the inequality in line (46) can be made strict. It follows that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

Thus, it remains for us to consider the case where $\hat{\theta}_{l_2} - a^* < \gamma\lambda$, which implies that $a^* - \hat{\theta}_{l_1} \geq \gamma\lambda$. We denote this case (A2). Now, from (45) we can obtain

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{\hat{\theta}_{l_2} - \gamma\lambda < \tilde{a} \leq \hat{\theta}_{l_2}} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\hat{\theta}_{l_2} - \tilde{a}). \qquad (48)$$

The objective is piecewise quadratic (and continuously differentiable), with two pieces: $[\hat{\theta}_{l_1}, \hat{\theta}_{l_2} - \gamma\lambda]$ and $(\hat{\theta}_{l_2} - \gamma\lambda, \hat{\theta}_{l_2}]$. On the first region, the objective is a convex quadratic with minimum at $\bar{Y}_{k_j} \in [\hat{\theta}_{l_1}, \hat{\theta}_{l_2} - \gamma\lambda]$.

By the assumption that $a^* > \hat{\theta}_{l_2} - \gamma\lambda$, we know that the objective must be concave on $(\hat{\theta}_{l_2} - \gamma\lambda, \hat{\theta}_{l_2}]$. It is clear that the derivative of the objective at $\hat{\theta}_{l_2} - \gamma\lambda$ is positive. Hence, if $\tilde{a}^* = \hat{\theta}_{l_2} - \gamma\lambda$, then the objective will take a strictly lower value at some $\tilde{a}^* \in (\hat{\theta}_{l_2} - \gamma\lambda - \epsilon, \hat{\theta}_{l_2} - \gamma\lambda)$ (for some small $\epsilon > 0$), contradicting optimality of $\tilde{a}^*$. It therefore follows that $\tilde{a}^* = \hat{\theta}_{l_2}$.

With this knowledge, we can further simplify (48) to obtain

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_2})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 > 0.$$

The second inequality follows from $\bar{Y}_{k_j} - \hat{\theta}_{l_1} \leq \zeta$ and $\hat{\theta}_{l_2} - \bar{Y}_{k_j} > \zeta$. Hence, we obtain that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

We now we direct our attention towards case (B), where similarly to before we observe that the penalty contributions between l_1 and l_2 in $Q(\hat{\boldsymbol{\theta}})$ are

$$\sum_{l=l_1}^{l_2-1} \rho(\hat{\theta}_{l+1} - \hat{\theta}_l) = \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}).$$

Similarly to (44) in case (A), we obtain

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq & \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1})^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 \\ & + \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \qquad (49) \\ \geq & \inf_{\bar{Y}_{k_j} < a \leq b < \bar{Y}_{k_j+1}} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 \\ & + \rho(\hat{\theta}_{l_2} - b) + \rho(b - a) + \rho(a - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2. \qquad (50) \end{aligned}$$

We specify the infimum in (50) because $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$ is not closed and therefore a minimum may not exist. Let (a_m, b_m) be a convergent sequence in $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$ whose limit achieves this infimum. We now define $(a^*, b^*) = \lim_{m\to\infty}(a_m, b_m)$. By assumption (43), we know that $\bar{Y}_{k_j+1} - \bar{Y}_{k_j} \geq 3\gamma\lambda$, which implies that $\hat{\theta}_{l_2} - \hat{\theta}_{l_1} \geq 3\gamma\lambda$. Thus, one of $\{(\hat{\theta}_{l_2} - b^*), (b^* - a^*), (a^* - \hat{\theta}_{l_1})\}$ must be at least $\gamma\lambda$.

We first consider if $b^* - a^* \geq \gamma\lambda$, and denote this case (B1). Here, (50) becomes

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq & \inf_{\bar{Y}_{k_j} < a \leq b < \bar{Y}_{k_j+1}} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\hat{\theta}_{l_2} - b) + \rho(a - \hat{\theta}_{l_1}) \qquad (51) \\ = & \inf_{a \in (\bar{Y}_{k_j}, \bar{Y}_{k_j+1})} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(a - \hat{\theta}_{l_1}) \\ & + \inf_{b \in (\bar{Y}_{k_j}, \bar{Y}_{k_j+1})} \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\hat{\theta}_{l_2} - b) \qquad (52) \end{aligned}$$

We can observe that (52) is the sum of two copies of (46) in case (A1). Hence, by following the same arguments as before, we see that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

It therefore remains for us to obtain the result in the case that $b^* - a^* < \gamma\lambda$, and we denote this case (B2). Using that the separation (43) $\geq 3\gamma\lambda + 2\zeta$, it is straightforward to see that one of $(\bar{Y}_{k_j+1} - b^*)$ and $(a^* - \bar{Y}_{k_j})$ must be at least $\gamma\lambda + \zeta$. By the symmetry of the problem, it is sufficient for us to consider the case where $\bar{Y}_{k_j+1} - b^* \geq \gamma\lambda + \zeta$. In this case, we can obtain from (50) that

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{(\tilde{a},\tilde{b})\in\mathcal{B}} & \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{b})^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 \\ & + \rho(\tilde{b} - \tilde{a}) + \rho(\tilde{a} - \hat{\theta}_{l_1}), \end{aligned} \tag{53}$$

where $\mathcal{B} = \Big\{(\tilde{a}, \tilde{b})\colon \hat{\theta}_{l_1} \leq \tilde{a} \leq \tilde{b} \leq \bar{Y}_{k_j+1} - \gamma\lambda - \zeta,\ \tilde{b} - \tilde{a} < \gamma\lambda\Big\}$. From this, we can extract the terms dependent on $\tilde{b}$ to obtain

$$\tilde{b}^* = \operatorname*{arg\,min}_{\tilde{a}^* \leq \tilde{b} < \tilde{a}^* + \gamma\lambda} \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{b})^2 + \rho(\tilde{b} - \tilde{a}^*). \tag{54}$$

This objective is piecewise quadratic (and continuously differentiable), with two pieces; $[\tilde{a}^*, \tilde{a}^* + \gamma\lambda)$ and $[\tilde{a}^* + \gamma\lambda, \hat{\theta}_{l_2}]$. Over the second region, the objective is a convex quadratic with minimum at $\bar{Y}_{k_j+1} \in [\tilde{a}^* + \gamma\lambda, \hat{\theta}_{l_2}]$. By following the same argument as for (48) in case (A2), we see that $\tilde{b}^* = \tilde{a}^*$.

With this knowledge, we can further simplify (53) to obtain

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{\hat{\theta}_{l_1} \leq \tilde{a} \leq \bar{Y}_{k_j+1} - \gamma\lambda - \zeta} & \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{a})^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}). \end{aligned}$$

Since $\bar{Y}_{k_j+1} - \tilde{a}^* > \zeta$, we can see that $(\bar{Y}_{k_j+1} - \tilde{a}^*)^2 - (\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 > 0$. Thus, it suffices for us to show that

$$\min_{\hat{\theta}_{l_1} \leq \tilde{a} \leq \bar{Y}_{k_j+1} - \gamma\lambda - \zeta} \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}) \geq 0.$$

This objective is exactly as in (47) in case (A1), minimised over a smaller feasible set. Hence, it follows immediately that this holds and we can conclude that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

We now have for all cases that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$, which contradicts the optimality of $\hat{\theta}$. Thus, we can conclude that for $j = 1, \ldots, s$, $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$ and $\hat{\theta}_{k_{j-1}+1} \geq \bar{Y}_{k_{j-1}+1}$. □

Lemma 9. *Consider the setup of Lemma 8. For each $j = 1, \ldots, s$, there exists k_j^* in $\{k_{j-1} + 1, \ldots, k_j\}$ such that $\hat{\theta}_{k_j^*} \in [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$.*

Proof. We first show that if $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$, then for any k with $k_{j-1} + 1 \leq k \leq k_j$, if $\hat{\theta}_k > \bar{Y}_{k_j}$ then $\hat{\theta}_k = \hat{\theta}_{k_j}$.

We prove the first case since the proof for the second is identical. Suppose that this does not hold, i.e. $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$ and there exists some (minimal) k in $\{k_{j-1} + 1, \ldots, k_j - 1\}$ with

$\bar{Y}_{k_j} < \hat{\theta}_k < \hat{\theta}_{k_j}$. Then we construct $\check{\boldsymbol{\theta}}$ by

$$\check{\theta}_l = \begin{cases} \hat{\theta}_k & \text{for } l = k, k+1, \ldots, k_j \\ \hat{\theta}_l & \text{otherwise.} \end{cases} \tag{55}$$

We observe that the penalty contribution from $\check{\boldsymbol{\theta}}$ is no more than that of $\hat{\boldsymbol{\theta}}$ and that the quadratic loss for $\check{\boldsymbol{\theta}}$ will be strictly less than that of $\hat{\boldsymbol{\theta}}$. This gives us that $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\boldsymbol{\theta}}$.

Similarly, if $\hat{\theta}_{k_{j-1}+1} < \bar{Y}_{k_{j-1}+1}$ then the corresponding statement that for any k with $k_{j-1} + 1 \leq k_j$, if $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$ then $\hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}$.

We now establish a simple preliminary result. Suppose that for some j in $\{1, \ldots, s\}$ there exists k in $\{k_{j-1} + 1, \ldots, k_j\}$ with $\hat{\theta}_k \notin [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$, such that $\sum_{\{l:\, \hat{\theta}_l = \hat{\theta}_k\}} w_l \geq \eta/2s$. We claim that if $\hat{\theta}_k > \bar{Y}_{k_j}$ then $\hat{\theta}_k \leq \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$. Similarly, if $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$ then $\hat{\theta}_k \geq \bar{Y}_{k_{j-1}+1} - (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$.

To prove the claim, we consider the case $\hat{\theta}_k > \bar{Y}_{k_j}$ (the other is identical). By the first observation, if $\hat{\theta}_l > \bar{Y}_{k_j}$ for l in $\{k_{j-1} + 1, \ldots k_j\}$ then $\hat{\theta}_l = \hat{\theta}_k$. Now, for contradiction, suppose $\hat{\theta}_k > \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$ and let this k be minimal. Then we can construct $\check{\boldsymbol{\theta}}$ by

$$\check{\theta}_l = \begin{cases} \sum_{l=k}^{k_j} w_l \bar{Y}_l / \sum_{l=k}^{k_j} w_l & \text{for } l = k, \ldots, k_j \\ \hat{\theta}_l & \text{otherwise.} \end{cases}$$

By appealing to the optimality of $\hat{\boldsymbol{\theta}}$, we can easily observe that $\hat{\theta}_{k-1} \leq \bar{Y}_{k-1}$ and therefore that the ordering of the entries of $\check{\boldsymbol{\theta}}$ matches that of $\hat{\boldsymbol{\theta}}$. Here, we use that $(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) \geq \gamma\lambda$.

We can now see that the loss term in $Q(\check{\boldsymbol{\theta}})$ is less than in $Q(\hat{\boldsymbol{\theta}})$, with a difference of more than $(\eta/4s)(\sqrt{2s/\eta}\sqrt{\gamma}\lambda)^2 = \gamma\lambda^2/2$, which outweighs the possible increase in the penalty contribution. This gives us that $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\boldsymbol{\theta}}$.

We now return to the proof of the main result. Suppose, for contradiction, that there exists some $j \in \{1, \ldots, s\}$ such that $\hat{\theta}_k \notin [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$ for all $k = k_{j-1} + 1, \ldots, k_j$ and let this j be minimal. By the first observation, we know that entries of $\hat{\boldsymbol{\theta}}$ corresponding to level j can take one of at most two distinct values. That is, for $k \in \{k_{j-1} + 1, \ldots, k_j\}$, if we have $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$, then it follows that $\hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}$. Similarly, if $\hat{\theta}_k > \bar{Y}_{k_j}$, then $\hat{\theta}_k = \hat{\theta}_{k_j}$.

By the assumption $w^0_{\min} \geq \eta/s$, we have that either

$$\sum_{k:\, \hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}} w_k \geq \frac{\eta}{2s} \quad \text{or} \quad \sum_{k:\, \hat{\theta}_k = \hat{\theta}_{k_j}} w_k \geq \frac{\eta}{2s}.$$

We will without loss of generality take the second statement to be true (the proof for the first case follows identically). Let k' denote the minimal element in $\{k_{j-1} + 1, \ldots, k_j\}$ such that $\hat{\theta}_{k'} = \hat{\theta}_{k_j}$. From the preliminary result established earlier, $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$. By appealing to the optimality of $\hat{\boldsymbol{\theta}}$, we see that $\hat{\theta}_{k_j+1} < \hat{\theta}_{k_j} + \gamma\lambda$ (otherwise, we could take $\hat{\theta}_{k_j}$ to be $\bar{Y}_{k_j}$ and strictly reduce the value of the objective).

Now, we will use that the separation is at least $2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda$. By our earlier observation (55), it is clear that any $l \in \{k_j + 1, \ldots, k_{j+1}\}$ with $\hat{\theta}_l < \bar{Y}_{k_j+1}$ has $\hat{\theta}_l = \hat{\theta}_{k_j+1}$. Note that since $\hat{\theta}_{k_j+1} - \bar{Y}_{k_j} < (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda$, it follows that $\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1} > (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta$ and therefore that $\sum_{\{k:\, \hat{\theta}_k = \hat{\theta}_{k_j+1}\}} w_k < \eta/2s$ by the preliminary result. Since $w^0_{\min} \geq \eta/s$ and separation (43) $\geq 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda + \zeta$, we can define $l' \in \{k_j + 1, \ldots, k_{j+1}\}$ minimal such that $\hat{\theta}_{l'} \geq \bar{Y}_{k_j+1}$.

Now, in order to contradict the optimality of $\hat{\boldsymbol{\theta}}$ we construct a new feasible point $\tilde{\boldsymbol{\theta}}$ by setting

$$\tilde{\theta}_l = \begin{cases} \bar{Y}_{k_j} & \text{for } l = k', \ldots, k_j \\ \hat{\theta}_{l'} & \text{for } l = k_j + 1, \ldots, l' - 1 \\ \hat{\theta}_l & \text{otherwise.} \end{cases}$$

It follows that for $l = k_j + 1, \ldots, l' - 1$ we have

$$\begin{aligned} |\hat{\theta}_l - \bar{Y}_l| &> (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta \\ |\tilde{\theta}_l - \bar{Y}_l| &\leq (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta. \end{aligned}$$

It is also straightforward to see that $|\hat{\theta}_{k_j} - \bar{Y}_l| \geq |\bar{Y}_{k_j} - \bar{Y}_l|$ for $l = k', \ldots, k_j$. If follows that the loss contribution in $Q(\tilde{\boldsymbol{\theta}})$ is strictly less than that in $Q(\hat{\boldsymbol{\theta}})$. Hence, using $\hat{\theta}_{l'} - \hat{\theta}_{k_j} > \gamma\lambda$, we obtain

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) >& \rho(\hat{\theta}_{l'} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{k'-1}) \\ & - \frac{1}{2}\gamma\lambda^2 - \rho(\bar{Y}_{k_j} - \hat{\theta}_{k'-1}) \\ \geq& 0, \end{aligned}$$

contradicting the optimality of $\hat{\boldsymbol{\theta}}$. We conclude that for $j = 1, \ldots, s$, there exists k_j^* in $\{k_{j-1} + 1, \ldots, k_j\}$ such that $\hat{\theta}_{k_j^*} \in [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$. □

Lemma 10. *Consider the univariate objective* (11)*, relaxing the normalisation constraint to* $\check{w} := \sum_k w_k \leq 1$*. Suppose that* $w^T\bar{Y} = 0$*, and that* $\|\bar{Y}\|_\infty < (2 \wedge \sqrt{\gamma\check{w}})\,\lambda/\check{w}$*. Then* $\hat{\boldsymbol{\theta}} = 0$*.*

Proof. Let $P_w = I - \mathbf{1}w^T/\check{w}$ and $D_w \in \mathbb{R}^{K\times K}$ be the diagonal matrix with entries $D_{kk}\sqrt{w_k}$. First note that

$$\begin{aligned} Q(\boldsymbol{\theta}) - Q(P_w\boldsymbol{\theta}) &= \frac{1}{2}\sum_{k=1}^K w_k(\bar{Y}_k - \theta_k)^2 - \frac{1}{2}\sum_{k=1}^K w_k(\bar{Y}_k - \theta_k + w^T\boldsymbol{\theta})^2 \\ &= -\frac{1}{2}\sum_{k=1}^K w_k(w^T\boldsymbol{\theta})(2\bar{Y}_k - 2\theta_k + w^T\boldsymbol{\theta}) \\ &= \left(1 - \frac{1}{2}\check{w}\right)(w^T\boldsymbol{\theta})^2 \ \geq 0. \end{aligned}$$

Thus for all $\boldsymbol{\theta} \in \mathbb{R}^K$, we have

$$\begin{aligned} Q(\boldsymbol{\theta}) - Q(0) &\geq \frac{1}{2}\|D_wP_w(\bar{Y} - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\bar{Y}\|_2^2 + \sum_{k=1}^{K-1}\rho(\theta_{(k+1)} - \theta_{(k)}) \\ &\geq \frac{1}{2}\|D_wP_w(\bar{Y} - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\bar{Y}\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)}) \\ &\geq \min_{\xi\in[-\tau,\tau]^K} F(\boldsymbol{\theta}, \xi, w) \end{aligned}$$

where

$$F(\boldsymbol{\theta}, \xi, w) = \frac{1}{2}\|D_wP_w(\xi - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\xi\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)}).$$

Consider minimising F over $\mathbb{R}^K \times [-\tau, \tau]^K \times S$, where $S \subseteq \mathbb{R}^K$ is the unit simplex scaled by $\check{w}$. We aim to show this minimum is 0. As with the first claim in the proof of Lemma 8, it

is straightforward to see that for any feasible $(\boldsymbol{\theta}, \xi, w)$, there exists $\boldsymbol{\theta}'$ with $\|\boldsymbol{\theta}'\|_\infty \leq \|\xi\|_\infty$ and $F(\boldsymbol{\theta}', \xi, w) \leq F(\boldsymbol{\theta}, \xi, w)$. Hence,

$$\inf_{(\boldsymbol{\theta},\xi,w)\in\mathbb{R}^K\times[-\tau,\tau]^K\times S} F(\boldsymbol{\theta}, \xi, w) = \inf_{(\boldsymbol{\theta},\xi,w)\in[-\tau,\tau]^K\times[-\tau,\tau]^K\times S} F(\boldsymbol{\theta}, \xi, w).$$

As on the RHS we are minimising a continuous function over a compact set, we know a minimiser must exist. Let $(\tilde{\boldsymbol{\theta}}, \tilde{\xi}, \tilde{w})$ be a minimiser (to be specified later). Observe that

$$\|D_{\tilde{w}} P_{\tilde{w}} (\xi - \boldsymbol{\theta})\|_2^2 - \|D_{\tilde{w}} P_{\tilde{w}} \xi\|_2^2 = -2\xi^T P_{\tilde{w}}^T D_{\tilde{w}}^2 P_{\tilde{w}} \boldsymbol{\theta} + \theta^T P_{\tilde{w}}^T D_{\tilde{w}}^2 P_{\tilde{w}} \boldsymbol{\theta}$$

is linear as a function of ξ. Hence it is minimised over the set

$$\{\xi : \|\xi\|_\infty \leq \tau\} = \mathrm{conv}(\{-\tau, \tau\}^K)$$

at some point in $\{-\tau, \tau\}^K$. Here $\mathrm{conv}(\cdot)$ denotes the convex hull operation. We thus have

$$Q(\boldsymbol{\theta}) - Q(0) \geq \min_{\xi\in\{-\tau,\tau\}^K} \frac{1}{2}\|D_{\tilde{w}} P_{\tilde{w}} (\xi - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_{\tilde{w}} P_{\tilde{w}} \xi\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)}).$$

Let us take $(\tilde{\boldsymbol{\theta}}, \tilde{\xi}) \in \mathbb{R}^K \times \{-\tau, \tau\}^K$ to be a minimiser of the RHS.

Note that if we have $\tilde{\xi}_j = \tilde{\xi}_k$ then we may take $\tilde{\theta}_j = \tilde{\theta}_k$. Indeed, we may construct $\check{\boldsymbol{\theta}} \in \mathbb{R}^K$ by setting

$$\check{\theta}_l = \begin{cases} \arg\min_{b\in\{\tilde{\theta}_j, \tilde{\theta}_k\}} (\tilde{\xi}_j - b)^2 & \text{for } l = j, k \\ \tilde{\theta}_l & \text{otherwise.} \end{cases}$$

Since the penalty contribution from $\check{\boldsymbol{\theta}}$ is not greater than that of $\tilde{\boldsymbol{\theta}}$, it follows that $Q(\check{\boldsymbol{\theta}}) \leq Q(\tilde{\boldsymbol{\theta}})$. Thus, we can assume that entries of $\tilde{\boldsymbol{\theta}}$ can take one of only two distinct values.

Next we write $\tilde{\alpha} = \sum_{k:\tilde{\xi}_k = -\tau} \tilde{w}_k$ and observe that $\tilde{w}^T \tilde{\xi} = (\check{w} - 2\tilde{\alpha})\tau$. Let us set $s = \min_k \tilde{\theta}_k$ and $x = \max_k \tilde{\theta}_k - \min_k \tilde{\theta}_k$. Then we have

$$\begin{aligned} F(\tilde{\boldsymbol{\theta}}, \tilde{\xi}, \tilde{w}) =& \frac{1}{2}\tilde{\alpha}\{(2\tilde{\alpha} - 1 - \check{w})\tau - s\}^2 + \frac{1}{2}(\check{w} - \tilde{\alpha})((2\tilde{\alpha} + 1 - \check{w})\tau - s - x)^2 \\ &+ \rho(x) - \frac{2}{\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})\tau^2 \\ =& \frac{1}{2\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})(2\tau - x)^2 + \rho(x) - \frac{2}{\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})\tau^2 \\ =& \frac{\check{w}}{8}(2\tau - x)^2 + \rho(x) - \frac{1}{2}\tau^2. \end{aligned} \tag{56}$$

In the second line above, we have solved for s to find that

$$s = \frac{1}{\check{w}}\{\tau(1 - \check{w})(\check{w} - 2\tilde{\alpha}) + (\tilde{\alpha} - \check{w})x\}.$$

In the third line above, we have solved for $\tilde{\alpha}$ to obtain $\tilde{\alpha} = \check{w}/2$ and hence $\tilde{\alpha}(\check{w} - \tilde{\alpha})/\check{w} = \check{w}/4$. These follow from optimality of $\tilde{\boldsymbol{\theta}}$ and $\tilde{w}$ respectively. The result follows from applying Lemma 7, setting $\kappa = \check{w}/4$. □

S2.2 Proof of Theorem 6

We begin by defining P^0 to be the orthogonal projection onto the linear space

$$V_0 = \left\{ \mu + \sum_{j=1}^{j} \sum_{k=1}^{K_j} \mathbb{1}_{\{X_{ij}=k\}} \theta_{jk} : (\mu, \theta) \in \mathbb{R} \times \Theta_0 \right\}.$$

The residuals from the oracle least-squares fit are $(I-P^0)Y = (I-P^0)\varepsilon$. The partial residuals $R^{(j)}$ as defined in (18) for the jth variable are therefore

$$R_i^{(j)} = \sum_{k=1}^{K_j} \mathbb{1}_{\{X_{ij}=k\}} \hat{\theta}^0_{jk} + \left[(I-P^0)\varepsilon\right]_i . \tag{57}$$

For $j = 1, \ldots, p$, we define $\bar{R}_k^{(j)} = \sum_{i=1}^n \mathbb{1}_{\{X_{ij}=k\}} R_i^{(j)}/n_{jk}$ for $k = 1, \ldots, K_j$, reordering the labels such that $\bar{R}_1^{(j)} \leq \cdots \leq \bar{R}_{K_j}^{(j)}$. We then aim to apply the arguments of Theorem 5 to $\hat{\boldsymbol{\theta}}_j$ defined by

$$\hat{\boldsymbol{\theta}}_j \in \operatorname*{arg\,min}_{\boldsymbol{\theta}_j \in \Theta_j} \frac{1}{2} \sum_{k=1}^{K_j} w_{jk} \left(\bar{R}_k^{(j)} - \theta_{jk}\right)^2 + \sum_{k=1}^{K_j - 1} \rho(\theta_{jk+1} - \theta_{jk}). \tag{58}$$

In order to do this, we define the events (for some τ_j to be determined later):

$$\Lambda_j^{(1)} = \left\{ |\hat{\theta}^0_{jk_l} - \theta^0_{jk_l}| \leq \tau_j : l = 1, \ldots, s_j \right\} \tag{59}$$

$$\Lambda_{jk}^{(2)} = \left\{ \left| \frac{1}{n_{jk}} \sum_{i=1}^n \mathbb{1}_{\{X_{ij}=k\}} ((I-P^0)\varepsilon)_i \right| < \frac{1}{2}\sqrt{\eta\gamma_{*j}s_j}\lambda_j \right\}. \tag{60}$$

On the intersection of events $\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)}$, we have that $|\bar{R}_k^{(j)} - \hat{\theta}^0_{jk}| < \sqrt{\eta\gamma_{*j}s_j}\lambda_j/2$. By following an identical approach to that involved in computing (35), we have that

$$\mathbb{P}\left(\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)}\right) \geq 1 - 2\exp\left(-\frac{n w_{j,\min} \eta\gamma_{*j} s_j \lambda_j^2}{8\sigma^2} + \log(K_j)\right),$$

where we recall that $w_{jk} = n_{jk}/n$.

We now turn our attention to the event $\Lambda_j^{(1)}$. Note that if $s_j = 1$, then this is immediately satisfied since $\hat{\boldsymbol{\theta}}_j^0 = \boldsymbol{\theta}_j^0 = 0$. If $s_j > 1$, we use that the oracle least squares estimate $\hat{\boldsymbol{\theta}}^0 = AY$ is a linear transformation A of the responses $(Y_i)_{i=1}^n$. For each $i = 1, \ldots, n$, Y_i has an independent (non-central) sub-Gaussian distribution with parameter σ. Therefore for each $k = 1, \ldots, K_j$, $\hat{\theta}^0_{jk} - \theta^0_{jk}$ also has a sub-Gaussian distribution, with parameter at most $\sigma c_{\min}^{-1/2}$ (recalling that $c_{\min} = (\max_l (AA^T)_{ll})^{-1}$). This enables us to show that

$$\mathbb{P}\left(\Lambda_j^{(1)}\right) \geq 1 - 2\exp\left(-\frac{c_{\min}\tau_j^2}{2\sigma^2} + \log(s_j)\right).$$

We can now set $\tau_j = \sqrt{\eta\gamma_{*j}s_j}\lambda_j/2$. From (26) and the triangle inequality, on the event $\Lambda_j^{(1)}$ we have that

$$\begin{aligned}\Delta(\hat{\boldsymbol{\theta}}_j^0) &\geq \Delta(\boldsymbol{\theta}_j^0) - \sqrt{\eta\gamma_{*j}s_j}\lambda_j \\ &\geq 3\left(1 + \frac{\sqrt{2}}{\eta}\right)\sqrt{\gamma_j\gamma_j^*}\lambda_j.\end{aligned}$$

Thus, on the intersection of events $\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)}\right)$, we can proceed as in the proof of Theorem 5 from (38), to conclude that $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}_j^0$.

It immediately follows that on the intersection of events $\cap_{j=1}^{p}\left(\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j}\Lambda_{jk}^{(2)}\right)\right)$, we have $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$. By a union bound, this occurs with probability at least

$$\begin{aligned}\mathbb{P}\left(\cap_{j=1}^{p}\left(\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j}\Lambda_{jk}^{(2)}\right)\right)\right) &\geq 1 - 2\sum_{j=1}^{p}\left[\exp\left(-\frac{n_{j,\min}\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(K_j)\right)\right. \\ &\qquad \left. + \exp\left(-\frac{c_{\min}\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(s_j)\right)\right] \\ &\geq 1 - 4\sum_{j=1}^{p}\exp\left(-\frac{(n_{j,\min} \wedge c_{\min})\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(K_j)\right),\end{aligned}$$

where in the final line we use $s_j \leq K_j$. □

S3 Additional experimental information

S3.1 Details of methods

Tree-based methods

We used the implementation of the random forest procedure [Breiman, 2001] in the R package `randomForest` [Liaw and Wiener, 2002] with default settings. CART [Breiman et al., 1984] was implemented in the R package `rpart` [Therneau and Atkinson, 2019], with pruning according to the 1-SE rule (as described in the package documentation).

CAS-ANOVA

The CAS-ANOVA estimator $\hat{\boldsymbol{\theta}}^{\text{cas}}$ optimises over $(\mu, \boldsymbol{\theta})$ a sum of a squared loss term (3) and an all-pairs penalty term (4). In particular, Bondell and Reich [2009] consider two regimes of weight vectors w. The first is not data-dependent and sets $w_{j,k_1k_2} = (K_j+1)^{-1}\sqrt{n_{jk_1} + n_{jk_2}}$. The second, 'adaptive CAS-ANOVA', uses the ordinary least squares estimate for $\boldsymbol{\theta}$ to scale the weights. Here, $w_{j,k_1k_2} = (K_j+1)^{-1}\sqrt{n_{jk_1} + n_{jk_2}}|\hat{\theta}_{jk_1}^{\text{OLS}} - \hat{\theta}_{jk_2}^{\text{OLS}}|^{-1}$.

Here we introduce a new variant of adaptive CAS-ANOVA, following ideas in Bühlmann and Van De Geer [2011] for a 2-stage adaptive Lasso procedure. Instead of using the ordinary least squares estimate $\hat{\boldsymbol{\theta}}^{\text{OLS}}$ in the above expression, an initial (standard) CAS-ANOVA estimate is used to scale the weights, with λ selected for the initial estimate by 5-fold cross-validation. In simulations, this outperformed the adaptive CAS-ANOVA estimate using ordinary least squares initial estimates so in the interests of time and computational resources this was omitted from the simulation study. Henceforth adaptive CAS-ANOVA will refer to this 2-stage procedure.

The authors describe the optimisation of $\hat{\boldsymbol{\theta}}^{\text{cas}}$ as a quadratic programming problem, which was solved using the R package `rosqp` [Anderson, 2018]. Here we used our own implementation of the quadratic programming approach described by the authors. We found it considerably faster than the code available from the authors' website, and uses ADMM-based optimisation [Boyd et al., 2011] tools not available at the time of its publication. We also found, as discussed in Section 5.1 of Maj-Kańska et al. [2015], that we could not achieve the best results using the publicly available code. Lastly, using our own implementation allowed us to explore a modification of CAS-ANOVA using the more modern approach of adaptive weights via a 2-stage procedure [Bühlmann and Van De Geer, 2011] to compare SCOPE to a wider class of all-pairs penalty procedures.

For large categorical variables, solutions are slow to compute and consume large amounts of memory. In the case of binary response, CAS-ANOVA models were fitted iterating a locally quadratic approximation to the loss function.

DMR

The DMR algorithm [Maj-Kańska et al., 2015] is implemented in the R package `DMRnet` [Prochenka-Sotys and Pokarowski, 2018]. The degrees of freedom in the model is decided by 5-fold cross-validation. It is based on pruning variables using the Group Lasso [Yuan and Lin, 2006] to obtain at a low-dimensional model, then performing backwards selection based on ranking t-statistics for hypotheses corresponding to each fusion between levels in categorical variables.

The cross-validation routine appeared to error when all levels of all categorical variables were not present in one of the folds. In Section 6.2, cross-validation was therefore not possible so model selection was performed based on Generalized Information Criterion (GIC) [Zheng and Loh, 1995]. In all other examples, models were selected via 5-fold cross-validation.

Bayesian effect fusion

In Section 6.1.1 we include Bayesian effect fusion [Pauger and Wagner, 2019], implemented in the R package `effectFusion` [Pauger et al., 2019]. Coefficients within each categorical variable were modelled with a sparse Gaussian mixture model. The posterior mean was estimated with 1000 samples.

Lasso

In Section 6.1.2 we also include Lasso [Tibshirani, 1996] fits, to serve as a reference point. Of course, this is unsuitable for models where levels in categorical variables should be clustered together, but the advanced development of the well-known R package `glmnet` [Friedman et al., 2010] nevertheless sees its use in practice.

In order to make the fit symmetric across the categories within each variable, models were fitted with an unpenalised intercept and featuring dummy variables for all of the categories within each variable. This is instead of the corner-point dummy variable encoding of factor variables that is commonly used when fitting linear models. Models are fitted and cross-validated with `cv.glmnet` using the default settings.

SCOPE

For SCOPE, we have provided the R package `CatReg` [Stokell, 2021]. The univariate update step (see Section 3.1) is implemented in C++ using Rcpp [Eddelbuettel and François, 2011], with models fitted using a wrapper in R. For the binary response case, the outer loop to iterate the local quadratic approximations in the proximal Newton algorithm are done within R. In the future, performance could be improved by iterating the univariate update step (and the local quadratic approximations, as in Sections 6.2 and 6.3) within some lower-level language. In higher-dimensional experiments, SCOPE was slowed by cycling through all the variables; an active-set approach to this could make it faster still.

S3.2 Further details of numerical experiments

For the experiments in Section 6.1, we define the signal-to-noise ratio (SNR) as σ_S/σ, where σ_S is the standard deviation of the signal $Y - \varepsilon$, and σ is the standard deviation of the noise ε.

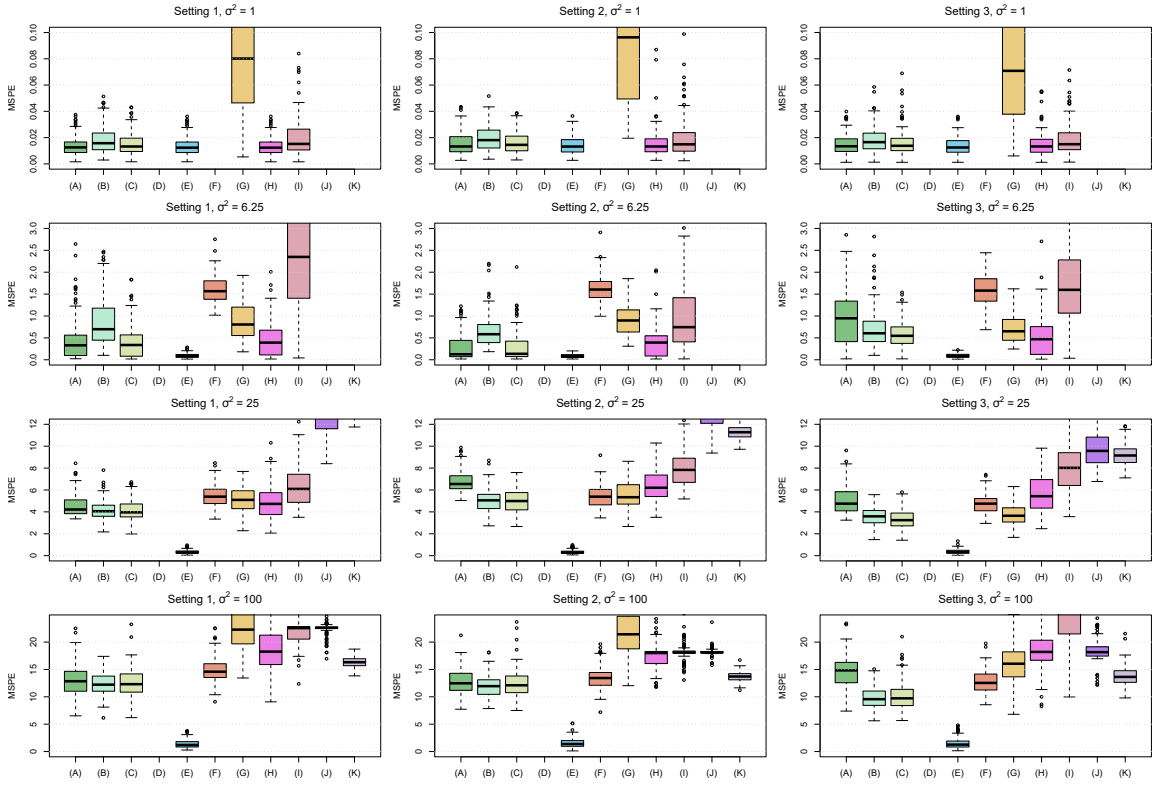

Figure 9: Prediction performance of various methods: (A) SCOPE-8; (B) SCOPE-32; (C) SCOPE-CV; (D) Linear regression; (E) Oracle least squares; (F) CAS-ANOVA; (G) Adaptive CAS-ANOVA; (H) DMR; (I) BEF; (J) CART; (K) RF. Note that some 'boxes' are not visible in some of the plots; this is due to the MSPE in the tests being beyond the range of the plot.

S3.2.1 Low-dimensional simulations

In Table 7 we include details of computation time and dimension of the fitted models. Figure 9 visualises the results also summarised in Table 1 in the main paper.

	Mean fitted model dimension				
σ^2:	1	6.25	25	100	Mean computation time (s)
SCOPE-8	7.2	8.5	4.7	4.3	16
SCOPE-32	9.6	12.6	13.2	9.8	48
SCOPE-CV	7.9	10.3	16.8	10.9	68
Oracle least squares	7.0	7.0	7.0	7.0	0.00
Linear regression	231.0	231.0	231.0	231.0	0.01
CAS-ANOVA	35.2	70.0	74.3	52.4	4679
Adaptive CAS-ANOVA	13.4	31.3	36.9	32.5	9659
DMR	7.0	7.2	5.3	2.7	21
BEF	7.3	6.3	4.1	2.0	975
CART					0.01
RF					0.66

Table 7: Mean fitted model dimension and computation time for the various methods.

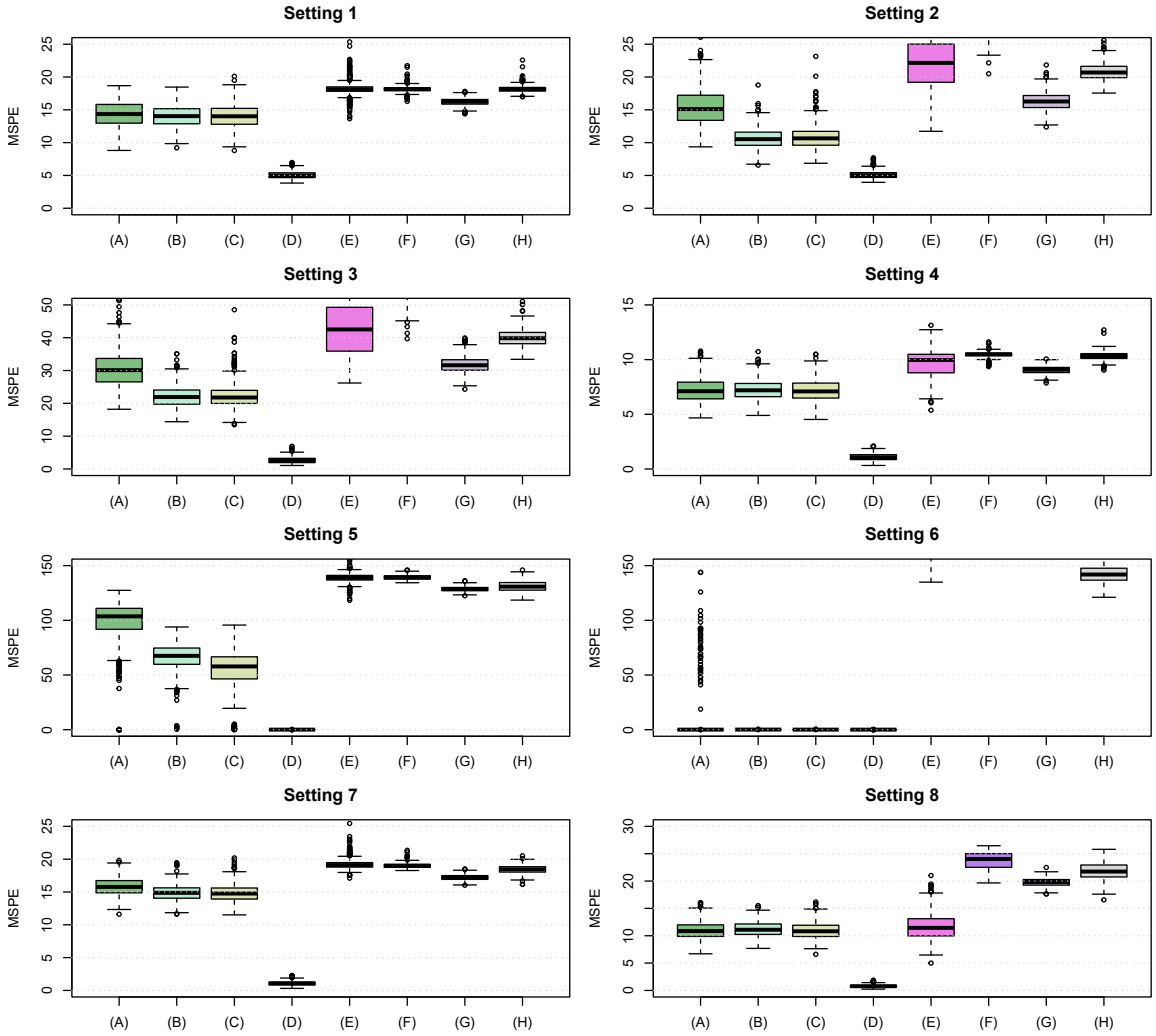

Figure 10: Prediction performance of various methods: (A) SCOPE-8; (B) SCOPE-32; (C) SCOPE-CV; (D) Oracle least squares; (E) DMR; (F) CART; (G) RF; (H) Lasso. Note that some 'boxes' are not visible in some of the plots; this is due to the MSPE in the tests being beyond the range of the plot.

S3.2.2 High-dimensional simulations

Here we include additional results relating to the high-dimensional experiments. Figure 10 visualises the results in Table 2 of the main paper.

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	224	322	348	76	234	518	209	175
SCOPE-32	134	341	502	51	283	650	113	161
SCOPE-CV	951	1739	2450	332	1516	2892	767	902
DMR	26	38	39	26	30	36	30	29
CART	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
RF	5.7	5.7	5.9	2.7	5.8	5.8	5.9	5.8
Lasso	1.5	1.5	1.6	1.2	1.4	1.5	1.5	1.5

Table 8: Mean computation time (s)

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	6.9	9.4	9.8	6.9	21.3	27.1	9.3	7.2
SCOPE-32	20.7	37.5	38.0	19.9	75.8	26.1	32.9	31.3
SCOPE-CV	21.4	40.4	40.8	19.5	103.7	26.2	36.6	17.9
DMR	1.9	4.9	4.7	3.4	3.7	22.8	2.3	7.5
Lasso	15.7	167.1	152.0	32.7	143.7	469.7	35.8	82.8

Table 9: Mean fitted model dimension

Setting	γ:	4	8	16	32	64
1		0.028	0.290	0.196	0.138	**0.348**
2		0.002	0.016	0.234	0.298	**0.450**
3		0.006	0.012	0.286	0.248	**0.448**
4		0.030	**0.356**	0.244	0.100	0.270
5		0.000	0.000	0.026	0.070	**0.904**
6		0.000	0.000	0.464	**0.534**	0.002
7		0.006	0.092	0.234	0.144	**0.524**
8		0.264	**0.446**	0.102	0.018	0.170

Table 10: Proposition of times each γ was selected by cross-validation.

References

E. Anderson. *rosqp: Quadratic Programming Solver using the 'OSQP' Library*, 2018. R package version 0.1.0.

H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1):169–177, 2009.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232, 2011.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

A. Maj-Kańska, P. Pokarowski, A. Prochenka, et al. Delete or merge regressors for linear model selection. *Electronic Journal of Statistics*, 9(2):1749–1778, 2015.

D. Pauger and H. Wagner. Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, 14(2):341–369, 2019.

D. Pauger, M. Leitner, H. Wagner, and G. Malsiner-Walli. *effectFusion: Bayesian Effect Fusion for Categorical Predictors*, 2019. R package version 1.1.1.

A. Prochenka-Sotys and P. Pokarowski. *DMRnet: Delete or Merge Regressors Algorithms for Linear and Logistic Model Selection and High-Dimensional Data*, 2018. R package version 0.2.0.

B. Stokell. CatReg: Solution Paths for Linear and Logistic Regression Models with Categorical Predictors, with SCOPE Penalty `https://CRAN.R-project.org/package=CatReg`, 2021.

T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

X. Zheng and W.-Y. Loh. Consistent variable selection in linear models. *Journal of the American Statistical Association*, 90(429):151–156, 1995.