ONLINE SPARSE ADDITIVE LEARNING WITHOUT A PRESPECIFIED REPORTING HORIZON

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating sparse nonlinear effects from a growing data stream requires updating both the selected variables and their uncertainty. Correlated covariates can give an irrelevant variable a larger initial effect than a relevant one, causing thresholding to fail. The sample means used to center effects also vary across samples. In this article, we propose an online sparse additive method in which selection, estimation, and centering share one kernel-weighted least-squares criterion computed from stored sums. Joint re-estimation with smoothly clipped absolute deviation weights consistently selects the relevant variables in constructed correlated models where initial thresholds fail. Under population selection margins and identifiable correlated effects, continuous-effect estimates attain the known-support optimal rate for a stationary repeat-or-redraw covariate process. With Gaussian autoregressive errors and bounded pairwise density ratios, we establish asymptotically valid joint confidence intervals for a fixed finite set of effect values at each prespecified reporting time. The final sample size need not be specified. For a fixed number of covariates, storage has an asymptotically sublinear bound. In simulations with strong signals, recomputed centers correct the undercoverage of fixed centers. Our method also lowers estimation error against the compared planned-epoch and stochastic-gradient estimators. Bike-rental and appliance energy analyses identify stable sparse nonlinear effects.

1 INTRODUCTION

In many applications, observations arrive over time, and the response depends nonlinearly on a few of many covariates. For example, bike rental demand and household energy use require tracking relevant variables and their nonlinear effects. Sparse additive models describe this dependence through one univariate effect per selected covariate, adjusted for the others.

Offline methods estimate these additive effects (Meier, van de Geer and Bühlmann, 2009; Raskutti, Wainwright and Yu, 2012) and select the relevant covariates (Ravikumar et al., 2009; Lee et al., 2024) by analyzing a fixed dataset. Sparse kernel expansions, tilted empirical risks, and random-feature pruning address nonlinear feature selection (Liu, Chen and Huang, 2020; Wang et al., 2023; Xie et al., 2022). For streaming data, online nonparametric methods estimate regression functions sequentially (Tarrès and Yao, 2014; Dieuleveut and Bach, 2016; Zhang and Simon, 2022; Chen and Klusowski, forthcoming), and online smooth backfitting adapts additive estimation and includes component tests for an identity link (Yang, Yao and Zhao, 2024). Other streaming studies analyze memory-constrained estimation (Quan and Lin, 2024), sparse linear prediction with limited covariate access (Kale et al., 2017), coefficient estimation (Yang et al., 2023), and sparse additive bandit decisions (Wang, Zhang and Zhang, 2026).

Offline inference concerns fixed-sample components (Gregory, Mammen and Wahl, 2021; Lu, Kolar and Liu, 2020) and derivatives (Guo, Yuan and Zhang, 2026). Online methods provide inference for regression functions, time series, smoothed nonstationary trends, single-index parameters, and regression and policy parameters (Liu, Shang and Yang, 2026; Palm and Nagler, 2024; Nagler et al., 2026; Han et al., 2024; Chen, Tamer and Yao, 2026). We study fully observed streams and the memory needed for component estimates. To the best of our knowledge, joint confidence intervals for selected, population-centered additive effects have not been established for streaming procedures with sublinear storage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this article, we propose an online sparse additive method that updates effects and intervals from stored sums without a prespecified final sample size. We study the additive model $Y_t = \mu + \sum_{j=1}^{d} f_j(X_{tj}) + \varepsilon_t$. Here Y_t is the response, X_{tj} are covariates, μ is the intercept, and the error ε_t has conditional mean zero given the covariates. Each effect $f_j(X_{tj})$ has mean zero under its population covariate distribution. Selection, estimation, and centering share one kernel-weighted least-squares criterion computed from stored additive sums. We represent each continuous effect by local lines at grid nodes, following the smooth backfitting criterion (Mammen, Linton and Nielsen, 1999). We select variables using smoothly clipped absolute deviation (SCAD) weights (Fan and Li, 2001; Zou and Li, 2008) computed from initial effect norms and refit the selected effects without penalty.

Two obstacles motivate this construction. First, correlated covariates can give an irrelevant variable a larger initial effect than a relevant one. No common cutoff on these initial sizes can then recover the active set. A cutoff that removes the irrelevant variable also removes the relevant one. Existing results show when selection can still succeed using one-step and nonconvex penalties (Zou and Li, 2008; Fan, Xue and Zou, 2014; Loh and Wainwright, 2017). We construct smooth correlated additive models in which thresholding and the adaptive group Lasso (Huang, Horowitz and Wei, 2010) with the specified weights fail. Our joint re-estimation with SCAD weights selects the relevant variables consistently in these models.

Second, adding a constant to one component and subtracting it from the intercept leaves the regression function unchanged. We fix these constants through zero population means. Because the covariate distribution is unknown, we use empirical centering. For continuous effects, center variation is of smaller asymptotic order than local estimation error. It remains relevant in finite samples and has the same order as categorical estimation error. We account for it by deleting each time block's sums and refitting with recomputed centers.

This article makes three contributions. First, our continuing estimator updates the selected set and component estimates without a prespecified horizon. Under the stated selection and dependence conditions, each continuous effect has expected integrated squared error of order at most $n^{-4/5}$. This known-support rate is optimal on a common signal class. The result permits identifiable correlated effects with the stated population selection margins. For fixed dimension, storage has an upper bound of order $n^{9/10}$ scalars up to iterated-logarithm factors, with probability tending to one. Second, we establish asymptotically valid joint intervals after selection for a fixed finite set of effect values at each prespecified reporting time. Inference under correlation also requires bounded pairwise density ratios. Block resampling (Künsch, 1989) with recomputed centers accounts for temporal dependence and center variation. In simulations with strong signals, fixing the centers reduces joint coverage from 0.930–0.975 to 0.505–0.790. Third, an exact representation of the stored sums reduces planned-epoch storage without changing estimates or intervals. Simulations compare selection, component error and coverage.

2 Estimation and Sequential Updates

At time $t = 1, 2, \ldots$, we observe a scalar response Y_t and the full covariate vector $X_t = (X_{t1}, \ldots, X_{td})^{\mathsf{T}}$ of dimension d. We assume the sequence is stationary and satisfies the additive regression model

$$Y_t = \mu + \sum_{j=1}^{d} f_j(X_{tj}) + \varepsilon_t, \qquad \mathbb{E}(\varepsilon_t \mid X_t) = 0.$$

Here ε_t is the error and μ is the intercept. The effect $f_j(x)$ gives the contribution of covariate j at x to the conditional mean. We estimate continuous curves and categorical level effects. Continuous covariates take values in $[0, 1]$ and have twice differentiable effects. Categorical covariates have finitely many levels.

For example, in the bike rental data of Section 4, Y_t is the log of one plus the rental count in hour t. The $d = 56$ candidates include six recorded variables and 50 synthetic nulls. The descriptive fit selects temperature X_{t1}, humidity X_{t2}, hour X_{t3}, and the demand 24 hours earlier X_{t4}. With hats for fitted quantities, the model is

$$\widehat{Y}_t = \widehat{\mu} + \widehat{f}_{\mathrm{temp}}(X_{t1}) + \widehat{f}_{\mathrm{hum}}(X_{t2}) + \widehat{f}_{\mathrm{hour}}(X_{t3}) + \widehat{f}_{\mathrm{lag}}(X_{t4}),$$

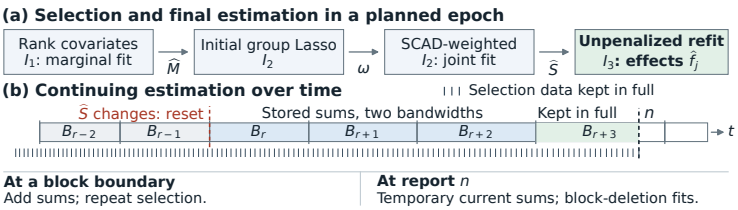


Figure 1: Variable selection, final estimation and sequential updates.

and $\widehat{f}_{\text{hour}}$ is one categorical effect with 24 levels. Appendix Section S3 shows that separate unrestricted sine and cosine effects would not be identifiable. The other 52 estimated effects are zero. Figure 1 outlines the three estimation steps.

We write P_j for the stationary marginal distribution of X_{tj}. We impose $\mathbb{E}f_j(X_{tj}) = 0$ for every j. This gives $\mu = \mathbb{E}Y_t$ and separates each effect from the overall mean. The active set $S = \{j : f_j \neq 0\}$ has size $s = |S|$ small relative to d. The population functions and active set remain fixed over time. Their estimates change as observations arrive.

2.1 A LEAST-SQUARES CRITERION FOR THE COMPONENT EFFECTS

We define G equally spaced grid nodes $u_1 < \cdots < u_G$ on $[0, 1]$ and a bandwidth $h > 0$. The local line at u_g is $\ell_{jg}(x) = \beta_{jg,0} + \beta_{jg,1}(x - u_g)/h$. Here $\beta_{jg,0}$ and $\beta_{jg,1}$ are the intercept and slope in rescaled distance. The nonnegative kernel weights $\pi_{jg}(x)$ sum to one over g. They are positive only for nodes within distance h of x.

With hats for estimated coefficients, the reported curve averages the neighboring local lines:

$$\widehat{f}_j(x) = \sum_{g=1}^{G} \pi_{jg}(x)\{\widehat{\beta}_{jg,0} + \widehat{\beta}_{jg,1}(x - u_g)/h\}.$$

For a categorical covariate, $\ell_{j1}(x)$ is its effect at level x. We set $\pi_{j1}(x) = 1$ for every level. The group β_j contains local intercepts and slopes for covariate j, or categorical level effects.

For an observation block I, we write $n_I = |I|$ and $\bar{Y}_I = n_I^{-1} \sum_{i \in I} Y_i$. For each group j, the centering constraint and weighted norm are

$$\frac{1}{n_I} \sum_{i \in I, g} \pi_{jg}(X_{ij})\ell_{jg}(X_{ij}) = 0, \qquad \|\beta_j\|_I^2 = \frac{1}{n_I} \sum_{i \in I, g} \pi_{jg}(X_{ij})\ell_{jg}(X_{ij})^2.$$

The zero average over observed covariate values separates the fitted component from the intercept $\bar{Y}_I$. The norm measures effect size as the weighted root mean square of local values. For a vector of nonnegative weights $\omega = (\omega_1, \ldots, \omega_d)$, we define the objective

$$L_I(\beta; \omega) = \frac{1}{2n_I} \sum_{i \in I} \sum_{g_1, \ldots, g_d} \Big\{ Y_i - \bar{Y}_I - \sum_{j=1}^{d} \ell_{jg_j}(X_{ij}) \Big\}^2 \prod_{j=1}^{d} \pi_{jg_j}(X_{ij}) + \sum_{j=1}^{d} \omega_j \|\beta_j\|_I. \tag{1}$$

We minimize $L_I(\beta; \omega)$ over $\beta = (\beta_1^\mathsf{T}, \ldots, \beta_d^\mathsf{T})^\mathsf{T}$ subject to the centering constraints. Here g_j indexes the G nodes for a continuous covariate or the single categorical term. With one continuous covariate and $\omega_j = 0$, we fit the local lines separately by least squares with weights $\pi_{jg}(X_{ij})$. For several covariates, product weights distribute observations over neighboring nodes for a joint fit. The first term discretizes the smooth backfitting integral of Mammen, Linton and Nielsen (1999) by summing over nodes. The second term is the local-linear counterpart of the functional Lasso penalty (Lee et al., 2024). This penalty removes covariates by setting coefficient groups to zero.

2.2 VARIABLE SELECTION AND FINAL ESTIMATES

For a planned reporting time, we use screening, selection, and estimation blocks I_1, I_2, I_3. Unused intervals between blocks reduce temporal dependence. We write $L_I(\beta; c) = L_I(\beta; (c, \ldots, c))$ for

$c \geq 0$. For a prescribed $1 \leq m \leq d$, we retain the m largest one-covariate loss reductions (Fan, Feng and Song, 2011):

$$\Delta_j = L_{I_1}(0;0) - \min_{\beta:\, \beta_k = 0 \ (k \neq j)} L_{I_1}(\beta;0), \qquad j = 1, \ldots, d,$$
$$\widehat{M} \in \operatorname*{arg\,max}_{M \subset \{1,\ldots,d\}:\, |M| = m} \sum_{j \in M} \Delta_j.$$

Here Δ_j is the loss decrease from fitting covariate j alone. On block I_2, we set $\beta_j = 0$ for $j \notin \widehat{M}$. We write $u \wedge v = \min(u, v)$ and $u \vee v = \max(u, v)$. For $\lambda > 0$, $a > 2$, and $t_+ = t \vee 0$, we compute the group-Lasso fit (Yuan and Lin, 2006), the SCAD fit, and the final estimate:

$$\widehat{\beta}^L \in \arg\min_{\beta} L_{I_2}(\beta;\lambda), \quad \omega_j = \lambda \wedge \frac{(a\lambda - \|\widehat{\beta}^L_j\|_{I_2})_+}{a-1},$$
$$\widehat{\beta}^{\mathrm{SCAD}} \in \arg\min_{\beta} L_{I_2}(\beta;\omega), \quad \widehat{S} = \{j \in \widehat{M} : \widehat{\beta}^{\mathrm{SCAD}}_j \neq 0\}, \qquad \widehat{\beta} \in \operatorname*{arg\,min}_{\beta:\, \beta_j = 0 \ (j \notin \widehat{S})} L_{I_3}(\beta;0).$$

We write $p_{\lambda,a}$ for the SCAD penalty (Fan and Li, 2001). The weight is its derivative, $\omega_j = p'_{\lambda,a}(\|\widehat{\beta}^L_j\|_{I_2})$. The weight equals λ when the initial norm of group j is at most λ. It decreases linearly to zero as the norm grows to $a\lambda$. Beyond this threshold, zero weights leave large initial effects unpenalized in the joint refit. Zero coefficient groups are removed. The unpenalized fit on I_3 removes shrinkage. Appendix Sections S1, S2, and S18 give the matrix formulation, its solutions, numerical tolerances and the centering error of exact affine fits.

In the population quadratic model of Appendix Proposition S5, the Gaussian score correlations are $(\rho_{12}, \rho_{13}, \rho_{23}) = (-0.7, 0.05, -0.55)$. The third covariate is inactive. For $a = 3.7$, the true coefficient magnitudes $|\theta^*|$, the initial group-Lasso magnitudes $|\theta^L|$, and the SCAD weights w satisfy

$$|\theta^*| \approx \lambda(5.26,\ 9.74,\ 0), \qquad |\theta^L| \approx \lambda(1.01,\ 5.00,\ 1.40), \qquad w \approx \lambda(0.996,\ 0,\ 0.853).$$

No cutoff on $|\theta^L|$ recovers the active pair. SCAD leaves the largest effect unpenalized and recovers this pair by weighted refitting. Adaptive weights $\lambda^2/|\theta^L_j|$ fail here.

2.3 Updating estimates and confidence intervals

After the planned reporting time, we update effects and intervals from stored sums. We define the weighted local mean and variance at observation i by

$$m_{ij} = \sum_g \pi_{jg}(X_{ij})\ell_{jg}(X_{ij}), \qquad v_{ij} = \sum_g \pi_{jg}(X_{ij})\{\ell_{jg}(X_{ij}) - m_{ij}\}^2.$$

Expanding (1) gives

$$L_I(\beta;0) = \frac{1}{2n_I} \sum_{i \in I} \Big[\Big\{ Y_i - \bar{Y}_I - \sum_j m_{ij} \Big\}^2 + \sum_j v_{ij} \Big].$$

Only within-group and pairwise cross moments enter this quadratic. We need not enumerate grid-node combinations. We write $\mathcal{M}_I$ for the count, response, design, and cross-product sums detailed in Appendix Section S20. Additivity gives $\mathcal{M}_{I \cup B} = \mathcal{M}_I + \mathcal{M}_B$ for disjoint blocks. It also gives $\mathcal{M}_{I \setminus B} = \mathcal{M}_I - \mathcal{M}_B$ for a stored block $B \subset I$.

We divide the stream into consecutive blocks B_r for $r = 1, 2, \ldots$ of increasing length. We retain about $n^{3/4}$ of the first n observations individually for selection. First, between reports, we store the full open block B_r without refitting. Second, we replace completed blocks by moments and add them to the totals. Screening and selection use the retained observations at the update times in Appendix Section S9. A support change clears the stored moments and restarts accumulation with block B_{r+1}. Third, reports use stored moments and temporary moments of the open block. The stored state is unchanged. The continuing estimate from observations I in these moments is

$$\widehat{\beta}^{\mathrm{C}} \in \operatorname*{arg\,min}_{\beta:\, \beta_j = 0 \ (j \notin \widehat{S})} L_I(\beta;0).$$

Completed blocks keep their original kernel weights. The incomplete block uses the reporting bandwidth. The local lines $\ell_{jg}(x) = \beta_{jg,0} + \beta_{jg,1}(x - u_g)$ use unscaled distance. Slopes have the same meaning across blocks. Finer node grids accumulate their own moments before use. Appendix Sections S9 and S20.8 give the block, grid, and storage rules.

Confidence intervals use separate moments for the same blocks with a smaller bandwidth. This undersmoothing keeps the smoothing bias small relative to standard errors. We call this set the inference state and the other the point-estimation state. We write $\widehat{f} \in \mathbb{R}^M$ for the estimates at M component-value pairs indexed by k. We compute these from the inference state. The N observations of the inference state form $J \geq 2$ consecutive blocks $I_1, \ldots, I_J$ of sizes $n_1, \ldots, n_J$. We subtract $\mathcal{M}_{I_b}$ and refit. We keep the selected set, grid, and kernel weights, but recompute the centers. This gives the deleted-block estimate $\widehat{f}_{-b}$. As in (S6), we define

$$V_b = (N - n_b)(\widehat{f} - \widehat{f}_{-b}), \qquad \widetilde{V}_b = V_b - \frac{n_b}{N} \sum_{a=1}^{J} V_a, \qquad \widehat{\Sigma} = \frac{J}{(J-1)N^2} \sum_{b=1}^{J} \widetilde{V}_b \widetilde{V}_b^{\mathsf{T}}.$$

The vector V_b approximates the contribution of block b to the error of $\widehat{f}$. Here $\widetilde{V}_b$ is the centered contribution. The matrix $\widehat{\Sigma}$ is their sample covariance times J/N^2. The standard error of $\widehat{f}_k$ is $\widehat{\Sigma}_{kk}^{1/2}$. Deleting whole blocks preserves dependence between nearby observations within each contribution (Künsch, 1989). Recomputed centers add the derivative term in Appendix Proposition S3.

Because the M intervals must hold at once, they share one critical value c^* replacing the normal quantile for one interval. To obtain this critical value, we draw J vectors independently with replacement from $\widetilde{V}_1, \ldots, \widetilde{V}_J$. We write $\bar{V}_k^*$ and s_k^* for the mean and sample standard deviation of their kth coordinates. The ratio $\sqrt{J}\,|\bar{V}_k^*|/s_k^*$ imitates the studentized error of $\widehat{f}_k$. The maximum over k imitates the largest error. For $0 < \alpha < 1$, we compute

$$T^* = \max_{1 \leq k \leq M} \frac{\sqrt{J}\,|\bar{V}_k^*|}{s_k^*}, \qquad c^* = \inf\{c : \overset{*}{\Pr}(T^* \leq c) \geq 1 - \alpha\}.$$

We write $\Pr^*$ for the resampling probability given the data. The intervals are $[\widehat{f}_k - c^* \widehat{\Sigma}_{kk}^{1/2}, \widehat{f}_k + c^* \widehat{\Sigma}_{kk}^{1/2}]$ for $k = 1, \ldots, M$. Section 3 gives conditions for joint coverage with probability tending to $1 - \alpha$. Appendix Sections S6 and S20 give resampling details and the identifiability test. Unidentified evaluations count as inference failures.

3 Statistical Guarantees

We write $\|\cdot\|_2$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_\infty$ for the Euclidean, spectral, and supremum norms. We write $\lambda_{\min}$ and $\lambda_{\max}$ for the smallest and largest eigenvalues. For positive sequences, $a_n \lesssim b_n$ means $a_n \leq C b_n$ for a constant independent of n. Its dependence on model parameters is specified in each result. $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. O and o denote bounded ratios and ratios tending to zero. O_p and o_p denote ratios bounded in probability and converging to zero in probability. $\longrightarrow_p$ and $\to_p$ denote convergence in probability. We fix constants $B > 0$ and $0 \leq \rho_0 < 1$.

Condition 1. (i) Continuous components have uniformly bounded second derivatives; their covariate densities are bounded above and away from zero. Categorical groups have a fixed number of at least two levels with probabilities uniformly bounded away from zero. (ii) $\mathbb{E} f_j(X_{tj}) = 0$ for every j and $\sum_j \|f_j\|_\infty^2 \leq B^2$. (iii) The coordinates of X_t are independent under the stationary distribution, and at each time the process retains the previous vector with probability $\rho \leq \rho_0 < 1$ and draws a new independent vector with the remaining probability. (iv) The errors form a stationary Gaussian autoregression independent of the covariate process, with autoregression coefficient of absolute value at most a fixed constant less than one and variance bounded above and away from zero.

Condition 1 is a reference model with independent covariate coordinates and temporal dependence. Selection, planned-epoch and continuing guarantees also allow the correlated designs specified below.

3.1 SCAD selection beyond thresholding

At selection sample size n, we use G, h_L, and λ_n for the grid size, bandwidth, and penalty. For a stationary observation (X, Y), $z_j(X_j)$ is the centered, normalized design vector for group j defined

in Appendix Section S10. We call $\|\mathbb{E}\{z_j(X_j)(Y-\mu)\}\|_2^2$ its marginal score. Screening estimates this score. Screening with these population scores retains candidates $M \supset S$, with $m = |M|$. The sampling-error bounds for moment matrices and scores use the scale

$$r_n = \left\{\frac{G\log(dn)}{n}\right\}^{1/2} + \frac{G\log n\log(dn)}{n} + h_L^2\left\{\frac{G\log(dn)}{n}\right\}^{1/2},$$

and this scale includes the contribution of the projection residual. The population quadratic loss has matrix Γ_M on M and minimizer θ^*. The population group-Lasso coefficient θ^L determines the SCAD weights $w_j = p'_{\lambda_n,a}(\|\theta_j^L\|_2)$. The vector $\bar{\theta}$ minimizes this weighted population criterion with zero coefficients outside S. Appendix Condition S1 states the remaining premises. These include the density and level-probability bounds of Condition 1(i). They also require normalized design vectors with norms $O(\sqrt{G})$, bounded coefficients, and an additive projection residual of order h_L^2 with $\lambda_n = O(1)$. Finally, they assume a stationary transition $P(x,\cdot) = (1-\rho)Q(\cdot) + \rho K(x,\cdot)$ with $0 \le \rho \le \rho_0 < 1$. Here Q is the stationary covariate distribution, and K is a Markov transition that preserves it. This transition structure permits correlated covariates through Q and temporal dependence through K. Appendix Proposition S4 verifies these bounds for a class of correlated designs.

The following theorem establishes exact selection when active coefficient norms and the gaps from inactive residual scores to their penalty weights exceed the sampling error.

Theorem 1. *Suppose that Appendix Condition S1 and Condition 1(iv) hold,* $\lambda_{\min}(\Gamma_M) \ge \kappa_\Gamma$, $mr_n = o(\lambda_n)$ *and*

$$\min_{j\in S}\|\bar{\theta}_j\|_2 \ge c\lambda_n, \qquad \min_{j\in M\setminus S}\left\{w_j - \|\{\Gamma_M(\bar{\theta}-\theta^*)\}_j\|_2\right\} \ge c\lambda_n$$

for fixed $\kappa_\Gamma, c > 0$. *If the population marginal scores select* M *with a fixed strict margin, then* $\Pr(\widehat{S} = S) = 1 - o(n^{-A})$ *for every fixed* $A > 0$.

Theorem 1 concerns exact minimizers. Its proof is in Appendix Section S14. This result does not require the initial norms to rank every active group above every inactive group. Appendix Proposition S8 specifies the optimization accuracy needed to reproduce this support. The following theorem proves the probability limits for the correlated three-covariate model of Section 2.2. We set $a = 3.7$. We write $\widehat{S}^{\mathrm{SCAD}}$ for SCAD selection and $\widehat{S}^{\mathrm{adap}}$ for selection with adaptive weights $\lambda_n^2/\|\widehat{\beta}_j^L\|_{I_2}$.

Theorem 2. *Suppose that* $\lambda_n = O(1)$ *and* $mr_n = o(\lambda_n)$. *There exists a sequence of positive-definite three-group Gaussian-copula additive models with active set* $S = \{1,2\}$ *and active components proportional to* λ_n *such that, if* $h_L \to 0$, $G \asymp h_L^{-1}$, *the covariate vectors are independent over time and the errors satisfy Condition 1(iv), then*

$$\Pr(\widehat{S}^{\mathrm{SCAD}} = S) \to 1, \qquad \Pr\{\exists t \ge 0 : \{j : \|\widehat{\beta}_j^L\|_{I_2} > t\} = S\} \to 0, \qquad \Pr(\widehat{S}^{\mathrm{adap}} = S) \to 0.$$

Appendix Proposition S5, Corollary S4 and Section S14.5 construct probit and smooth versions of the model and prove the limits. Appendix Corollary S5 gives the same separation for a sufficiently large fixed penalty multiplier. Appendix Corollaries S1 and S2 give the selection conditions used below. These conditions require independent coordinates, a bandwidth of order $n^{-1/5}$, and a penalty and projection error small relative to the smallest active signal. They require $m^2 G\log^3(dn)/n \to 0$ for growing d. For fixed d, they require a sufficiently large penalty constant with a fixed signal separation.

3.2 Component Estimation

In a planned epoch, N is the final estimation block size. We write s_c for the number of active continuous components and S_d for the active categorical set. We measure component error by $\|g\|_{P_j}^2 = \int g^2\, dP_j$ with $g = \widehat{f}_j - f_j$. For expected risk, we truncate the point estimate at $\pm\log N$ to give $\widehat{f}_j^T$. This bounds loss on rare unstable samples, as specified in Appendix Section S4. The following corollary bounds total expected error, including selection failures. The bound is uniform over distributions P with given s_c and $|S_d|$ satisfying Condition 1 and the selection conditions below. We write $\mathbb{E}_P$ for the expectation under P.

Corollary 1. *Suppose that Condition 1 and the selection conditions of Appendix Corollary S1 or Corollary S2 hold. If the point-estimation bandwidth is of order $N^{-1/5}$ with grid size $G \asymp N^{1/5}$, then*

$$\sup_P \mathbb{E}_P \sum_{j=1}^{d} \|\widehat{f}_j^T - f_j\|_{P_j}^2 \lesssim s_c N^{-4/5} + |S_d| N^{-1}.$$

Corollary 1 gives the known-support rates. The reason is that selection failures have probability $o(N^{-A})$ for every fixed $A > 0$. The proof is in Appendix Section S10. Appendix Proposition S9 permits correlated covariates, including Gaussian copulas with well-conditioned latent correlation matrices. This result keeps the selection error explicit. Proposition S2 permits non-Gaussian or nonstationary errors for risk.

Appendix Theorem S6 gives a matching lower bound on a class with s active smooth components, bounded total signal, component norms above a small threshold, and independent uniform covariates. No estimator on this class attains worst-case component error of smaller order than $sN^{-4/5}$. The planned-epoch estimator is thus rate optimal on the intersection of this class with the class of Corollary 1.

3.3 Simultaneous intervals

Planned-epoch intervals use final block size N, inference bandwidth h and consecutive blocks of length L_N. The following theorem allows polynomially many reported values. Reported values of every candidate group are fixed in advance. We define a nonempty set $\mathcal{T}_N$ of the reported pairs (j, x) of the active groups, with at most N^{A_0} pairs for a fixed $A_0 > 0$. Each x is a categorical level or a continuous value in a fixed interior interval. We form $\widehat{\Sigma}$ and c^* as in Section 2.3, with the maximum over the reported values of the selected groups, and an unselected group receives an empty interval.

Theorem 3. *Suppose that Condition 1 and the selection conditions of Appendix Corollary S1 or Corollary S2 hold, s is fixed, $h \asymp N^{-\kappa}$ and $L_N \asymp N^{\gamma}$ with $1/5 < \kappa < 1/2$ and $\kappa < \gamma < 1$. Then*

$$\Pr\left\{ f_j(x) \in [\widehat{f}_j(x) - c^* \widehat{\Sigma}^{1/2}_{(j,x),(j,x)}, \widehat{f}_j(x) + c^* \widehat{\Sigma}^{1/2}_{(j,x),(j,x)}] \text{ for every } (j,x) \in \mathcal{T}_N \right\} \longrightarrow 1 - \alpha.$$

Theorem 3, proved in Appendix Section S12, uses an inference bandwidth of smaller order than the point-estimation bandwidth $N^{-1/5}$. This undersmoothing makes the bias negligible relative to the standard errors. Each block has many more than $1/h$ observations but a fraction of the sample tending to zero, which keeps the deleted-block covariance accurate. The probability includes selection errors. Appendix Theorem S3 permits shorter blocks for a fixed evaluation set. Appendix Theorem S5 permits correlated covariates with finite smooth expansions of their pairwise densities and identifiable additive effects. Its coverage error includes the selection-error probability.

Appendix Corollary S3 extends the intervals on a fine grid to a band over a covariate interval under the specified second-derivative bound. Appendix Proposition S7 gives the variances of categorical levels and of level differences. The estimated center cancels from these differences.

3.4 Continuing estimation and inference

At time n, we define $N_{E,n}$ and $N_{I,n}$ as the counts of observations in the point-estimation and inference states of Section 2.3. For point estimates, the reporting bandwidth is $h_n = c_E n^{-1/5}$ with $c_E > 0$. For expected risk, $\widehat{f}_j^{\mathrm{C,T}}$ truncates the estimate at $\pm \log N_{E,n}$ with the unstable-sample rule in Appendix Section S9. For correlated covariates, an additive lower bound prevents component effects from canceling each other. Selection requires weak correlation and fixed signals, or strict population margins with a checked optimization tolerance. Appendix Condition S2 specifies these premises. The following theorem retains almost all observations and the known-support rate.

Theorem 4. *Suppose that s is fixed, Condition 1 holds, and the selection conditions of Appendix Corollary S1 or Corollary S2 hold uniformly at retained sample sizes $(n/\log n)^{3/4} \le k \le n^{3/4}$. Alternatively, suppose Appendix Condition S2 holds and use its specified procedure. Then $N_{E,n}/n \longrightarrow_p 1$, $N_{I,n}/n \longrightarrow_p 1$ and, for each active continuous component,*

$$\mathbb{E}\|\widehat{f}_j^{\mathrm{C,T}} - f_j\|_{P_j}^2 \le C\{h_n^4 + (nh_n)^{-1}\} \lesssim n^{-4/5}.$$

Table 1: Selection and component estimation over 100 replications per design.

	Exact support			Component error		
Design	SCAD	Best cutoff	Adaptive	SCAD	Lasso-refit	Adaptive
H1	.99	.01	.02	.0182	.0210	.0209
H2	1.00	.00	.00	.0137	.0155	.0155
H3	1.00	.00	.00	.0074	.0083	.0083

		Exact support		Component error		
Design	Used fraction	Continue	Planned	Continue	Planned	Coverage
A1	.833	1.000	1.000	.0094	.0156	.970
A2	.865	1.000	1.000	.0073	.0121	.940
A3	.894	1.000	1.000	.0055	.0089	.920

Appendix Sections S15 and S30 prove the independent and correlated alternatives. Selection stabilizes before the current grid accumulates moments. Earlier bandwidths preserve squared bias $O(h_n^4)$ and variance $O((nh_n)^{-1})$. The correlated risk proof needs no pairwise density. For inference, we assume bounded pairwise density ratios and a fixed active distribution.

For intervals, $\widehat{f}_j^{\mathrm{C}}(x)$ is the inference-state estimate with reporting bandwidth $c_I n^{-1/4}$, $c_I > 0$. We write $\widehat{\mathrm{se}}_j^{\mathrm{C}}(x)$ and $c_{1-\alpha}^{\mathrm{C}}$ for its standard error and critical value from Section 2.3 for $0 < \alpha < 1$. The following theorem establishes joint coverage at a fixed finite set of pairs (j, x) admitted in $\mathcal{T}_N$, with the critical value computed over the selected groups. The reporting time n is fixed in advance. An unavailable inference calculation makes the event false.

Theorem 5. *Suppose the conditions of Theorem 4 hold and the continuous marginal densities have two bounded derivatives. Then*

$$\Pr\left\{\max_{j,x}\frac{|\widehat{f}_j^{\mathrm{C}}(x) - f_j(x)|}{\widehat{\mathrm{se}}_j^{\mathrm{C}}(x)} \le c_{1-\alpha}^{\mathrm{C}}\right\} = 1 - \alpha + o(1).$$

Theorem 5 includes selection errors and dependence between selection and estimation. Its correlated proof is in Appendix Section S30. Theorem 3 applies to a planned epoch.

Under the conditions of Theorem 4, Appendix Proposition S6 bounds storage with probability tending to one. Selection observations use $O(dn^{3/4})$ numbers. The states use $O(s^2 n^{9/10})$ numbers, up to iterated-logarithm factors. This is an asymptotic bound for fixed d and s; it does not imply memory savings at the sample sizes studied here.

4 Empirical Evidence

We study selection, continuing estimation, and joint coverage through simulations with at least 100 independent replications per design. Exact support is the fraction of replications with $\widehat{S} = S$. Component error equals $\sum_j \|\widehat{f}_j - f_j\|_{P_j}^2$ in one replication, and tables report its mean. Joint coverage is the fraction of replications covering all designated component values in the same run. An omitted active component counts as noncoverage. F-SGD and Sieve-SGD denote functional and sieve stochastic gradient descent (SGD).

In H1–H3, SCAD has the best support recovery and lowest mean component error in Table 1. These designs use three correlated covariates and two bounded smooth effects as in Theorem 2. The selectors share observations, candidate set, optimization and support tolerances, and final unpenalized fit. Best cutoff is the frequency with which some common cutoff on the initial norms recovers S. Except in one H1 replication, every cutoff that keeps both active groups also keeps the inactive group, and its refit equals Lasso-refit. Adaptive uses weights $\lambda^2/\|\widehat{\beta}_j^L\|_{I_2}$. Appendix Table S8 gives the full comparison. Appendix Tables S7 and S14 report the probit-score G1–G3 designs.

In A1–A3, we compare continuing, planned-epoch and latest-block fits with five active effects among $d = 20$, 50, and 50 covariates. A2 and A3 retain the preceding covariate vector with probability 0.3.

A3 has an error autoregression coefficient 0.3. Continue and Planned use the same stream and penalty multiplier 0.8, with different selection sample sizes. Both recover the exact support in every run. Used fraction is the proportion contributing to the continuing point state. Table 1 gives continuing errors 39.8, 39.5, and 38.1 percent below planned-epoch errors. With the same support and point resolution, latest-block errors are 0.1715, 0.4020, and 0.1515 in Appendix Table S15.

In the third study, we compare our continuing estimator with F-SGD (Chen and Klusowski, forthcoming) and Sieve-SGD (Zhang and Simon, 2022) in the independent A1 design. Both competitors use order-one cosine Sobolev tuning, which all five effects satisfy. Twenty independent pilot streams choose interior tuning values from expanded grids. Continuing error is 75.2 and 77.5 percent lower than F-SGD and Sieve-SGD. It is 60.8 percent below online smooth backfitting (Yang, Yao and Zhao, 2024) with oracle pilot tuning. On the active components alone, the three reductions are 51.5, 53.6 and 37.0 percent (Appendix Tables S17 and S18). Appendix Table S16 reports the comparison under order-two cosine Sobolev tuning, a class that excludes the first effect.

In the fourth study, we examine an inactive binary covariate associated with two active effects in R1–R3. These designs follow Appendix Theorem S4. Both methods use the same unpenalized final fit. Appendix Tables S5 and S13 give exact recovery 0.92–0.99 for SCAD against 0.21–0.28 for Lasso-refit. The binary selection frequency is zero for SCAD and 0.65–0.77 for Lasso-refit. SCAD reduces component error by 7.7–9.5 percent and gives joint coverage 0.93–0.96.

In D1–D8, joint coverage is 0.935–0.980, near the nominal level, and SCAD recovers the exact support in 0.810–0.975 of replications against 0.600–0.935 for Lasso-refit. These designs have five active effects among $d = 200$ or 500 covariates. D1–D6 use Gaussian copulas, D7 has a nonuniform marginal, and D8 has a categorical effect. D7 and D8 follow the proved dependence model. Appendix Sections S22 and S25 specify the designs. In a paired centering comparison with strong linear or categorical effects, Appendix Table S11 shows undercoverage with fixed centers: coverage is 0.930–0.975 with recomputed centers against 0.505–0.790 with fixed centers. Appendix Table S12 reports grid coverage 0.920–0.950 and narrowing interpolation bands under a known curvature bound.

In B1–B3, reduced moments lower storage, mean time, and peak memory among three equivalent planned-epoch implementations. On 300 datasets with up to $n = 6{,}000$ and $d = 1{,}000$, reduced moments use 11.3–15.3 percent of dense moment storage. The supports agree, with estimate and interval differences below 3×10^{-14} (Appendix Table S6). With 200–1,000 candidate groups, the reduced summaries need 14–69 percent of the memory of the raw observations. The continuing storage bound is asymptotic.

Finally, we analyze bike-rental and appliance energy data with planned-epoch fits. The bike-rental analysis uses 7,965 hourly observations and 56 candidate groups, including 50 synthetic nulls. The appliance analysis uses 102 candidate groups and 15,672 observations. All 100 null replications in the bike data select the four effects of Section 2 and no synthetic null. In the appliance data, SCAD selects 14 groups. Eleven appear in at least 95 of 100 consecutive-block deletion fits. Appendix Sections S23 and S24 give the curves, hour contrasts, and simultaneous uncertainty.

5 Limitations and Discussion

We update nonlinear effects and simultaneous intervals from stored sums while discarding historical observations. Continuing inference permits the correlated designs in Appendix Condition S2. The proofs use the regeneration times of the redraw transition, and the D1–D6 designs and calendar applications lie outside these proved classes. Coverage in D1–D6 is empirical, and the application intervals are nominal. The fitted curves have no causal interpretation. Inference requires Gaussian autoregressive errors, a fixed active set, and prespecified reporting times, whereas the trend bands of Nagler et al. (2026) are uniform in time. Storage optimality remains unproved (Quan and Lin, 2024). The compared SGD estimators use less storage but provide neither selection nor intervals. Dynamic active sets and data-dependent reporting times require further theory.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and manuscript editing. We verified the cited publications and reviewed all AI-assisted material. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The empirical analyses use public datasets that contain no personal identifiers. We report conditional associations and do not give the fitted components a causal interpretation.

REPRODUCIBILITY STATEMENT

The appendices give the complete estimator, proofs, simulation distributions, seed rules, data processing and additional results. Every reported production simulation design uses at least 100 independent replications. The numerical scripts and stored outputs determine all numerical tables and empirical figures, and the two public datasets are identified by persistent digital object identifiers. An anonymous supplementary archive contains the registered scripts, production outputs, deterministic tests and public-data retrieval instructions. The same archive includes the certified continuing selector and its deterministic checks.

REFERENCES

Chen, X. and Klusowski, J. M. (forthcoming). Stochastic gradient descent for nonparametric additive regression. To appear in *Bernoulli*. arXiv:2401.00691.

Chen, X., Tamer, E. and Yao, Q. (2026). Fast online inference on semiparametric models. Cowles Foundation Discussion Paper 2548, Yale University. arXiv:2603.08614.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics* **44** 1363–1399.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106** 544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42** 819–849.

Gregory, K., Mammen, E. and Wahl, M. (2021). Statistical inference in sparse high-dimensional additive models. *Annals of Statistics* **49** 1514–1536.

Guo, Z., Yuan, W. and Zhang, C.-H. (2026). Decorrelated local linear estimator: Inference for non-linear effects in high-dimensional additive models. *Journal of Machine Learning Research* **27**(27) 1–79.

Han, D., Xie, J., Liu, J., Sun, L., Huang, J., Jiang, B. and Kong, L. (2024). Inference on high-dimensional single-index models with streaming data. *Journal of Machine Learning Research* **25**(337) 1–68.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38** 2282–2313.

Kale, S., Karnin, Z., Liang, T. and Pál, D. (2017). Adaptive feature selection: Computationally efficient online sparse linear regression under RIP. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR **70** 1780–1788.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17** 1217–1241.

Lee, E. R., Park, S., Mammen, E. and Park, B. U. (2024). Efficient functional Lasso kernel smoothing for high-dimensional additive regression. *Annals of Statistics* **52** 1741–1773.

Liu, G., Chen, H. and Huang, H. (2020). Sparse shrunk additive models. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR **119** 6194–6204.

Liu, M., Shang, Z. and Yang, Y. (2026). Scalable inference for nonparametric stochastic approximation in reproducing kernel Hilbert spaces. *Annals of Statistics* **54** 771–795.

Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics* **45** 2455–2482.

Lu, J., Kolar, M. and Liu, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association* **115** 2084–2099.

Mammen, E., Linton, O. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27** 1443–1490.

Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37** 3779–3821.

Nagler, T., Brock, T. and Palm, N. (2026). Online bootstrap inference for the trend of nonstationary time series. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, PMLR **337** 4684–4711.

Palm, N. and Nagler, T. (2024). An online bootstrap for time series. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, PMLR **238** 190–198.

Quan, M. and Lin, Z. (2024). Optimal one-pass nonparametric estimation under memory constraint. *Journal of the American Statistical Association* **119** 285–296.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **13**(13) 389–427.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B* **71** 1009–1030.

Tarrès, P. and Yao, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory* **60** 5716–5735.

Wang, Y., Chen, H., Liu, W., He, F., Gong, T., Fu, Y. and Tao, D. (2023). Tilted sparse additive models. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR **202** 35579–35604.

Wang, W., Zhang, Q. and Zhang, X. (2026). SPARKLE: A nonparametric approach for online decision-making with high-dimensional covariates. *arXiv preprint arXiv:2503.16941v3*.

Xie, Y., Shi, R., Schaeffer, H. and Ward, R. (2022). SHRIMP: Sparser random feature models via iterative magnitude pruning. In *Proceedings of Mathematical and Scientific Machine Learning*, PMLR **190** 303–318.

Yang, S., Yan, Y., Zhu, X. and Sun, Q. (2023). Online linearized LASSO. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, PMLR **206** 7594–7610.

Yang, Y., Yao, F. and Zhao, P. (2024). Online smooth backfitting for generalized additive models. *Journal of the American Statistical Association* **119** 1215–1228.

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

Zhang, T. and Simon, N. (2022). A sieve stochastic gradient descent estimator for online nonparametric regression in Sobolev ellipsoids. *Annals of Statistics* **50** 2848–2871.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36** 1509–1533.

APPENDIX

Every component of the additive model is centered to have mean zero under the marginal distribution of its own covariate. This population centering determines both the component target and its sampling variation, and additive moment equations preserve this target when observations arrive or a time block is removed. The results below connect these equations to selection, estimation error and simultaneous confidence intervals.

For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite. We write I_q for the q-dimensional identity matrix and I when its dimension is clear. For a distribution P, $L^2(P)$ is the space of functions with $\|g\|_P^2 = \int g^2\, dP < \infty$, and $L_0^2(P)$ is its mean-zero subspace. We also write $\|g\|_{L^2(P)}$ for this norm. For functions, $\|g\|_2$ also denotes this norm under the specified distribution. Inner products and adjoints refer to the indicated Euclidean or L^2 space. C^r denotes functions with r continuous derivatives, and C^∞ denotes infinitely differentiable functions.

We write $\mathbb{1}(A)$ for the indicator of an event A and $\mathrm{tr}(A)$ for the trace of a square matrix. Unless specified otherwise, positive constants c, C may change between bounds and depend only on the fixed parameters in the stated conditions. A subscript A in C_A or O_A permits dependence on the fixed probability exponent A. The notation O_{op} measures the order in operator norm.

PROCEDURE SUMMARY

The following two lists give the operations from observed data to the reported estimates and intervals. The conditions of the main results are listed after them.

Planned epoch.

1. Screening block: accumulate the marginal moments of all d groups, compute the one-covariate loss reductions Δ_j, and keep the m groups with the largest reductions.
2. Selection block: accumulate the cross moments of the kept groups, fit the group Lasso with penalty $\widehat{\lambda} = c_\lambda \widehat{\sigma}_Y \{\log d/(N_L h_L)\}^{1/2}$, compute the SCAD weights $\omega_j = p'_{\widehat{\lambda},a}(\|\widehat{\beta}_j^L\|)$, refit the weighted criterion once, and let $\widehat{S}$ be its nonzero groups.
3. Final block: on $\widehat{S}$, accumulate moments at the point bandwidth and, separately for J consecutive blocks, at the smaller inference bandwidth. Unused intervals of order $\log n$ separate the three blocks.
4. Report: the unpenalized fit from the point moments is the estimate. For the intervals, delete each inference block from the moments, refit with recomputed centers, and form V_b, $\widehat{\Sigma}$ and the resampling critical value c^* of Section S6.

Continuing estimation.

1. Keep the observations at the indices $\lfloor k^{4/3} \rfloor$, about $n^{3/4}$ of the first n, for selection only.
2. Divide the other observations into blocks ending at $t_r = \lfloor r^{5/2} \rfloor$ and keep the open block until its boundary.
3. At a boundary, add the completed block's moments, computed at its own point and inference bandwidths $c_E t_r^{-1/5}$ and $c_I t_r^{-1/4}$ with unscaled local lines, to every active grid state. Keep its inference moments as a deletion block and discard its observations. Update the selected set on the retained observations when the update rule applies, and restart the states when the selected set changes.
4. Each nested grid starts to accumulate moments before it becomes the reporting grid, as specified in Section S9.
5. At a reporting time, add temporary moments of the open block, compute the estimate from the point state, and compute the intervals from the inference state by exact block deletion and resampling.

Conditions of the main results. Theorem 1 uses Condition S1, Gaussian autoregressive errors and population margins for the active norms and inactive scores. Condition S1 permits correlated

covariates through the stationary distribution and a transition with independent redraws. Theorem 2 concerns a constructed three-group Gaussian-copula model whose active components are proportional to the penalty level, with covariate vectors independent over time. Corollary 1 and Theorem 3 use Condition 1, with independent coordinates and the redraw transition, and the selection conditions of Corollary S1 or Corollary S2. Theorem 3 adds a fixed active set, undersmoothing and the block-length condition. Theorems 4 and 5 use either Condition 1 with selection conditions at the retained sizes or Condition S2, which permits correlated active covariates with bounded pairwise density ratios. Theorem 5 concerns a fixed finite set of values at a prespecified reporting time.

S1 Local-polynomial component estimation

The matrix formulation below covers local-polynomial continuous effects and categorical levels in the criterion of Section 2.

S1.1 Continuous and categorical groups

We write $Y_i = \mu + \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i$ for $i = 1, \ldots, N$. Here N is the number of observations in the current block, and $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$ is its response mean. We center f_j with respect to the marginal distribution of X_{ij}. For a continuous predictor, we fix a grid $u_{j1}, \ldots, u_{jG_j}$ on $[0, 1]$ with positive quadrature weights q_{jg}. We define $K(t) = 3(1 - t^2)_+/4$ as the Epanechnikov kernel and fix the bandwidth $h_j > 0$ before a data block begins. The weights assigned to neighboring grid nodes define a probability distribution by

$$\pi_{jg}(x) = \frac{q_{jg} K\{(x - u_{jg})/h_j\}}{\sum_{a=1}^{G_j} q_{ja} K\{(x - u_{ja})/h_j\}}.$$

The theoretical selection and planned-epoch grids are equally spaced with $4 \le (G_j - 1)h_j \le 8$. This spacing makes the denominator positive on $[0, 1]$. Their trapezoidal weights equal the spacing at interior nodes and half the spacing at the boundaries. Continuing estimation uses the nested grids in Section S9. Numerical quadrature rules are specified with each study. The identity $\sum_g \pi_{jg}(x) = 1$ normalizes over nodes and does not make the estimated marginal density equal to one.

We write r for the local-polynomial degree. We use an auxiliary grid index U_j to express the weighted average over local polynomials as a conditional expectation. Conditional on $X_j = x$, the index U_j equals g with probability $\pi_{jg}(x)$. This notation represents a weighted sum. The estimator evaluates the sum directly and does not sample these auxiliary indices. We define the vector $v_j(U_j, x)$ of length $(r + 1)G_j$ by setting its block at $U_j = g$ to $\{1, t, \ldots, t^r\}^\mathsf{T}$. Here $t = (x - u_{jg})/h_j$. Its remaining entries are zero. We compute

$$a_j(x) = \mathbb{E}\{v_j(U_j, x) \mid x\}, \qquad D_j(x) = \mathbb{E}\{v_j(U_j, x) v_j(U_j, x)^\mathsf{T} \mid x\}.$$

Here $a_j(x)$ averages the local basis vectors, and $D_j(x)$ averages their outer products. The first quantity determines the reported curve, and the second determines the within-group loss and penalty norm. For a categorical variable with K_j levels, we set $v_j(U_j, x) = a_j(x) = e_x$ and $D_j(x) = e_x e_x^\mathsf{T}$. The vector e_x is the corresponding coordinate vector in $\mathbb{R}^{K_j}$. No density on an interval is used for this group. The auxiliary grid indices are conditionally independent across groups given the covariates.

S1.2 One quadratic criterion

We write $a_{ij} = a_j(X_{ij})$, $D_{ij} = D_j(X_{ij})$, $\widehat{D}_j = N^{-1} \sum_i D_{ij}$, and $\widehat{c}_j = N^{-1} \sum_i a_{ij}$. The vector β_j contains the local-polynomial coefficients or the categorical coefficients of group j. For nonnegative weights ω_j, we minimize the criterion over the intercept α and coefficient blocks β_j:

$$\frac{1}{2N} \sum_{i=1}^{N} \mathbb{E}\Big[\Big\{Y_i - \alpha - \sum_{j=1}^{d} v_j(U_j, X_{ij})^\mathsf{T} \beta_j\Big\}^2 \mid X_i, Y_i\Big] + \sum_{j=1}^{d} \omega_j (\beta_j^\mathsf{T} \widehat{D}_j \beta_j)^{1/2}, \qquad \widehat{c}_j^\mathsf{T} \beta_j = 0. \quad \text{(S1)}$$

Both the loss and group norm use the same smoothed empirical measure. The centering constraint gives $\widehat{\alpha} = \bar{Y}$.

In the quadratic form of the criterion, the unnormalized diagonal block is $\widehat{D}_j$. For $j \neq k$, the cross block is $N^{-1}\sum_i a_{ij}a_{ik}^{\mathsf{T}}$. The auxiliary indices are conditionally independent across groups. This independence gives products of conditional means in the cross blocks. The constraint $\widehat{c}_j^{\mathsf{T}}\beta_j = 0$ permits only coefficients whose reported curve has empirical mean zero. After removing zero-norm directions, we choose vectors spanning this coefficient space and place them in the columns of a matrix R_j. The substitution $\beta_j = R_j\theta_j$ then satisfies the centering constraint for every vector θ_j. We choose these columns to satisfy $R_j^{\mathsf{T}}\widehat{D}_j R_j = I$. The penalty norm in (S1) becomes $\|\theta_j\|_2$. Section S20 constructs R_j from the covariate moments. An evaluation outside the estimable space requires an explicit extrapolation convention and does not identify a population component there. We collect the group coordinates in θ and define

$$H_{jj} = I, \qquad H_{jk} = R_j^{\mathsf{T}}\left\{N^{-1}\sum_i a_{ij}a_{ik}^{\mathsf{T}}\right\}R_k, \qquad b_j = R_j^{\mathsf{T}}N^{-1}\sum_i a_{ij}(Y_i - \bar{Y}). \tag{S2}$$

We write $H = (H_{jk})_{j,k}$ for the block matrix and b for the vector stacking the b_j. With $n = N$, the criterion (S1) equals, up to a constant,

$$\mathcal{L}_n(\theta) = \frac{1}{2}\theta^{\mathsf{T}}H\theta - b^{\mathsf{T}}\theta + \sum_{j=1}^{d}\omega_j\|\theta_j\|_2. \tag{S3}$$

Equation (S3) is the normalized-coordinate form of (1). The squared-residual criterion makes H symmetric and positive semidefinite without an additional density condition. The identity diagonal blocks of H in (S2) give a group soft-thresholding update for each component.

We write $\widehat{\theta}_j$ for the block of the minimizer for group j. The reported component is

$$\widehat{f}_j(x) = a_j(x)^{\mathsf{T}}R_j\widehat{\theta}_j, \qquad N^{-1}\sum_i \widehat{f}_j(X_{ij}) = 0.$$

For an identifiable noiseless affine response, the local-linear criterion has zero residual error and reproduces the response exactly. Empirical centering still differs from population centering. The calculation in Section S18 bounds this difference.

S2 Streaming selection and final estimation

Screening, selection and final estimation use consecutive blocks separated by unused intervals. Selection precedes the final responses. The gaps reduce temporal dependence but do not imply exact independence.

S2.1 Additive summaries

Screening reduces the number of covariate pairs whose moments must be stored. The raw moments in Section S20 give the marginal score $\|b_j\|_2^2$, twice the decrease in the unpenalized loss from fitting covariate j alone. We retain the largest scores and accumulate cross moments only for these candidates. For fixed degree and grid size G, screening uses $O(dG)$ numbers and selection uses $O(m^2G^2)$ on m candidates. The grid and candidate count may depend on the planned epoch size. Section S6 specifies the additional deleted-block summaries. No completed-phase observations are needed by a later phase.

S2.2 Group selection

We use an initial group-Lasso fit to set group penalties and then re-estimate all retained groups jointly. On the selection block, we compute a group-Lasso estimate $\widehat{\theta}^L$ with penalty $\lambda > 0$. For $a > 2$, the SCAD derivative is

$$p'_{\lambda,a}(t) = \begin{cases} \lambda, & 0 \le t \le \lambda, \\ (a\lambda - t)/(a-1), & \lambda < t < a\lambda, \\ 0, & t \ge a\lambda. \end{cases}$$

We use $\omega_j = p'_{\lambda,a}(\left\|\widehat{\theta}_j^L\right\|_2)$ in one additional convex minimization of (S1) on the same selection block. The nonzero groups of its solution define $\widehat{S}$. After the second unused interval, we fit (S1) on $\widehat{S}$ with all penalty weights equal to zero. Thus, SCAD determines the selected support, and the final unpenalized fit estimates its component functions. We recompute H and b from the final block at its estimation bandwidth. In normalized coordinates, the final solution satisfies

$$\widehat{\theta} \in \underset{\theta:\,\theta_j=0\ (j\notin\widehat{S})}{\arg\min} \left\{\frac{1}{2}\theta^\top H\theta - b^\top\theta\right\}.$$

We transform back by $\widehat{\beta}_j = R_j\widehat{\theta}_j$ to obtain the component estimates defined in Section S1.

Selection, point estimation and inference use bandwidths h_L, h_E and h_I, with separate grids fixed before their blocks begin. Final-block observations update both estimation states on the selected support. The smaller h_I reduces inference bias; the expected risk uses h_E. Both estimates target the same population components.

The numerical procedure removes raw moment directions whose eigenvalues are at most N_B^{-3} times the larger of one and the largest eigenvalue in a block of size N_B. A final normalized matrix with smallest eigenvalue below N_B^{-1} is rejected, in which case the point estimate is zero and the uncertainty calculation is reported as a failure. The population results prove that these numerical decisions are inactive with probability tending to one; they are not assumed to be inactive.

The following proposition shows when one SCAD step equals the restricted unpenalized solution. It is a deterministic result about the quadratic criterion. Population probability bounds for its matrix and score conditions are established in Section S10. For a block vector v, we write $\|v\|_{2,\infty} = \max_j \|v_j\|_2$. We write $\|H_{jk}\|_{\mathrm{op}}$ for the spectral norm. We define θ^* as a fixed coefficient vector with nonempty support $S = \{j : \theta_j^* \neq 0\}$ and write $e = b - H\theta^*$. Here S^c is the complement of S among the candidate groups, and $\mathrm{supp}(\theta)$ denotes the set of its nonzero groups.

Proposition S1. *Suppose H is symmetric, $H_{jj} = I$, and*

$$\max_j \sum_{k\neq j} \|H_{jk}\|_{\mathrm{op}} \leq \tfrac{1}{4}, \qquad \|e\|_{2,\infty} \leq \lambda/4.$$

Suppose $\min_{j\in S} \left\|\theta_j^*\right\|_2 > (a+5/3)\lambda$. Then the group-Lasso solution has support S, and its SCAD weights equal zero on S and λ on S^c. The one-step SCAD solution equals $H_{SS}^{-1}b_S$ on S and zero on S^c.

Proposition S1 removes active-group shrinkage while keeping inactive groups zero under its matrix and score bounds. The general one-step oracle principle is established in Fan, Xue and Zou (2014).

S3 Identification of the hour effect

Hour of day is often represented by sine and cosine coordinates. Allowing a separate unrestricted effect of each coordinate can make those effects unidentifiable, because both coordinates describe the same hour. We write Θ for the corresponding angle. If $U = (1+\sin\Theta)/2$ and $V = (1+\cos\Theta)/2$, the centered functions

$$g(U) = (2U-1)^2 - \mathbb{E}\sin^2\Theta, \qquad q(V) = (2V-1)^2 - \mathbb{E}\cos^2\Theta$$

satisfy $g(U) + q(V) = 0$ even when both functions are nonzero. Thus separate unrestricted effects of U and V are not identifiable. This argument does not apply when both effects are restricted to be linear.

We represent hour by one categorical variable $C \in \{0,\ldots,23\}$. The centering constraint is $\sum_{c=0}^{23}\Pr(C=c)f_C(c) = 0$, giving 23 free coordinates when every level has positive probability. The empirical version replaces these probabilities by the observed frequencies. The procedure gives values and uncertainty assessments at the 24 hours only.

This representation removes the preceding null direction, but it does not by itself prove identification relative to all other predictors. For example, we write P_j for the marginal distribution of each

group and suppose the joint distribution has a density at least $c > 0$ with respect to $\bigotimes_j P_j$. For population-centered square-integrable functions g_j,

$$\mathbb{E}\left\{\sum_j g_j(X_j)\right\}^2 \geq c\int\left\{\sum_j g_j(x_j)\right\}^2 \prod_j dP_j(x_j) = c\sum_j \|g_j\|^2_{L^2(P_j)}.$$

The expected squared value of an additive sum is then at least a constant multiple of the sum of the squared component norms, and the additive sum equals zero only when every component equals zero. This population condition establishes identification for mixed categorical and continuous groups and holds with $c = 1$ for independent groups. Positive definiteness of one fitted finite-grid matrix does not establish this population condition for the bike-rental data.

S4 Population component risk

We bound population integrated component error, including empirical centering, incorrect selection and unstable fits. Even exact prediction does not eliminate component error, as Section S18 shows.

For the expected-risk calculation, we write N for the final block size, define $T_N = \log N$, and truncate the point estimate by $\widehat{f}_j^T(x) = \max\{-T_N, \min\{T_N, \widehat{f}_j(x)\}\}$. The reported estimate of an unselected component is zero. If the selected normal equation cannot be solved under the stated numerical tolerance, all components of the reported estimate are set to zero. This rule is specified before observing the final responses. The truncation is used for point estimation and not as a modification of the derivative in Proposition S3. The inferential calculation remains centered at the untruncated estimate. The truncation makes an expected-risk statement possible without an unproved integrability assertion for the inverse of a random matrix. Truncated least squares has been studied by Cohen, Davenport and Leviatan (2013); the component-centering term below is specific to the present additive criterion.

We use Condition 1. We write P_j for the marginal distribution of covariate j and p_j for its density when continuous. We define $S = \{j : f_j \neq 0\}$ as the active set, and write $s = |S|$ for its size. The fixed population bounds are $\sum_j \|f_j\|^2_\infty \leq B^2$, $\|f_j''\|_\infty \leq L$, $c_p \leq p_j \leq C_p$ for continuous groups, and level probabilities at least c_p for categorical groups. The redraw probability is $1-\rho$, with $0 \leq \rho \leq \rho_0 < 1$. The independent stationary Gaussian error process has variance σ^2 and autoregression coefficient ϑ. Here $0 < \sigma_- \leq \sigma \leq \sigma_+$ and $|\vartheta| \leq \vartheta_0 < 1$. Taking $\rho = \vartheta = 0$ gives the independent experiment used for the lower bound.

We write m for the maximum number of selected groups. The quantity s_c is the number of continuous active groups, S_d is the categorical active set, and q is the dimension of the prescribed centered representation on the active set, including its intercept, $q = 1 + s_c(2G-1) + \sum_{j\in S_d}(K_j - 1)$, where K_j is the number of levels of categorical group j. Here the final point estimator uses local-linear polynomials, bandwidth $0 < h \leq 1/2$, and an equally spaced grid with $4 \leq (G-1)h \leq 8$. The upper constant in this grid condition only simplifies the stated dimension bound, and the proof applies to any fixed bounded ratio with a sufficiently fine grid. Put $\delta_{\text{sel}} = \Pr(\widehat{S} \neq S)$, the probability that the selected set differs from the active set, and let $v = \sigma_+^2(1+\vartheta_0)/(1-\vartheta_0)$ be the bound on the eigenvalues of the error covariance matrices that the risk statement uses. The following theorem establishes the expected-risk bound without assuming that δ_{sel} is small or that a sample concentration event holds.

Throughout the risk statements, c, C denote positive constants independent of sample size, dimension and the regression function, but their values may depend on the fixed population bounds.

Theorem S1. *Suppose the preceding population model holds, $s \leq m$, $T_N \geq B$, and $G \leq N$. For every integer $\ell \in [1, N/2]$, for sufficiently large N,*

$$\mathbb{E}\sum_{j=1}^d \|\widehat{f}_j^T - f_j\|^2_{P_j} \leq C\left\{s_c L^2 h^4 + \frac{vq + B^2}{N}\right\} + C(m+s)T_N^2\left\{\delta_{\text{sel}} + 2\ell q \exp\left(-\frac{cN}{q\ell}\right) + N\rho^\ell\right\}. \tag{S4}$$

The constants are uniform over the stated class.

Theorem S1 bounds the population component loss used in the minimax comparison. The first two terms are the squared approximation error and the noise contribution. The term B^2/N accounts for estimation of the component centers. The last line of (S4) retains the loss from selection failures and exceptional design samples. Independence between selection and the final fit is unnecessary: on $\widehat{S} = S$, the computed fit equals the known-support fit for the same final observations.

Proposition S9 extends the point-risk calculation under a population L^2 comparison for additive functions. Bounded density ratios and Gaussian copulas with latent correlation eigenvalues bounded above and away from zero satisfy this comparison. The selection-error probability remains explicit.

The following theorem shows that the selection-error probability δ_{sel} in (S4) is negligible under population conditions. Let n denote the complete epoch size. The three retained block sizes are proportional to n, and the two discarded gaps have order $\log n$. Let I_L be the selection block, $N_L = |I_L|$ its size, h_L its bandwidth, and G_L its grid size. The screening block uses the same group representations. Define $b_n = \min_{j \in S} \|f_j\|_{P_j}$ and $\lambda_n^0 = c_\lambda \sigma_Y \{\log d/(N_L h_L)\}^{1/2}$, where $\sigma_Y^2 = \mathrm{Var}(Y_i)$ and $c_\lambda > 0$ is fixed. We write $\widehat{\sigma}_Y$ for the screening-block sample standard deviation. The implemented penalty is $\widehat{\lambda} = c_\lambda \widehat{\sigma}_Y \{\log d/(N_L h_L)\}^{1/2}$. This estimate is part of the procedure, not a primitive condition.

Theorem S2. *Suppose the preceding population model holds,* $1 \le s \le m$, *and*

$$h_L \asymp n^{-1/5}, \qquad G_L \asymp h_L^{-1}, \qquad \log n \lesssim \log d = o(G_L), \qquad \frac{m^2 G_L \log^3(dn)}{n} \longrightarrow 0,$$

$$s h_L^2 \longrightarrow 0, \qquad \frac{h_L^2 + \lambda_n^0}{b_n} \longrightarrow 0.$$

For every fixed $A > 0$, *the screening, group-Lasso and one-step SCAD procedure satisfies*

$$\Pr(\widehat{S} \neq S) = o(n^{-A}).$$

The assertion is uniform over classes satisfying the rate limits uniformly.

Theorem S2 allows increasing sparsity and decreasing minimum signal b_n, provided that smoothing error and penalty are smaller. It uses independent coordinates, not a claim about arbitrary correlation. We derive score and matrix bounds for the local-polynomial criterion used in marginal screening and one-step selection. These steps adapt the screening of Fan, Feng and Song (2011) and the one-step selection of Zou and Li (2008) and Fan, Xue and Zou (2014) to local lines.

Because the matrix rate already controls the approximation score, the following corollary relaxes the separate dimension conditions in Theorem S2 while retaining the same penalty and algorithm.

Corollary S1. *Suppose the population model of Section S4 holds,* $1 \le s \le m$, *and*

$$h_L \asymp n^{-1/5}, \qquad G_L \asymp h_L^{-1}, \qquad \log d \longrightarrow \infty,$$

$$\frac{m^2 G_L \log^3(dn)}{n} \longrightarrow 0, \qquad \frac{h_L^2 + \lambda_n^0}{b_n} \longrightarrow 0.$$

Then $\Pr(\widehat{S} \neq S) = o(n^{-A})$ *for every fixed* $A > 0$.

Corollary S1 permits dimension growth both slower and faster than the preceding theorem allows. For fixed m and fixed positive minimum signal, its conditions allow $d = \lceil \log n \rceil$ and $d = \lceil \exp(n^{1/4}) \rceil$. In the second case the matrix rate has order $n^{-1/20}$. The separate condition $s h_L^2 \to 0$ follows from the matrix rate and $s \le m$. The result does not cover fixed d with an arbitrary fixed penalty constant, because that case does not make the selection score negligible relative to the stated penalty.

The fixed-dimensional case requires a different comparison: the norm of a selection score is of the same order as the penalty. A sufficiently large fixed penalty constant controls that norm without a diverging dimension. The following corollary also replaces a diverging signal-to-penalty ratio by a fixed separation. The constants below depend only on the fixed dimension, the population bounds, the bandwidth and grid constants, and the SCAD parameter.

Corollary S2. *Under the population model of Section S4, let*

$$d \geq 2 \text{ be fixed}, \qquad 1 \leq s \leq m \leq d, \qquad h_L \asymp n^{-1/5}, \qquad G_L \asymp h_L^{-1}.$$

There exists a fixed $C < \infty$ such that

$$c_\lambda \geq C, \qquad b_n \geq C(h_L^2 + \lambda_n^0) \quad \Longrightarrow \quad \Pr(\widehat{S} \neq S) = o(n^{-A}) \quad \text{for every fixed } A > 0.$$

Corollary S2 uses the same procedure and permits active norms of order $n^{-2/5}$. Its sufficiently large fixed penalty constant replaces dimension growth in Corollary S1. The proof does not cover an arbitrarily small constant.

We write $\mathrm{i} = \sqrt{-1}$ for the imaginary unit in the Fourier formulas. The expected-risk bound uses only a uniform upper bound on the eigenvalues of the population error covariance matrices, without requiring stationarity or Gaussianity. The selection results use, in addition, Gaussian concentration and a common error variance. The errors remain independent of the complete covariate process. A stationary error process with spectral density g_ε gives one sufficient condition for this covariance bound. We use the convention

$$\mathrm{Cov}(\varepsilon_0, \varepsilon_k) = \int_{-\pi}^{\pi} e^{\mathrm{i}k\omega} g_\varepsilon(\omega)\, d\omega.$$

For a stationary error process, suppose

$$2\pi \|g_\varepsilon\|_\infty \leq v < \infty.$$

Then Parseval's identity implies the population covariance bound in the following proposition. The design, signal and tuning conditions of Theorem S1 are retained. The constant v is fixed over the population class; it is not estimated from a realized covariance matrix.

Proposition S2. *Suppose the centered errors are independent of the complete design process and $\mathrm{Cov}(\varepsilon_1, \ldots, \varepsilon_N) \preceq vI_N$ for every N. Then (S4) holds with this v. For Gaussian errors with $\mathrm{Var}(\varepsilon_i) = \sigma^2 \geq \sigma_-^2 > 0$ for every i, the conclusions of Theorem S2 and Corollaries S1 and S2 hold under their respective selection conditions.*

Proposition S2 includes independent errors with changing variances bounded by v. Its risk bound requires no positive variance or spectral lower bound. In the Gaussian selection case, Corollary 1, Theorem 4 and Proposition S6 retain their conclusions and population-centered targets. The inference results retain their stated error conditions; a bounded spectrum alone does not justify resampling.

For a fixed maximum number of categorical levels, these selection results give Corollary 1, proved in Section S10. The final-block fraction changes only its constant.

Corollary 1 concerns the point-estimation bandwidth. The smaller bandwidth used for inference has a different variance order and is not claimed to give an optimal point estimator. In the continuous Gaussian subclass, the lower-bound construction with $\sum_j \|f_j\|_\infty^2 \leq B^2$ and decreasing b_n gives the matching $sN^{-4/5}$ lower bound when $\sqrt{s}[b_n + \{\log(esN^{1/5})/N\}^{2/5}]$ is sufficiently small. Theorem S6 proves this lower bound. The upper and lower experiments then have the same signal condition and the same loss. For example, $d = n^2$, $m \asymp n^{1/5}$, $s \leq cn^{1/5}$ and $b_n = n^{-1/5}$ satisfy both the upper- and lower-bound conditions for sufficiently small fixed c. These conclusions do not use a fixed positive minimum signal with an unbounded number of active components.

S5 Centering and Uncertainty Calculations

An observation weight changes both the fit and its empirical center. We derive both contributions for the unpenalized fit on a fixed support.

We write $\widehat{\beta}_j = R_j \widehat{\theta}_j$ for the local-polynomial coefficients, $\widehat{f}_{ij} = a_{ij}^{\mathsf{T}} \widehat{\beta}_j$ for the fitted component value, and $r_i = Y_i - \bar{Y} - \sum_j \widehat{f}_{ij}$ for the response residual. We define the score s_i by stacking the group blocks

$$s_{ij} = R_j^{\mathsf{T}} \left[a_{ij} r_i - \{ D_{ij} \widehat{\beta}_j - a_{ij} \widehat{f}_{ij} \} \right].$$

The term in braces accounts for variation of the local predictions across grid nodes. It is present because the loss averages squared local residuals, and r_i uses the averaged fitted curve. Multiplying r_i by a regression weight alone would omit this part of the derivative.

We restrict evaluation to linear functionals that are identified on the estimable spaces. The empirical distribution is P_N, with $P_N g = N^{-1}\sum_{i=1}^{N} g(X_i, Y_i)$. To measure the effect of observation i, we perturb P_N to $(1-t)P_N + t\delta_{(X_i,Y_i)}$ and differentiate at $t = 0$. Here $\delta_{(X_i,Y_i)}$ is the point mass at that observation. The following proposition gives the resulting derivative.

Proposition S3. *Suppose that the group moment matrices are positive definite on their estimable spaces and that H is nonsingular. Then the derivative of the fitted component is*

$$\dot{f}_{j,i}(x) = a_j(x)^{\mathsf{T}} R_j [H^{-1} s_i]_j - \widehat{f}_{ij}, \qquad N^{-1}\sum_i \dot{f}_{j,i}(x) = 0. \tag{S5}$$

The term $-\widehat{f}_{ij}$ in Proposition S3 is the empirical center contribution. Theorem S3 controls the approximation, covariance and bias errors needed for coverage.

S6 Inference from blockwise moment summaries

Block deletion recomputes both contributions in (S5) from stored moments. We first treat a fixed evaluation set and then extend coverage to increasing grids and intervals.

In this section, we write N for the final block size, h for the inference bandwidth, and the number of fitting-grid nodes satisfies $G \asymp h^{-1}$. These nodes specify the local polynomials; the evaluation points specify where intervals are reported. We define $\mathcal{T}$ as a nonempty deterministic collection of M pairs (j, x) with $j \in S$. Here x is an interior continuous covariate value or a categorical level. We write $\widehat{f}$ for the vector of untruncated final component estimates at these pairs. The selected support is fixed during the following calculations. The deletions do not repeat selection.

We split the ordered observations into $J = \lceil N/L_N \rceil$ consecutive blocks that include every observation and whose sizes n_b differ by at most one. Here L_N is the target block length. We write $\widehat{f}_{-b}$ for the exact fit after deleting block b, with the bandwidth, grid and support unchanged. The vectors V_b below measure the change in the fit after deletion, and $\widetilde{V}_b$ centers them across blocks. We define

$$V_b = (N - n_b)(\widehat{f} - \widehat{f}_{-b}), \qquad \widetilde{V}_b = V_b - \frac{n_b}{N}\sum_{a=1}^{J} V_a, \qquad \widehat{\Sigma} = \frac{J}{(J-1)N^2}\sum_{b=1}^{J} \widetilde{V}_b \widetilde{V}_b^{\mathsf{T}}. \tag{S6}$$

The standard error at a reported value is the square root of the corresponding diagonal entry of $\widehat{\Sigma}$. This matrix is positive semidefinite and equals the usual delete-group jackknife covariance when the block sizes are equal. For dependent observations, the role of a growing deleted block is studied by Künsch (1989) and Lahiri (2003).

Intervals that hold at all M reported values at once need a single critical value, and we obtain that value by resampling the centered deleted-block vectors. For each resampling draw, we sample J vectors independently with replacement from $\widetilde{V}_1, \ldots, \widetilde{V}_J$. We denote their coordinatewise mean and sample standard deviation by $\bar{V}^*$ and s^*, and calculate

$$T^* = \max_{1 \le k \le M} \frac{\sqrt{J}\,|\bar{V}_k^*|}{s_k^*}.$$

For $0 < \alpha < 1$, we define c^* as the conditional $(1-\alpha)$ quantile of T^*. The confidence interval at coordinate k is $[\widehat{f}_k - c^* \widehat{\Sigma}_{kk}^{1/2}, \widehat{f}_k + c^* \widehat{\Sigma}_{kk}^{1/2}]$. We recompute the standard errors in every resampling draw. For probability statements, failed fits, unidentified evaluations, and nonpositive or nonfinite standard errors give empty intervals and count as noncoverage. Before studentization, the implementation divides each coordinate of the block contributions by its maximum absolute value across blocks. It uses a positive machine tolerance only to avoid division by zero in a resample whose sample standard deviation equals zero at some coordinate; such resamples have asymptotically zero probability under the following theorem.

The following theorem gives asymptotically nominal joint coverage on $\mathcal{T}$ with fixed active-set and evaluation-set sizes, and the candidate dimension may grow.

Theorem S3. *Suppose the population model and selection conditions of Corollary S1 or Corollary S2 hold. Suppose s and M are fixed, each categorical group has at least two levels, and*

$$h \asymp N^{-\kappa}, \qquad L_N \asymp N^{\gamma}, \qquad 1/5 < \kappa < 1/2, \qquad \max\{0, 3\kappa - 1\} < \gamma < 1.$$

For the exact conditional resampling quantile,

$$\Pr\left\{ f_j(x) \in [\widehat{f}_j(x) - c^* \widehat{\Sigma}^{1/2}_{(j,x),(j,x)}, \widehat{f}_j(x) + c^* \widehat{\Sigma}^{1/2}_{(j,x),(j,x)}] \text{ for every } (j,x) \in \mathcal{T} \right\} \longrightarrow 1 - \alpha.$$

The same limit holds for simultaneous coverage at every specified value of all selected groups, with the resampling maximum taken over exactly that collection.

Theorem S3 gives an unconditional statement: the probability includes the selection step. It applies to fixed active effects and either a fixed or increasing number of screened covariates. Its population class uses independent redraws of the covariate vector and Gaussian errors. It does not prove a uniform band for arbitrarily growing active sets or for arbitrary mixing covariates. For $\kappa < 1/3$, any fixed $0 < \gamma < 1$ is allowed. For larger κ, the proof uses $\gamma > 3\kappa - 1$ because deleting a block changes the empirical inverse equation as well as the score.

Finite resampling introduces a separate numerical approximation. Suppose B_* independent resampling statistics are generated and their empirical quantile uses linear interpolation between order statistics. We define $r_* = 1 + \lfloor (B_* - 1)(1 - \alpha) \rfloor$. The proof of the theorem shows that the limiting coverage, including the resampling randomness, lies between $r_*/(B_* + 1)$ and $(r_* + 1)/(B_* + 1)$. For $B_* = 999$ and $\alpha = 0.05$, these bounds are 0.949 and 0.950. This approximation is distinct from Monte Carlo uncertainty in the reported simulation coverage.

S6.1 Storage and the choice of block length

We write p for the total number of raw coefficients on the selected set. A direct representation of the moments for a final block uses a $p \times p$ cross-product matrix, two vectors of length p, the nodewise polynomial moment matrices, and three response summaries. The implementation also stores one response origin per moment state for numerical stability. A deleted-block fit is obtained by subtracting the corresponding block summaries from the complete summaries. The centering matrices are recomputed from this difference, not retained from the complete sample. This implies that (S6) is determined by $O(Js^2G^2)$ stored numbers and requires no historical response or covariate vector.

This direct representation stores cross moments that the criterion does not need. We retain each between-group block once and replace short-block cross products by an exact QR factor when the QR factor uses fewer entries. The diagonal local-polynomial moments remain separate. The resulting inference summaries require

$$O\left\{ p^2 + Jp + \sum_{b=1}^{J} \min(p^2, pn_b) \right\} \tag{S7}$$

numbers. Section S20.8 derives the representation and its sequential update. The representation reconstructs the same criterion after every block deletion, including the changed component centers. The reduction in (S7) changes storage, not the estimator or the resampling distribution. The numerical study measures both persistent moment storage and peak process memory.

For fixed s, the choice $h \asymp N^{-1/4}$ and $L_N \asymp N^{3/5}$ gives $O(N^{9/10})$ inference storage. The exponent follows from $JG^2 \asymp N^{1-3/5+1/2} = N^{9/10}$. A block length of order $N^{1/3}$ gives $O(N^{7/6})$ storage for the same dense block summaries. Both choices satisfy the conditions of Theorem S3, but only the former is sublinear in N for fixed s. The implemented rule is $L_N = \lceil 0.4N^{3/5} \rceil$. This choice changes the number of stored summaries and the finite-sample resampling distribution; it is not a multiplier applied to a critical value.

Screening stores $O(dG_L)$ marginal summaries and selection stores $O(m^2G_L^2)$ cross summaries. The point estimate has its own $O(s^2G_E^2)$ summaries. Here $G_E \asymp N^{1/5}$. For the candidate rule $m = O(n^{1/5})$, the selection state has size of order $n^{4/5}$. These expressions distinguish the dependence on the full dimension, the screened dimension, and the active dimension. They do not imply that the entire algorithm uses constant memory when the grid resolution increases with the planned epoch size.

S6.2 Interpolation on an interval

A finite set of confidence intervals extends to the values between its reported values when a numerical curvature bound is specified for the parameter class. Suppose $|f_j''| \le L$ and the simultaneous intervals at consecutive values $x_a < x_{a+1}$ are $[l_a, u_a]$ and $[l_{a+1}, u_{a+1}]$. For $t = (x - x_a)/(x_{a+1} - x_a)$, an interval valid throughout this segment is

$$\left[(1-t)l_a + tl_{a+1} - \frac{L}{2}(x - x_a)(x_{a+1} - x), (1-t)u_a + tu_{a+1} + \frac{L}{2}(x - x_a)(x_{a+1} - x)\right].$$

To verify the extension, we subtract from the component function the line through its true values at the two ends. The resulting twice differentiable function is zero at both ends and has second derivative bounded in absolute value by L. Subtracting or adding $L(x - x_a)(x_{a+1} - x)/2$ makes this function convex or concave, respectively. This gives the two inequalities. The extra half-width is at most $L(x_{a+1} - x_a)^2/8$. This deterministic argument does not estimate an unknown curvature bound from the same data. With a fixed evaluation grid it also does not establish a shrinking-band optimality result. Categorical effects need no interpolation.

S6.3 An increasing evaluation grid

We define $\mathcal{T}_N$ to be a nonempty set of at most N^{A_0} deterministic pairs (j, x) for fixed $A_0 > 0$, with j in a fixed active set. Continuous values lie in a fixed closed subinterval of $(0, 1)$; categorical groups have fixed level counts. The deleted-block estimator and studentized statistic are unchanged.

Theorem 3 gives simultaneous coverage under its population, bandwidth and block-length conditions; its proof is in Section S12. The coverage condition $\gamma > \kappa$ differs from $\gamma > 2\kappa$ in Section S20. The latter gives sublinear storage for a fixed active set.

For a continuous group j, suppose the grid includes the ends of the interval of interest and has largest adjacent spacing $\Delta_N \le h$. If the curvature bound L is specified for the parameter class, we apply the interpolation intervals of the preceding subsection between all adjacent grid values. We denote this continuous band by $\mathcal{C}_j(x)$ and write $\mathcal{C}_j(c)$ for the level interval of a categorical group. The following corollary establishes that these bands cover every continuous component over its whole interval and every categorical level at once, and it bounds their maximum length.

Corollary S3. *Under the conditions of Theorem 3, for these interpolation bands,*

$$\Pr\{f_j(x) \in \mathcal{C}_j(x) \text{ for all continuous } j \in S \text{ and all } x \text{ in their intervals}\} \ge 1 - \alpha + o(1),$$

$$\max_{j \in S} \sup_x \operatorname{length}\{\mathcal{C}_j(x)\} = O_p\left\{\sqrt{\frac{\log N}{Nh}} + L\Delta_N^2\right\}.$$

The same joint guarantee includes the categorical intervals.

Corollary S3 gives simultaneous bands over covariate intervals when a curvature bound is specified. The corollary does not estimate that bound and does not claim adaptive optimality of the band width. Choosing a number of evaluation values of order G preserves the moment-storage order of the procedure. The finite-resampling bounds stated after Theorem S3 apply to the increasing grid as well.

The estimated component center affects each categorical level interval, and it cancels from the difference between two levels of the same group; Proposition S7 gives the corresponding influences and variances.

S7 Correlated predictors

We construct a correlated model in which SCAD removes an inactive variable retained by Lasso. We then extend component risk and inference to a class with dependent covariates.

We write Q for the stationary covariate distribution, and $P(x, \cdot)$ for the distribution of the next covariate vector given that the current one equals x. We consider the transition

$$P(x, \cdot) = (1 - \rho)Q(\cdot) + \rho K(x, \cdot), \qquad QK = Q, \qquad 0 \le \rho \le \rho_0 < 1. \tag{S8}$$

Here K is any specified Markov transition preserving Q, and the process starts in Q. The earlier independent-redraw model uses a transition K that retains the current covariate vector. Equation (S8) also permits the covariates to change between independent draws. The Gaussian error process remains independent of the entire covariate sequence, with the variance and autoregression bounds stated in Section S4.

S7.1 A VARIABLE RETAINED BY LASSO AND REMOVED BY SCAD

We define $u(x) = \sqrt{3}(2x - 1)$. The first two variables are continuous and the third is a two-level variable $Z \in \{-1, 1\}$. Relative to the product of two uniform marginals on $[0, 1]$ and a symmetric Bernoulli marginal, the joint density of these three variables is

$$q(x_1, x_2, z) = \{1 + 0.1u(x_1)u(x_2)\}\{1 + 0.99z \operatorname{sign}(x_1 + x_2 - 1)\}. \tag{S9}$$

The first density factor makes the continuous variables dependent; the second associates the binary variable with their centered sum. Under Q, the other coordinates are independent of these three and of one another, and they have the continuous or categorical marginals already specified. Suppose $f_1 = f_2 = b_0 u$, $f_3 = 0$, and b_0 lies in a fixed positive compact interval. The other active functions satisfy the earlier smoothness and bounded-signal conditions. Their norms are bounded below by a fixed positive constant, and the active set has fixed size s.

The density in (S9) is positive and integrates to one. In particular, $\mathbb{E}\{u(X_1)u(X_2)\} = 0.1$, and Z is associated with both effects although it has no additive regression effect of its own. Section S13 derives the population selection equations. We write $\widehat{S}^L$ for the support of the initial group-Lasso fit on the selection block. The following theorem establishes that the initial group Lasso retains the binary variable and that one SCAD step removes it.

Theorem S4. *Suppose the error process of Section S4, (S8), (S9) and the preceding regression conditions hold, $m \geq s + 1$, and the selection bandwidth, grid and dimension rates of Theorem S2 hold. Then, for every fixed $A > 0$,*

$$\Pr\{\widehat{S}^L = S \cup \{3\},\ \widehat{S} = S\} = 1 - o(n^{-A}).$$

Theorem S4 isolates the selection gain from SCAD: the initial Lasso support is incorrect with probability tending to one. Its proof concerns the implemented local-linear criterion. The one-step principle is due to Zou and Li (2008), and its general oracle theory is established in Fan, Xue and Zou (2014). The strict population inequalities persist in an open neighborhood of the two density constants. The example is not an isolated distribution.

S7.2 COMPONENT INFERENCE BEYOND INDEPENDENT COVARIATES

We next permit a larger class of stationary distributions Q. For each active group j, we define χ_j as a nonempty finite vector of centered orthonormal functions in $L^2(P_j)$. For a continuous group, these functions and their first two derivatives are uniformly bounded; for a categorical group, they are bounded contrasts. We write Q_{jk} for the joint distribution of the covariates of groups j and k under Q, and $P_j \otimes P_k$ for the product of their marginals. The density of Q_{jk} relative to that product satisfies

$$\frac{dQ_{jk}}{d(P_j \otimes P_k)}(x, y) = 1 + \chi_j(x)^{\mathsf{T}} \mathcal{R}_{jk} \chi_k(y), \qquad j \neq k, \quad j, k \in S. \tag{S10}$$

The matrix $\mathcal{R}$ of fixed dimension has identity diagonal blocks and off-diagonal blocks $\mathcal{R}_{jk}$. This matrix is a population parameter, not an estimated or selected Gram matrix. The condition $\lambda_{\min}(\mathcal{R}) > 0$ ensures identification of the additive components. The proof of the inference theorem below derives the corresponding empirical stability bounds. Equation (S10) is a finite orthonormal expansion of each pairwise density. The corresponding general expansions are treated by Dueck, Edelmann and Richards (2017, Section 2). Here the smoothness of the continuous dependence functions controls the component projection bias.

For the density in (S9) on its active set, we take $\chi_1 = \chi_2 = u$ and one bounded smooth function or contrast for each remaining group. The only nonzero off-diagonal entries of $\mathcal{R}$ are 0.1 between the first two groups.

We retain the marginal density, smoothness and bounded-signal conditions but do not require independence between coordinates. For each candidate group j, we fix a deterministic collection $\mathcal{X}_{N,j}$ of reported values and define $\mathcal{T}_N = \bigcup_{j\in S}(\{j\} \times \mathcal{X}_{N,j})$. This nonempty collection satisfies the size and location restrictions of Theorem 3. We calculate the interval $\widehat{I}_j(x)$ by Section S6, with the resampling maximum taken over $\bigcup_{j\in\widehat{S}}(\{j\} \times \mathcal{X}_{N,j})$. For an unselected group, we set $\widehat{I}_j(x) = \varnothing$. The following theorem establishes the component-risk bound and simultaneous coverage under this pairwise dependence.

Theorem S5. *Suppose the population model of Section S4 holds with coordinate independence replaced by (S8) and (S10), s is fixed, $\lambda_{\min}(\mathcal{R}) \geq c > 0$, and the point-estimation grid and truncation of Theorem S1 are used. If $m \geq s$, $T_N \geq B$ and $G_E \leq N$, then (S4) holds for every integer $\ell \in [1, N/2]$ and sufficiently large N with $h = h_E$ and $G = G_E$. If*

$$h_I \asymp N^{-\kappa}, \qquad L_N \asymp N^{\gamma}, \qquad 1/5 < \kappa < 1/2, \qquad \kappa < \gamma < 1,$$

then, for the exact conditional resampling quantile,

$$\left|\Pr\{f_j(x) \in \widehat{I}_j(x) \text{ for every } (j,x) \in \mathcal{T}_N\} - (1-\alpha)\right| \leq \delta_{\text{sel}} + o(1).$$

The same inequality holds when coverage includes selected inactive groups.

Theorem S5 keeps selection error explicit. Theorem S4 makes it negligible for (S9), giving selection, risk and inference for one correlated, temporally dependent model. More generally, the inference conclusion holds under (S10) when $\delta_{\text{sel}} \to 0$.

Smooth pairwise dependence permits the same inference bandwidth as in the independent case, because its contribution to population projection bias is $O(h_I^4)$, and the direct continuous smoothing bias is $O(h_I^2)$. The smaller order follows from the orthogonality of two local-polynomial projection errors. This distinction also controls the bias of a correlated categorical component at its root-N scale. The finite-resampling quantile bounds and the known-curvature interpolation argument remain valid. Neither the finite-rank pairwise condition nor the transition (S8) is asserted to hold for every Gaussian copula or for a deterministic calendar sequence.

S8 Population selection margins

We express selection margins through the population initial and weighted fits, then verify the projection bounds for correlated designs.

We define M as a population candidate set containing S, and write $m = |M|$. Here M is unrelated to the number of evaluation pairs in Section S6. In centered and within-group standardized coordinates, we write its population criterion as

$$\mathcal{Q}_M(\theta) = \frac{1}{2}(\theta - \theta^*)^{\mathsf{T}}\Gamma_M(\theta - \theta^*), \qquad \theta^L = \arg\min_{\theta}\left\{\mathcal{Q}_M(\theta) + \lambda \sum_{j\in M}\|\theta_j\|_2\right\}. \tag{S11}$$

Here Γ_M is the population Gram matrix, θ^* is the unpenalized population coefficient, and θ^L is the population group-Lasso minimizer. We define $w_j = p'_{\lambda,a}(\|\theta_j^L\|_2)$. The function $p'_{\lambda,a}$ is the SCAD derivative with parameter $a > 2$. We define $\bar{\theta}$ to minimize the weighted criterion on S, with zero blocks on $M \setminus S$, and write $\bar{\theta}_S$ for its restriction to S.

Correct selection follows when the active blocks stay away from zero and each inactive score $\{\Gamma_M(\bar{\theta} - \theta^*)\}_j$ has norm strictly below its penalty weight. Both conditions concern the population criterion. We write z_j for a centered, within-group standardized population coordinate. The population marginal score of group j is $\|\mathbb{E}\{z_j(X_j)(Y-\mu)\}\|_2^2$.

We use r_n and the population coordinates of Section 3, with $\lambda = \lambda_n$. The coordinate and coefficient bounds, projection-residual bound and stationary transition are those of Condition S1. The Gaussian errors satisfy Condition 1(iv).

The implemented penalty equals λ multiplied by the ratio of the screening-block sample standard deviation to σ_Y, and $\widehat{S}$ is the support of the exact one-step minimizer on the screened set. Proposition S4 derives the coordinate, projection and coefficient bounds for a class of correlated population

models. Marginal regularity alone does not imply the joint projection bound. The screening and weighted-score margins remain separate population conditions. The Gaussian error process and its independence of the complete covariate process remain as specified in Section S4. Theorem 1 gives support recovery from the population margins; its proof is in Section S14.

Theorem 1 uses inactive scores after joint re-estimation. It does not require initial norms to separate the groups, as every common threshold does.

We verify the projection condition in Theorem 1 without assuming a property of a fitted residual. We consider a candidate set $M \supset S$ and the marginal and bounded-signal conditions of Section S4, with fixed s. We extend the pairwise representation (S10) from S to M. We permit an empty vector χ_j for a group independent of the other candidate groups. We require the total number of dependence functions to be fixed, their continuous values and first two derivatives to be uniformly bounded, and their categorical values to be uniformly bounded. The matrix $\mathcal{R}$ is defined by (S10) on the nonempty vectors. If all vectors are empty, its eigenvalue condition below is omitted. We write r_M for the augmented local-polynomial population projection residual of the conditional mean on M at bandwidth h_L. The following proposition gives uniform bounds on the resulting criterion.

Proposition S4. *Under the preceding population conditions, if $\lambda_{\min}(\mathcal{R}) \geq c_0 > 0$, then, uniformly in m and G,*

$$cI \preceq \Gamma_M \preceq CI, \qquad \|r_M\|_\infty \leq Ch_L^2, \qquad \|\theta^*\|_2 + \|\theta^L\|_2 + \|\bar{\theta}_S\|_2 \leq C,$$
$$\max_j \sup_x \|z_j(x)\|_2 \leq C\sqrt{G}, \qquad \max_{j,k} \left\|\mathbb{E}\{\|z_j\|_2^2 z_k z_k^\mathsf{T}\}\right\|_{\mathrm{op}} \leq CG.$$

The constants in Proposition S4 do not depend on the number of independent inactive candidates. Screening and weighted-score margins remain separate conditions. The proof is in Section S14.

We give a population submodel in which an inactive initial Lasso coefficient exceeds an active coefficient in absolute value. We define

$$\Gamma_0 = \begin{pmatrix} 1 & -7/10 & 1/20 \\ -7/10 & 1 & -11/20 \\ 1/20 & -11/20 & 1 \end{pmatrix}, \qquad \beta^* = \lambda \left(-\frac{256187}{48700}, -\frac{4745}{487}, 0\right)^\mathsf{T}.$$

The matrix satisfies $\lambda_{\min}(\Gamma_0) > 0.13$. It is the correlation matrix of three Gaussian scores and hence defines a Gaussian-copula additive model after marginal transformation to $[0, 1]$.

The following proposition identifies a selection gain that no threshold of the initial Lasso norms can reproduce.

Proposition S5. *For the criterion (S11) with $\Gamma_M = \Gamma_0$, $\theta^* = \beta^*$, $S = \{1, 2\}$ and $a = 3.7$,*

$$\theta^L = \lambda \left(-\frac{101}{100}, -5, \frac{680}{487}\right)^\mathsf{T}, \qquad \mathrm{supp}(\bar{\theta}) = S.$$

No $t \geq 0$ satisfies $\{j : |\theta_j^L| > t\} = S$. For weights $\lambda^2/|\theta_j^L|$, the weighted population minimizer has support different from S.

The inactive initial norm in Proposition S5 exceeds an active norm. Joint SCAD re-estimation satisfies the inactive Karush–Kuhn–Tucker (KKT) inequality. The adaptive fit restricted to the active set violates it.

The submodel becomes a sequence of sampling models when its coefficients are scaled by the penalty level. For sample size n, we set $\lambda = \lambda_n$ in β^*. We take $Z \sim N(0, \Gamma_0)$ and write Φ for the standard normal distribution function. We define $X_j = \Phi(Z_j)$ and $Y = \beta_1^* Z_1 + \beta_2^* Z_2 + \varepsilon$. The error process has the variance and autoregression bounds in Section S4 and is independent of the covariates. The Gaussian conditional expectation maps each Hermite degree into the same degree. The first-degree subspace gives the matrix Γ_0, and the response score is zero in every higher degree. We write $\widehat{S}^{\mathrm{SCAD}}$ and $\widehat{S}^{\mathrm{adap}}$ for the supports selected by the SCAD step and by the adaptive weights $\lambda_n^2/\|\widehat{\theta}_j^L\|_2$, respectively. The following corollary connects the population calculation to the local-polynomial estimator.

Corollary S4. *Suppose $h_L \to 0$, $G \asymp h_L^{-1}$, $mr_n = o(\lambda_n)$, $\lambda_n = O(1)$, the covariate vectors are independent over time and ε is Gaussian with bounded autoregressive parameter. Then*

$$\Pr(\widehat{S}^{\mathrm{SCAD}} = S) \to 1, \qquad \Pr\{\exists t \geq 0 : \{j : \|\widehat{\theta}_j^L\|_2 > t\} = S\} \to 0, \qquad \Pr(\widehat{S}^{\mathrm{adap}} = S) \to 0.$$

Corollary S4 permits shrinking signals for which the minimum component norm is a fixed multiple of λ_n. In this regime, the order of the initial block norms matters. The condition $mr_n = o(\lambda_n)$ requires $\lambda_n\{n/(G\log n)\}^{1/2} \to \infty$, and the fixed-$c_\lambda$ penalty λ_n^0 of Theorem S2 does not satisfy this condition for fixed d, because the selection score then has the same order as the penalty. Corollary S5 gives a separate comparison at the usual penalty order for the bounded smooth version of this three-group model. Neither result claims that SCAD dominates every adaptive penalty for every population quadratic.

S9 Continuing estimation without a prespecified reporting horizon

The continuing estimator accumulates moments while the selected support is unchanged. It reserves a fraction tending to zero for selection and uses the remaining observations for component estimation.

We fix $\nu \in (0,1)$ and retain the observations at indices

$$\mathcal{I}_{\mathrm{sel}} = \{\lfloor k^{1/\nu} \rfloor : k \geq 1\}. \tag{S12}$$

At time n, their number is $n^\nu\{1+o(1)\}$, and they are used only for screening and one-step selection. These selection observations remain stored individually for later selection updates. We divide all other observations into blocks at the times $t_r = \lfloor r^{5/2} \rfloor$. At a selection update, we apply the screening, group-Lasso and one-step SCAD procedure of Section S2 to the k retained selection observations, with the bandwidth, grid and penalty evaluated at sample size k. The resulting support is used for the next block. The numerical implementation starts with a boundary at 1,000 observations and uses the subsequent t_r. It updates selection when $k \geq 150$ and at least 75 more selection observations have arrived since the previous update; at other boundaries it retains the previous support. A block ending near time t contains about $(15/8)t^{7/20}$ selection indices. For t above about 4×10^4, each block contains more than 75 of them, and the rule updates at every block boundary, as the proofs require. A block is fitted with the support available before that block. When a support update changes the selected set, the stored estimation moments are reset. If selection cannot be computed under the stated tolerance, the point and inference states are cleared until a later selection update succeeds.

Completed estimation blocks retain only their moments; blocks with fewer than 80 nonselection observations are omitted. The observations in the current incomplete block are retained until its scheduled boundary.

At an intermediate reporting time, a temporary calculation fits the observed part of the current incomplete block at the reporting-time bandwidth and adds its moments to the stored states for that report only. The temporary calculation also checks selection when the same observation-count rule permits an update. A changed or unavailable selection clears only the temporary states. The numerical point query then returns zero components, and the interval-report query raises an error. The report does not alter the completed blocks or the support used for the next scheduled block.

The bandwidth in a block ending at t_r is $c_E t_r^{-1/5}$ for point estimation and $c_I t_r^{-1/4}$ for inference. Here $c_E, c_I > 0$. Because local-polynomial powers are not scaled by the bandwidth in this calculation, their coefficients have the same meaning across blocks and the raw quadratic moments are added exactly. The continuing estimates minimize (S1) with $t = x - u_{jg}$ and with h_j replaced by the bandwidth of observation i in the ith summand.

The grid resolution also changes without a prespecified horizon. For each state, h_t denotes its reporting bandwidth, $c_E t^{-1/5}$ for point estimation or $c_I t^{-1/4}$ for inference. We define $\ell_t = 3 \vee \log\log\log(t + e^{e^e})$ and use the nested grids with $G_l = 2^l + 1$ points. The reporting grid is the smallest such grid satisfying $(G_l - 1)h_t \geq \ell_t$. We define u_l as the first time at which grid l is selected as the reporting grid. The moments of this grid start at

$$t_l^{\mathrm{a}} = \left\lceil \frac{u_l}{1 \vee \log\log(u_l \vee e^e)} \right\rceil.$$

This definition gives $t_l^{\mathrm{a}} = o(u_l)$: a resolution accumulates moments well before it becomes the reporting grid. Every active resolution has its own moment state, and at a reporting time the procedure uses the state for the current reporting grid. At a block boundary or reporting time t, we stop adding

moments to grid l when $2(G_l - 1)h_t \leq 1$ and retain its accumulated moments. The current and future reporting grids never satisfy this stopping condition. Under the conditions of Theorem 4, the fraction of observations retained in this state converges in probability to one. Deleted-block fits subtract the corresponding stored moments. The finite numerical designs have $\ell_t = 3$.

We write $N_{E,n}$ and $N_{I,n}$ for the counts of the observations whose moments form the current point-estimation and inference states, respectively. Each state accumulates observations after the later of the last support change and the start of moment collection for its current reporting grid. A missing state has count zero. Thus the two counts refer to the observations actually used by their respective fits. We use Condition 1 with a fixed active-set size s. The risk proof requires marginal densities bounded above and away from zero and two bounded component derivatives; it derives invertibility of the pooled normal matrix. Local-linear smoothing also appears in batch smooth backfitting (Mammen, Linton and Nielsen, 1999), in its Nadaraya–Watson-type variant with local-linear marginal fits (Mammen and Park, 2006), and in its online extension (Yang, Yao and Zhao, 2024). Inference also requires two bounded density derivatives.

We define $h_n = c_E n^{-1/5}$. An empty selected set gives zero components. For expected risk, all point components are also set to zero if the current point state is unavailable, contains fewer than two observations or its normal equation cannot be solved under the stated numerical tolerance. For every other state, $\widehat{f}_j^{\mathrm{C,T}}$ truncates the continuing point estimate $\widehat{f}_j^{\mathrm{C}}$ at $\log N_{E,n}$ as in Section S4.

For $\nu = 3/4$, Theorem 4 gives the component-risk rate and shows that both states retain fractions tending to one. Its proof in Section S15 establishes that support stabilization precedes grid activation with probability tending to one. Section S15.2 derives the interior bias when earlier blocks use larger bandwidths.

The selection rates permit a growing number of inactive covariates and hold uniformly over the retained sample sizes used by the risk theorem. For example, we set $d = \lceil n^{1/10} \rceil$, a fixed active set with component norms bounded away from zero, selection bandwidth $0.85k^{-1/5}$ and candidate size $m_k = d \wedge \lceil 4k^{1/5} \rceil$. Uniformly for $(n/\log n)^{3/4} \leq k \leq n^{3/4}$, the matrix restriction is bounded by $Ck^{-2/5}\log^3 n = o(1)$ and the minimum-signal error is bounded by $Ck^{-2/5}\sqrt{\log n} = o(1)$. These rates use the same selection settings as the continuing numerical procedure.

Inference uses bandwidth $c_I t^{-1/4}$, and $\widehat{f}_j^{\mathrm{C}}$ in Theorem 5 is the estimate from this inference state. It differs from the point estimate of Theorem 4 through the bandwidth and the state. The number of estimation blocks is of order $n^{2/5}$, and their maximum length is $O(n^{3/5})$. The exact deletion calculation recomputes the centering and within-group normalization after each block is removed. The coverage event in that theorem concerns a fixed nonempty collection of active component values at a deterministic reporting time, with each continuous value in a fixed interior interval. The resampling maximum is taken over the reported values of the selected groups. We write $\widehat{\mathrm{se}}_j^{\mathrm{C}}(x)$ for the deleted-block standard error. All intervals are empty if the inference state or a required fit is unavailable, if fewer than two deletion blocks remain, or if a requested evaluation is unidentified or has a nonpositive or nonfinite standard error; in each of these cases the coverage event in that theorem is false.

Theorem 5 includes dependence between selection and estimation. Correct selection gives the known-support fit; selection error bounds the coverage difference. The finite-resampling bounds after Theorem S3 also apply.

The following proposition bounds the memory of the continuing calculation, counted as the number of stored real numbers.

Proposition S6. *Suppose that the population selection conditions of Theorem 4 hold. With probability tending to one, the number of stored scalars is*

$$O\{dn^{3/4} + s^2\ell^2_{n\log\log n}n^{2/5}(\log\log n)^{2/5} + s^2\ell^2_{n\log\log n}n^{9/10}(\log\log n)^{1/2}\}.$$

Completed estimation blocks retain no individual observations. The current incomplete block and the $n^{3/4}$ selection observations are retained.

Proposition S6, proved in Section S15, counts the sparse selection observations, moments at each active resolution and deleted-block inference states. For fixed d and s, the bound is sublinear, of order $n^{9/10}$ up to iterated-log factors. Its $dn^{3/4}$ term can dominate when d grows; the bound does

not assert memory optimality. The lower-bound analysis of Quan and Lin (2024) concerns global one-pass nonparametric estimation under a memory constraint counted in stored real numbers and provides a different benchmark.

S10 Proofs of population component-risk results

Under the product model, the population quadratic separates the component effects. We bound the change from this quadratic to its empirical version and account for the estimated component centers. The selection proof uses the same coordinates.

S10.1 Population local-polynomial coordinates

We use population coordinates that center each group and normalize its second-moment matrix. We define $D_j^0 = \mathbb{E}D_j(X_j)$ and $c_j^0 = \mathbb{E}a_j(X_j)$. We write e_j for the coefficient vector of the constant polynomial one. Then $D_j^0 e_j = c_j^0$ and $(c_j^0)^\mathsf{T} e_j = 1$. We choose R_j^0 such that $(R_j^0)^\mathsf{T} D_j^0 R_j^0 = I$ and $(c_j^0)^\mathsf{T} R_j^0 = 0$. We set $z_j(x) = (R_j^0)^\mathsf{T} a_j(x)$ and $A_j(x) = (R_j^0)^\mathsf{T} D_j(x) R_j^0$.

For the equally spaced grid, we define $\Delta = (G-1)^{-1}$, and write $q_g = q_{jg}$ and $\pi_g = \pi_{jg}$ for the fixed group j. The trapezoidal weights are Δ except for weight $\Delta/2$ at zero and one. Bounds for a Riemann sum of the nonnegative Epanechnikov kernel give

$$ch \le \sum_g q_g K\{(x-u_g)/h\} \le Ch, \qquad 0 \le x \le 1.$$

For the lower bound, the interval $[x-h/2, x+h/2] \cap [0,1]$ has length at least $h/2$ and contains a total quadrature weight at least ch when $h/\Delta \ge 4$. On this interval, $K \ge 9/16$. The upper bound follows from $K \le 3/4$ and support in an interval of length $2h$. For each grid node and every $w \in \mathbb{R}^2$, we change variables $x = u_g + ht$ to obtain

$$\frac{c}{G}\|w\|_2^2 \le \int \pi_g(x)(w_0 + w_1 t)^2 p_j(x)\, dx \le \frac{C}{G}\|w\|_2^2.$$

At every node, the integration interval contains $[0, 1/2]$ or $[-1/2, 0]$. The 2×2 polynomial moment matrices of these two intervals are positive definite, with constants independent of g, j, G, and h. The disjoint node blocks give

$$cG^{-1} I \preceq D_j^0 \preceq CG^{-1} I, \quad \|R_j^0\|_{\mathrm{op}} \le C\sqrt{G}, \quad \sup_x\{\|z_j(x)\|_2^2 + \|A_j(x)\|_{\mathrm{op}}\} \le CG.$$

For categorical groups the same assertions hold with G replaced by the fixed number of levels. The lower bound on the density p_j used above is replaced by the lower bound on the level probabilities.

We compare the population local-polynomial projection with the target f_j. We write $t_j^*(U_j, X_j)$ for the population orthogonal projection of $f_j(X_j)$ onto the local-polynomial group, with coefficient vector $(D_j^0)^{-1}\mathbb{E}\{a_j(X_j) f_j(X_j)\}$ and population mean zero. We write this coefficient vector as $R_j^0 \eta_j^*$ and set $g_j^*(x) = z_j(x)^\mathsf{T} \eta_j^*$. The local-linear Taylor error about u_g is bounded by $Lh^2/2$ on the support of π_g. The inverse moment bound controls both coefficients of the projection of this Taylor error, including at the boundary. Taking the supremum over the conditional support of U_j given x gives

$$\sup_{x, U_j} |t_j^*(U_j, x) - f_j(x)| \le CLh^2, \quad \|g_j^* - f_j\|_{P_j} \le Lh^2/2, \quad \|\eta_j^*\|_2 \ge \|f_j\|_{P_j} - Lh^2/2. \tag{S13}$$

Conditional Jensen and projection give the middle bound. Categorical projection is exact. The coefficient bound uses fixed moment matrices, without differentiable densities.

We collect the auxiliary indices as $U = (U_j)_{j \in S}$ and write $\mathbb{E}_U$ for expectation over these indices conditional on the covariates. For the active set, the full feature vector $\phi(U, X)$ has an intercept and centered group blocks $(R_j^0)^\mathsf{T} v_j(U_j, X_j)$. Independence of the coordinates implies

$$\mathbb{E}\{\phi\phi^\mathsf{T}\} = I_q, \qquad M_i = \mathbb{E}\{\phi(U, X_i)\phi(U, X_i)^\mathsf{T} \mid X_i\} \succeq 0, \qquad \|M_i\|_{\mathrm{op}} \le Cq.$$

The vector η^* contains μ and the η_j^*. We define $r_i(U) = m(X_i) - \phi(U, X_i)^\mathsf{T}\eta^*$. Here $m(X_i) = \mu + \sum_{j\in S} f_j(X_{ij})$. The separate projection residuals have mean zero and are independent across coordinates. Hence

$$\mathbb{E}r_i(U)^2 \le s_c L^2 h^4/4, \quad \mathbb{E}\{\phi r_i(U)\} = 0, \quad \sum_j \|g_j^*\|_{P_j}^2 \le B^2. \tag{S14}$$

S10.2 Comparison of quadratic forms

The unpenalized fit is stable when its empirical second-moment matrix is close to the population matrix I_q. We bound the probability that their spectral-norm difference exceeds one half, which gives a uniform bound on the empirical inverse. For independent observations, the positive-semidefinite matrix Chernoff inequality of Tropp (2012, Theorem 5.1) applies to $M_i/(Cq)$, because M_i has mean I_q and upper bound CqI_q. In particular,

$$\Pr\left(\left\|N^{-1}\sum_i M_i - I_q\right\|_{\mathrm{op}} > 1/2\right) \le 2q\exp(-cN/q).$$

We separate the indices into their ℓ residue classes. Successive observations within a class have no independent redraw between them with probability ρ^ℓ. We couple each class to independent stationary copies; the total probability that the coupling fails is at most $N\rho^\ell$. Each class has size at least $N/(2\ell)$. We apply the preceding matrix bound to the copies before conditioning on agreement. On agreement, when every class average is within $1/2$ of I_q, the weighted average of the class averages is within $1/2$. We define $\widehat{M} = N^{-1}\sum_i M_i$. Thus,

$$\Pr(\|\widehat{M} - I_q\|_{\mathrm{op}} > 1/2) \le 2\ell q\exp\{-cN/(q\ell)\} + N\rho^\ell. \tag{S15}$$

The coupling statement also follows from the independent-redraw construction directly, without an assumption about a fitted or selected quantity.

S10.3 Proof of Theorem S1

We first bound the coefficient error when the design is well conditioned, then account for selection and truncation. We define $\mathcal{G} = \{\|\widehat{M} - I_q\|_{\mathrm{op}} \le 1/2\}$ and write $\widehat{\eta}$ for the known-support unpenalized fit in population coordinates. Empirical and population centering represent the same additive polynomial space with an intercept. On $\mathcal{G}$, their fitted arrays coincide and

$$\widehat{\eta} - \eta^* = \widehat{M}^{-1}(b_r + b_\varepsilon), \quad b_r = N^{-1}\sum_i \mathbb{E}\{\phi r_i(U) \mid X_i\}, \quad b_\varepsilon = N^{-1}\sum_i \mathbb{E}(\phi \mid X_i)\varepsilon_i.$$

For every $u \in \mathbb{R}^q$, conditional Cauchy–Schwarz gives

$$|u^\mathsf{T} b_r|^2 \le (u^\mathsf{T}\widehat{M}u)N^{-1}\sum_i \mathbb{E}\{r_i(U)^2 \mid X_i\}.$$

It follows that $\|b_r\|_2^2 \le (3/2)N^{-1}\sum_i \mathbb{E}(r_i^2 \mid X_i)$ on $\mathcal{G}$. We define $\bar{\phi}_i = \mathbb{E}(\phi \mid X_i)$ and $V_N = N^{-1}\sum_i \bar{\phi}_i\bar{\phi}_i^\mathsf{T}$. Conditional Jensen gives $V_N \preceq \widehat{M}$. The error covariance is bounded by vI_N, independently of the design. This gives

$$\mathbb{E}\{\|b_\varepsilon\|_2^2 \mid X\} \le (v/N)\,\mathrm{tr}(V_N), \qquad \mathbb{E}[\mathbf{1}_\mathcal{G}\|\widehat{\eta} - \eta^*\|_2^2] \le C\{s_c L^2 h^4 + vq/N\}.$$

This calculation uses $\|\widehat{M}^{-1}\|_{\mathrm{op}} \le 2$ and (S14).

It remains to convert this coefficient error into the component loss, which also accounts for the estimated center of each group. We write the population-centered fitted component as $\widehat{g}_j(x) = z_j(x)^\mathsf{T}\widehat{\eta}_j$ and define $\bar{z}_j = N^{-1}\sum_i z_j(X_{ij})$. The reported untruncated component is exactly $\widehat{f}_j(x) = \widehat{g}_j(x) - \bar{z}_j^\mathsf{T}\widehat{\eta}_j$. Because $\widehat{g}_j - f_j$ has population mean zero,

$$\sum_j \|\widehat{f}_j - f_j\|_{P_j}^2 = \sum_j \|\widehat{g}_j - f_j\|_{P_j}^2 + \sum_j (\bar{z}_j^\mathsf{T}\widehat{\eta}_j)^2.$$

The first sum is bounded by $2\|\widehat{\eta}-\eta^*\|_2^2+s_cL^2h^4/2$. The intercept row of $\widehat{M}-I_q$ gives $\sum_j\|\bar{z}_j\|_2^2\le 1/4$ on $\mathcal{G}$. Hence

$$\sum_j\{\bar{z}_j^{\mathsf{T}}(\widehat{\eta}_j-\eta_j^*)\}^2\le\tfrac{1}{4}\|\widehat{\eta}-\eta^*\|_2^2.$$

The covariance of the centered projection g_j^* at lag k equals $\rho^k\|g_j^*\|_{P_j}^2$ under the redraw transition. Thus

$$\mathbb{E}\sum_j(\bar{z}_j^{\mathsf{T}}\eta_j^*)^2\le\frac{1+\rho_0}{1-\rho_0}\frac{B^2}{N}.$$

These three bounds prove the first line of (S4) on the event $\mathcal{G}$ for the known support.

On $\mathcal{G}\cap\{\widehat{S}=S\}$, the selected and known-support estimates coincide. Truncation does not increase squared error when $|f_j|\le B\le T_N$. On the complement, at most $m+s$ differences are nonzero, each bounded by $2T_N$, giving loss at most $4(m+s)T_N^2$. The raw and normalized eigenvalue thresholds N^{-3} and N^{-1} are eventually below the corresponding eigenvalue bounds on $\mathcal{G}$. Indeed, $E_j=[e_j,R_j^0]$ is nonsingular because R_j^0 has full column rank, $(c_j^0)^{\mathsf{T}}R_j^0=0$ and $(c_j^0)^{\mathsf{T}}e_j=1$. Because $E_j^{\mathsf{T}}\widehat{D}_jE_j$ is a principal submatrix of $\widehat{M}$, $\lambda_{\min}(\widehat{D}_j)\ge\lambda_{\min}(E_j^{\mathsf{T}}\widehat{D}_jE_j)/\|E_j\|_{\mathrm{op}}^2\ge c/G\ge c/N$ by $\|E_j\|_{\mathrm{op}}^2\le CG$ and $G\le N$. The normalized final matrix is obtained from a Schur complement of $\widehat{M}$ by block normalization. On $\mathcal{G}$ the eigenvalues of $\widehat{M}$ lie in $[1/2,3/2]$, and the normalized matrix has smallest eigenvalue at least $(1/2)/(3/2)=1/3$. The same loss bound covers the zero decision on other samples. Adding (S15) and δ_{sel} proves (S4). This completes the proof.

S10.4 Sample matrix and score bounds for selection

To transfer population separation to sample selection, we control errors in the marginal scores, the joint matrix and the residual scores. We set $G=G_L$ and $h=h_L$, with block sizes of order n. We use population coordinates to establish the bounds for Proposition S1.

The independent-redraw process gives constant design segments of lengths $L_1,\dots,L_K$. Except for two possibly incomplete boundary segments, these lengths are geometric with success probability $1-\rho$, independently of the draws from $\prod_j P_j$. For fixed $A>0$, truncating the lengths at $C_A\log n$ and applying the scalar Bernstein inequality give

$$\max_k L_k\le C_A\log n,\qquad\sum_k L_k^r\le C_{A,r}n\quad(r=2,4),\tag{S16}$$

except with probability $o(n^{-A})$. For the last bound, we truncate up to n independent geometric lengths at $C_A\log n$. Their finite moments give variance at most Cn and summands bounded by $C_A^r\log^r n$. For a deviation equal to a fixed multiple of n, Bernstein's exponent is at least $cn/\log^r n$. We take a union bound over terminal indices and the two boundary pieces. For $\rho=0$, every length equals one.

Conditional on these lengths, we apply the rectangular matrix Bernstein inequality of Tropp (2012, Theorem 6.1 and Remark 6.3) to the independent design vectors. The dilation of a rectangular matrix U is $\left(\begin{smallmatrix}0&U\\U^{\mathsf{T}}&0\end{smallmatrix}\right)$. For $j\ne k$, $\mathbb{E}z_jz_k^{\mathsf{T}}=0$ and

$$\|z_jz_k^{\mathsf{T}}\|_{\mathrm{op}}\le CG,\qquad\max\{\|\mathbb{E}\|z_k\|_2^2z_jz_j^{\mathsf{T}}\|_{\mathrm{op}},\|\mathbb{E}\|z_j\|_2^2z_kz_k^{\mathsf{T}}\|_{\mathrm{op}}\}\le CG.$$

The second bound uses the independence of the two coordinates and the inequality $\mathbb{E}z_jz_j^{\mathsf{T}}\preceq I$. The same variance order holds for A_j-I because $0\preceq A_j\preceq CGI$ and $\mathbb{E}A_j=I$. For the empirical means $\bar{z}_j$, we apply the same inequality to the rectangular dilation of z_j, whose variance parameter is at most CG. The length bounds in (S16) give, uniformly over j,k, with probability $1-o(n^{-A})$,

$$\max_j\|\bar{A}_j-I\|_{\mathrm{op}}+\max_j\|\bar{z}_j\|_2+\max_{j\ne k}\|\overline{z_jz_k^{\mathsf{T}}}\|_{\mathrm{op}}\le C_A\left\{\sqrt{\frac{G\log(dn)}{n}}+\frac{G\log n\log(dn)}{n}\right\}=:a_n.\tag{S17}$$

Bars denote block averages. We apply the bounds separately to screening and selection, without conditioning on a selected set. The rate condition of Theorem S2 gives $ma_n\to0$.

For marginal screening, we need a bound on one further weighted matrix. Conditional on the lengths, the same matrix inequality applied to $L_k^2 z_j(X_j^{(k)}) z_j(X_j^{(k)})^\mathsf{T}/n$, where $X_j^{(k)}$ is the constant covariate value of group j on the kth segment, gives

$$\max_j \left\| n^{-1} \sum_k L_k^2 z_j(X_j^{(k)}) z_j(X_j^{(k)})^\mathsf{T} \right\|_{\mathrm{op}} \le C_A \tag{S18}$$

with probability $1 - o(n^{-A})$. The conditional mean is bounded by $C_A I$, the variance by $C_A G/n$ using $\sum_k L_k^4 \le C_A n$, and each summand by $C_A G \log^2 n/n$. The rate condition permits a union bound over d groups. By (S17), empirical centering changes the bound by a constant factor.

We bound the noise and approximation parts of the marginal screening score separately. Conditional on the redraw lengths and the observed path of X_j, the sums of the other component values are independent across redraw segments. Hoeffding's bound and $\sum_k \|f_k\|_\infty^2 \le B^2$ give sub-Gaussian parameter CB^2 for these sums. The Gaussian errors are independent of the design with covariance at most vI. By (S18), each unit contrast of the normalized marginal noise score has sub-Gaussian variance parameter $C(B^2 + v)/n$. We pass from single contrasts to the norm by a net argument: a $1/2$-net in dimension at most $2G$ has at most 5^{2G} elements, and $\|u\|_2$ is at most twice the largest $|t^\mathsf{T} u|$ on the net. A union bound over the nets and the d groups bounds the norm of every normalized marginal noise score by

$$C_A(B + \sqrt{v})\sqrt{\{G + \log(dn)\}/n} \tag{S19}$$

with probability $1 - o(n^{-A})$. Conditional means of the Gaussian errors remain zero under this conditioning, without requiring independent error observations.

By (S13), the marginal approximation score is centered with summand norm at most $CLh^2\sqrt{G}$. Rectangular dilation, vector Bernstein and (S16) bound the average of this score by $C_A L h^2 \{G \log(dn)/n\}^{1/2}$ plus a smaller bounded-summand term. The intercept correction, a product of $\bar{z}_j$ and a centered response average, is also smaller than (S19).

For the joint selection score, we define $R_i = Y_i - \mu - \sum_{j \in S} g_j^*(X_{ij})$ and

$$u_{ij} = z_j(X_{ij}) R_i - \{A_j(X_{ij})\eta_j^* - z_j(X_{ij}) g_j^*(X_{ij})\}, \qquad \eta_j^* = 0 \quad (j \notin S).$$

Population projection and coordinate independence give $\mathbb{E} u_{ij} = 0$. Conditional on the design, the Gaussian part of $\bar{u}_j$ has covariance at most CvI/n on the event that (S17) holds. The sphere-net bound for the Gaussian part is $C_A\sqrt{v\{G + \log(dn)\}/n}$ uniformly over j. We substitute $t_k^* = f_k + O(Lh^2)$ to cancel the exact f_k terms and bound the design part by $CLsh^2\sqrt{G}$. Vector Bernstein then bounds the block average of the design part by

$$C_A L s h^2 \left\{ \sqrt{\frac{G \log(dn)}{n}} + \frac{\sqrt{G} \log n \log(dn)}{n} \right\}.$$

These bounds also apply to inactive groups but do not replace the inactive score by the marginal response score.

The penalty uses the screening-block response standard deviation, whose consistency follows from the bounds below. We write $Y_i - \mu = F_i + \varepsilon_i$. Here F_i is the sum of the active components. Conditional on the redraw lengths, the values of F_i from different redraw segments are independent, centered, and sub-Gaussian with parameter CB^2. We apply the scalar Bernstein inequality to the weighted F_i and F_i^2 using the length bounds. Diagonalizing the Gaussian error covariance gives

$$\mathbb{E} \exp\{t(\varepsilon^\mathsf{T}\varepsilon - \mathbb{E}\varepsilon^\mathsf{T}\varepsilon)\} \le \exp(Cnt^2), \qquad |t| \le c,$$

because the covariance eigenvalues are bounded by v. Conditional on F, $n^{-1}\sum_i F_i\varepsilon_i$ is centered Gaussian with variance at most $v\sum_i F_i^2/n^2$. These bounds give convergence of the sample mean and the sample second moment at every fixed tolerance, with exceptional probability $o(n^{-A})$. Because $\sigma_Y \ge \sigma_-$, the sample standard deviation is relatively consistent and the implemented penalty is $(1 + o_p(1))\lambda_n^0$ at this probability level.

S10.5 Proof of Theorem S2

We use the bounds of the preceding subsection simultaneously, with a fixed exponent above A. We define $F_j = \bar{A}_j - \bar{z}_j\bar{z}_j^{\mathsf{T}}$. We center z_j by $\bar{z}_j$ and normalize by T_j with $T_j^{\mathsf{T}} F_j T_j = I$. The cross block is

$$H_{jk} = T_j^{\mathsf{T}}\{\overline{z_j z_k^{\mathsf{T}}} - \bar{z}_j\bar{z}_k^{\mathsf{T}}\}T_k.$$

Equation (S17) implies $\|T_j\|_{\mathrm{op}} \le 2$ and the row-sum bound $\max_j \sum_{k\neq j,\ k\in M'} \|H_{jk}\|_{\mathrm{op}} \le Cma_n = o(1)$ for every set M' of at most m groups. This uniform assertion applies to the data-dependent screening set.

The marginal population score norm $\|\eta_j^*\|_2$ is zero for inactive groups and at least $b_n - Lh^2/2$ for active groups by (S13). In (S17)–(S19), $\log(dn) = o(G)$ and $\log d \to \infty$ give a noise score of order $o(\lambda_n^0)$, and $h^2 + \lambda_n^0 = o(b_n)$ controls the approximation error. Every active score eventually exceeds every inactive score; if $m \ge s$, retaining the m largest scores includes S.

In the selection coordinates, the empirically recentered population vector has block $\theta_j^* = T_j^{-1}\eta_j^*$ and norm $(\eta_j^{*\mathsf{T}} F_j \eta_j^*)^{1/2} \ge (1 - o(1))\|\eta_j^*\|_2$. The empirical normal-equation residual at this coefficient vector is

$$e_j = T_j^{\mathsf{T}}(\bar{u}_j - \bar{z}_j\bar{R}).$$

Transferring each empirical component mean to the intercept gives this identity, in which $\bar{R}$ is the block average of the residuals R_i. The joint score bounds, $sh^2 \to 0$, and $\log(dn) = o(G)$ give $\max_j \|e_j\|_2 = o(\lambda_n^0)$. The implemented penalty is asymptotically equivalent to λ_n^0, and the minimum active norm divided by the penalty tends to infinity. Proposition S1 then shows that the SCAD-selected support equals S.

The score and nonzero-group tolerance N_L^{-3} of Section S20 is $o(\lambda_n^0)$ and $o(b_n)$. Group soft thresholding is nonexpansive. With row sum at most $1/4$, induction over updated coordinates contracts the maximum block error by a factor $1/4$ per sweep toward the unique minimizer. After $C\log n$ sweeps, the error is $O(n^{-4})$. The inactive dual gap in Proposition S1 is at least $\lambda_n^0/3$, giving exactly zero inactive updates once the error is sufficiently small. Active norms remain above $b_n/2$. The sweep limit $100 + \lceil 30\log N_L \rceil$ eventually exceeds the required number of sweeps. Positive population moments and (S17) ensure that the eigenvalue thresholds retain all estimable directions on these samples. All exceptional probabilities are $o(n^{-A})$, proving the uniform selection bound. This completes the proof.

S10.6 Proof of Corollary S1

The preceding matrix and score bounds are stated in terms of $\log(dn)$ and do not use $\log n \lesssim \log d = o(G_L)$. The matrix rate gives $ma_n = o(1)$ in (S17). By $N_L \asymp n$ and $\sigma_Y \ge \sigma_- > 0$, the Gaussian and marginal noise score in (S19) satisfies

$$\frac{\sqrt{\{G_L + \log(dn)\}/n}}{\lambda_n^0} \le C\left\{\frac{1}{\log d} + \frac{1}{G_L} + \frac{\log n}{G_L \log d}\right\}^{1/2} \longrightarrow 0.$$

For the centered approximation part of the joint score,

$$sh_L^2\left\{\frac{\log(dn)}{\log d}\right\}^{1/2} \le \frac{C\{m^2 G_L \log^3(dn)/n\}^{1/2}}{\log(dn)\sqrt{\log d}} \longrightarrow 0.$$

This gives $sh_L^2 \to 0$. The matrix rate bounds the ratio of the bounded-summand and square-root terms by $C\log n\{\log(dn)/n\}^{1/2} = o(1)$, also for the marginal approximation score. The empirical mean correction divided by λ_n^0 is at most $C_A\{\log(dn)\log n/(n\log d)\}^{1/2} = o(1)$. These bounds show that screening retains S and that the maximum joint score is $o(\lambda_n^0)$. Response-scale consistency uses only the bounded signal and redraw lengths, without a lower bound on d.

The signal and matrix bounds give the active norm and row-sum inequalities in Proposition S1. The strict inactive inequalities and numerical tolerances follow as in Theorem S2, including the sweep contraction and eigenvalue thresholds. Each bound fails with probability $o(n^{-A})$ for arbitrary fixed A; their intersection proves the assertion. This completes the proof.

S10.7 Proof of Corollary S2

In fixed dimension a sufficiently large penalty constant controls the selection score. We use the population coordinates of Theorem S2, with block sizes proportional to n and $G = G_L \asymp n^{1/5}$. Constants are independent of A unless subscripted by A.

We write X_{bj} for the value of coordinate j on redraw segment b. The geometric redraw lengths L_b have uniformly bounded moments of every fixed order. Truncation at $C_A \log n$ has exceptional probability $o(n^{-A})$ for sufficiently large C_A. The scalar Bernstein inequality has variance $O(n)$ and a bounded-summand term that is a fixed power of $\log n$. This gives leading constants independent of A. These bounds give

$$\sum_b L_b^2 \le Cn, \qquad \sum_b L_b^4 \le Cn$$

with probability $1 - o(n^{-A})$, with a fixed leading constant C. Boundary pieces satisfy the same bounds. Conditional on these lengths, matrix Bernstein applied to $L_b^2 z_j(X_{bj}) z_j(X_{bj})^\mathsf{T}$ gives

$$\max_j \left\| \sum_b L_b^2 z_j(X_{bj}) z_j(X_{bj})^\mathsf{T} \right\|_{\mathrm{op}} \le Cn.$$

Its variance parameter is at most CGn, and its largest summand is at most $C_A G \log^2 n$. A deviation of order n has probability at most $CG \exp\{-c_A n/(G \log^2 n)\} = o(n^{-A})$. The unweighted empirical normalizations have eigenvalues in $[1/2, 2]$ at the same probability level. We replace z_j by its empirically centered and normalized version, with $F_j = \bar{A}_j - \bar{z}_j \bar{z}_j^\mathsf{T}$, to preserve the weighted matrix bound: $\|F_j^{-1/2}\|_{\mathrm{op}} \le \sqrt{2}$, $\|\bar{z}_j\|_2 = o(1)$, and $\sum_b L_b^2 \le Cn$ bound the two terms from $(z_j - \bar{z}_j)(z_j - \bar{z}_j)^\mathsf{T}$ by a fixed multiple of n.

Conditional on redraw lengths and coordinate j at every redraw, the other effects have a centered sub-Gaussian sum with variance proxy CB^2 by coordinate independence. The weighted matrix bound gives variance proxy CB^2/n for each unit contrast of the normalized marginal score; the Gaussian part contributes Cv/n. A union bound over a $1/2$-net of at most 5^{2G+C} points and the fixed number of groups bounds the norm of the normalized marginal score by $C(B + \sqrt{v})\sqrt{G/n}$ with probability $1 - C\exp(-cG) - o(n^{-A})$ for sufficiently large fixed C. The projection and centering corrections are at most $C_A h_L^2 \sqrt{G \log n/n} + o(\sqrt{G/n})$ by (S13) and the preceding subsection. Under normalization $F_j^{-1/2}$, the own-component score is $F_j^{1/2}\eta_j^*$, with norm at least $(1 - o(1))(b_n - Lh_L^2/2)$ for active groups and zero for inactive groups. We take b_n a sufficiently large fixed multiple of $h_L^2 + \lambda_n^0$ to separate the scores. Screening retains S for every $m \ge s$.

A sharper Gaussian bound suffices for the joint selection score. In the empirically centered coordinates, we write $\widetilde{a}_{ij} = R_j^\mathsf{T} a_j(X_{ij})$ and

$$e_j^\varepsilon = N_L^{-1} \sum_{i \in I_L} \widetilde{a}_{ij} \varepsilon_i.$$

Conditional Jensen and the exact empirical normalization give

$$\frac{1}{N_L} \sum_{i \in I_L} \widetilde{a}_{ij} \widetilde{a}_{ij}^\mathsf{T} \preceq I, \qquad \mathrm{Cov}(e_j^\varepsilon \mid X) \preceq \frac{v}{N_L} I.$$

There are at most $2G + C$ coordinates in a group. The moment-generating function of a squared Gaussian norm gives, for a sufficiently large fixed C_1,

$$\Pr\left(\|e_j^\varepsilon\|_2^2 > C_1 vG/N_L \mid X\right) \le \exp(-C_1 G/4)\, 2^{(2G+C)/2} \le C\exp(-cG).$$

To see this, we diagonalize the conditional covariance and apply Markov's inequality at $N_L/(4v)$, and every factor of the moment-generating function of the squared norm is at most $\sqrt{2}$.

The non-Gaussian joint score is the centered projection residual. For fixed active size, Bernstein bounds its maximum norm by

$$C_A h_L^2 \left\{ \sqrt{\frac{G \log n}{n}} + \frac{\sqrt{G} \log^2 n}{n} \right\} + o(\lambda_n^0) = o(\lambda_n^0).$$

These bounds include the empirical-intercept correction. The bounded signal and Gaussian quadratic-form bound give relative consistency of the screening estimate of σ_Y with probability $1 - o(n^{-A})$, without requiring d to grow. The bounds on Gh_L and $\sigma_Y \geq \sigma_- > 0$ allow a fixed c_λ such that the Gaussian bound is below $\lambda_n^0/16$. Eventually, the following inequalities hold simultaneously:

$$\tfrac{1}{2}\lambda_n^0 \leq \widehat{\lambda} \leq 2\lambda_n^0, \qquad \max_j \|b_j - (H\theta^*)_j\|_2 \leq \widehat{\lambda}/4, \qquad \max_j \sum_{k \neq j} \|H_{jk}\|_{\mathrm{op}} \leq 1/4.$$

The last inequality follows from the fixed dimension and the matrix bound $O\{\sqrt{G \log n/n} + G \log^2 n/n\} = o(1)$. Empirical normalization changes each population coefficient norm by a factor $1 + o(1)$, and a sufficiently large fixed constant in $b_n \geq C(h_L^2 + \lambda_n^0)$ gives the minimum-signal condition $\min_{j \in S} \|\theta_j^*\|_2 > (a + 5/3)\widehat{\lambda}$. Proposition S1 proves the support assertion.

The $1/4$ contraction per sweep gives the same support under the prescribed sweep limit, group tolerance and eigenvalue cutoffs as in Theorem S2. The exceptional probability is $C \exp(-cG) + o(n^{-A}) = o(n^{-A})$, with penalty and signal constants independent of A.

On a deterministic subsequence, retention probabilities are at most ρ_0 and the error covariance remains bounded by vI. Coupling with redraw probability $1 - \rho_0$ dominates retained segment lengths; splitting segments decreases their squared and fourth-power sums. The bounds are uniform over the population class and retained sizes in continuing estimation. This completes the proof.

S10.8 Proof of Proposition S2

We first verify the stationary sufficient condition. For deterministic $u = (u_1, \ldots, u_N)^\mathsf{T}$, the spectral representation and Parseval's identity give

$$\mathrm{Var}\left(\sum_{i=1}^N u_i \varepsilon_i\right) = \int_{-\pi}^{\pi} \left|\sum_{i=1}^N u_i e^{\mathrm{i}i\omega}\right|^2 g_\varepsilon(\omega)\, d\omega \leq v \sum_{i=1}^N u_i^2.$$

Hence $\mathrm{Cov}(\varepsilon_1, \ldots, \varepsilon_N) \preceq vI_N$. Next, we assume this population covariance bound directly. The covariance of any finite deterministic subsequence is a principal submatrix of a prefix covariance and satisfies the same bound. Independence of the errors and design gives the same conditional bound and conditional mean zero. In the proof of Theorem S1, these imply a score covariance of at most $(v/N)\widehat{M}$; the approximation, matrix and centering bounds depend only on the design. Truncation bounds the exceptional loss, proving (S4) without stationarity or Gaussianity, with the actual selection error $\Pr(\widehat{S} \neq S)$. Independent centered errors with variances bounded by v have diagonal covariance matrices bounded by vI_N. This verifies the stated example.

Suppose now that the errors are Gaussian with common variance $\sigma^2 \geq \sigma_-^2 > 0$. The conditional score bounds in all three selection proofs use only the Gaussian distribution of the errors and the upper covariance bound just established. The response-scale estimate also has the required concentration. We write Σ_N for the error covariance and $\sigma^2 = \mathrm{Var}(\varepsilon_i)$. The eigenvalues λ_r of the error covariance satisfy $0 \leq \lambda_r \leq v$ and $\sum_r \lambda_r = N\sigma^2$. Diagonalizing Σ_N gives, for $|u| < (2v)^{-1}$,

$$\log \mathbb{E} \exp\{u(\|\varepsilon\|_2^2 - N\sigma^2)\} = \sum_r \{-u\lambda_r - \tfrac{1}{2}\log(1 - 2u\lambda_r)\} \leq \frac{Nv^2u^2}{1 - 2v|u|}.$$

Chernoff's inequality bounds the probability of a fixed positive deviation of $N^{-1}\|\varepsilon\|_2^2$ from σ^2 by $2e^{-cN}$. Also, $\bar{\varepsilon}$ is Gaussian with variance at most v/N.

We define $F_i = \sum_{j \in S} f_j(X_{ij})$. Conditional on the design, $N^{-1}\sum_i F_i\varepsilon_i$ is Gaussian with variance at most $vN^{-2}\sum_i F_i^2$. By the Bernstein bounds for F_i and F_i^2 in Section S10.4, $N^{-1}\sum_i F_i^2$ and $\bar{F}$ converge to their expectations at every fixed tolerance with probability $1 - o(N^{-A})$. The Gaussian bound gives the same order for the cross term. Expanding the sample variance gives $\widehat{\sigma}_Y/\sigma_Y \to 1$ with exceptional probability $o(N^{-A})$, using a denominator at least σ_- and no decay rate of the error covariance.

The remaining selection inequalities depend only on the design or deterministic optimality equations, proving all three selection results under their respective conditions. For a deterministic subsequence,

the error covariance is a principal submatrix, still bounded by vI with diagonal σ^2. The Gaussian quadratic and score bounds are uniform over the retained sizes in continuing estimation.

The risk bound and the applicable selection probability give Corollary 1. For Theorem 4, uniform selection bounds give support stabilization and hence the deterministic oracle estimation set. The score covariance bound $(v/N)\widehat{M}_n$ gives the integrated variance and risk bounds. Observation counts and storage depend only on the design schedule and support stabilization.

The following example shows that no positive lower bound on the spectral density is needed. For $\varepsilon_i = \sigma(\xi_i - \xi_{i-1})/\sqrt{2}$ with independent standard Gaussian ξ_i, $2\pi g_\varepsilon(\omega) = \sigma^2(1-\cos\omega)$ is bounded by $2\sigma^2$ and zero at the origin. The proposition applies without a positive spectral lower bound and makes no additional inference claim. This completes the proof.

S10.9 Proof of Corollary 1

We choose ℓ to make the selection and matrix-error terms negligible in the risk bound. Because $s \le m$, $N \asymp n$ and $G \asymp G_L$, the matrix rate in Corollary S1 implies $sG\log^2 N/N \le Cm^2 G_L \log^3(dn)/n = o(1)$. For Corollary S2, the same bound follows from fixed s and $G \asymp N^{1/5}$.

We choose $\ell = \lceil C_A \log N \rceil$ in (S4). The constant C_A is chosen uniformly for $\rho \le \rho_0 < 1$ such that $N\rho^\ell \le N^{-A}$. The condition $sG\log^2 N = o(N)$ makes $2\ell q \exp\{-cN/(q\ell)\} = o(N^{-A})$ for every fixed A. Each applicable selection result in Section S4 controls δ_{sel} at the same order. Under either corollary, $m = o(\sqrt{n})$, and the exceptional loss is smaller than N^{-1} when A is large. Finally, $h^4 \asymp N^{-4/5}$ and, because $s \ge 1$, $(q+B^2)/N \lesssim s_c N^{-4/5} + |S_d|N^{-1}$. Equation (S4) gives the assertion. This completes the proof.

S11 Proof of finite-grid inference

We express the estimation error as a sum of observation contributions, including changes in both the regression fit and the component centers. Block deletion estimates these same sums, and resampling them gives the joint distribution needed for simultaneous confidence intervals.

S11.1 Population derivative and variance

We work in the population local-polynomial coordinates of Section S10, now at the inference bandwidth. A reported value is indexed by a pair $k = (j, x)$ of an active group and one of its covariate values. For $k = (j, x)$, we define e_k to equal $z_j(x)$ on block j and zero elsewhere, including the intercept. We define u_i to have the joint residual-score blocks u_{ij} from Section S10 and intercept $Y_i - \mu - \sum_j g_j^*(X_{ij})$. The first-order contribution of observation i to the centered component estimate at k is

$$\psi_{i,k} = e_k^{\mathsf{T}} u_i - g_j^*(X_{ij}). \tag{S20}$$

This is the population version of (S5) because the normalized population quadratic matrix is the identity. The two terms of $\psi_{i,k}$ are the change in the regression fit and the change in the component center, the two contributions to the uncertainty about a population-centered component.

We write ψ_i for the vector of $\psi_{i,k}$ over the reported pairs. We define $d_k = \sqrt{h}$ for continuous groups and $d_k = 1$ for categorical groups, and set D_h as the diagonal matrix with these entries. This scaling puts the continuous and categorical contributions on a common variance scale. There are functions $b_h(X)$ and $v_h(X)$ such that

$$D_h\psi_i = b_h(X_i) + v_h(X_i)\varepsilon_i, \qquad \mathbb{E}b_h = \mathbb{E}v_h = 0. \tag{S21}$$

To identify these two functions, we use the identity $R_j^0 R_j^{0\mathsf{T}} = (D_j^0)^{-1} - e_j e_j^{\mathsf{T}}$.

We define $k_j(x,u) = a_j(x)^{\mathsf{T}}(D_j^0)^{-1}a_j(u)$ as the population smoothing kernel. Then $v_{h,k}(X) = d_k\{k_j(x, X_j) - 1\}$; $b_{h,k}$ is $d_k\psi_{i,k}$ after setting $\varepsilon_i = 0$ in (S20). This definition specifies b_h without a fitted random quantity. For a categorical group, $k_j(x,u) = \mathbf{1}\{x = u\}/P_j(\{x\})$.

The moment bounds for D_j^0 give $|k_j(x,u)| \le C/h$ for continuous groups, with $k_j(x,u)$ supported on $|x-u| \le 2h$. In addition, $\int k_j(x,u)\, dP_j(u) = 1$. The approximation bound (S13) implies

$$\|b_h\|_\infty \le C, \qquad \mathbb{E}\|v_h\|_2^2 \le C, \qquad \mathbb{E}\|v_h\|_2^4 \le C/h, \qquad \min_k \mathbb{E} v_{h,k}^2 \ge c. \tag{S22}$$

Here the number of active groups s and the number of reported values M are fixed. The uniform bound on b_h holds because exact component terms cancel in the score, leaving continuous smoothing terms bounded by $Csh^{3/2}$ after scaling by $\sqrt{h}$, categorical terms by Csh^2, and bounded centering terms. For the variance lower bound at a continuous value, Cauchy–Schwarz gives $\int k_j(x,u)^2 dP_j(u) \ge 1/P_j([x-2h, x+2h]) \ge c/h$. Subtracting one and multiplying by h leaves a positive bound for small h. The categorical variance is $1/P_j(\{x\}) - 1 \ge c$, because another level has probability at least c_p.

The independent-redraw transition satisfies $\mathbb{E}\{g(X_{i+r}) \mid X_i\} = \rho^r g(X_i)$ for every centered function g. Because the design and the errors are independent, the long-run covariance of (S21) is explicit:

$$\Omega_h = \frac{1+\rho}{1-\rho}\mathbb{E}(b_h b_h^\mathsf{T}) + \sigma^2 \frac{1+\rho\vartheta}{1-\rho\vartheta}\mathbb{E}(v_h v_h^\mathsf{T}). \tag{S23}$$

The diagonal entries of Ω_h are uniformly bounded above and below. No consistent estimate of ρ or ϑ is used in the procedure; this formula verifies the covariance properties needed in the proof.

For every consecutive segment I of length a, these identities also give

$$\mathbb{E}\Big\| \sum_{i\in I} D_h \psi_i \Big\|_2^2 \le Ca, \qquad \mathbb{E}\Big\| \sum_{i\in I} D_h \psi_i \Big\|_2^4 \le C(a^2 + a/h). \tag{S24}$$

Conditional Gaussian fourth moments give $\mathbb{E}(\sum_i v_{h,k}(X_i)\varepsilon_i)^4 \le 3v^2\mathbb{E}(\sum_i v_{h,k}(X_i)^2)^2$. By (S22), squared functions have variance at most C/h and covariance decreasing as ρ^r, bounding the expectation by $C(a^2 + a/h)$. Conditional on redraw lengths, independence between redraw segments bounds the fourth moment of the centered design sum by $C\{(\sum L_b^2)^2 + \sum L_b^4\}$. The moments of the geometric redraw lengths give Ca^2 after taking the expectation. Summing over the fixed M coordinates proves (S24).

S11.2 Linearization of the complete fit

We approximate the estimation error by the average of the population influences. We write $g^* = (g_j^*(x))_{(j,x)\in\mathcal{T}}$ for the population projection values in the same order as $\widehat{f}$. We write $\widehat{M}$ for the population-coordinate empirical matrix and $\bar{u} = N^{-1}\sum_i u_i$ for the final block. The normal equations give $\widehat{\eta} - \eta^* = \widehat{M}^{-1}\bar{u}$ exactly. The matrix and score bounds of Section S10.4, now with a fixed active set and $G \asymp h^{-1}$, imply

$$\|\widehat{M} - I\|_{\mathrm{op}} = O_p\{(G\log N/N)^{1/2} + G\log^2 N/N\}, \qquad \|\bar{u}\|_2 = O_p\{(G\log N/N)^{1/2}\}.$$

The extra logarithms permit simultaneous control of all block deletions. Because $d_k\|e_k\|_2 \le C$ and the empirical mean vectors satisfy the same concentration bound, inversion of the normal equation and empirical recentering give

$$\Big\| \sqrt{N} D_h(\widehat{f} - g^*) - N^{-1/2}\sum_i D_h\psi_i \Big\|_\infty = O_p\{G\log^3 N/\sqrt{N}\} = o_p(1). \tag{S25}$$

We write $\widehat{f}_k - g_j^*(x) = e_k^\mathsf{T}(\widehat{\eta} - \eta^*) - \bar{z}_j^\mathsf{T}\eta_j^* - \bar{z}_j^\mathsf{T}(\widehat{\eta}_j - \eta_j^*)$. The first two terms give (S20). The last term is the product of two factors of order $\{G\log N/N\}^{1/2}$ and satisfies (S25) after scaling. The inverse remainder is at most $2\|\widehat{M} - I\|_{\mathrm{op}}\|\bar{u}\|_2$. After scaling, both the centering and inverse remainders tend to zero for $\kappa < 1/2$. Deleted fits also require $3\kappa - 1 - \gamma < 0$ in (S27), allowing $\kappa \ge 1/3$ for sufficiently large deletion blocks.

It remains to replace the population projection by the true components. For continuous components, $\sqrt{Nh}\|g_j^* - f_j\|_\infty \le C\sqrt{Nh^5} \to 0$. For categorical components $g_j^* = f_j$ exactly. This implies that (S25) holds with g^* replaced by the vector of true component values.

S11.3 Gaussian approximation for a fixed collection of values

We derive the Gaussian limit separately for the noise and design parts. Conditional on the design, the sum of the noise terms in (S21) is exactly Gaussian. Its conditional covariance is

$$Q_N = \frac{\sigma^2}{N} \sum_{i,l=1}^{N} \vartheta^{|i-l|} v_h(X_i) v_h(X_l)^{\mathsf{T}}.$$

At lag r, overlapping segments contribute at most $r+1$ covariance terms of order $1/h$; covariances between disjoint segments decrease geometrically by the redraw transition. Thus the variance of the average of $v_{h,k}(X_i)v_{h,l}(X_{i+r})$ is at most $C(r+1)/(Nh)$. Minkowski's inequality and $\sum_{r\geq 0} |\vartheta|^r \sqrt{r+1} < \infty$ give

$$Q_N - \sigma^2 \frac{1+\rho\vartheta}{1-\rho\vartheta} \mathbb{E}(v_h v_h^{\mathsf{T}}) \longrightarrow 0 \quad \text{in probability.}$$

The lag sum is first truncated at $C \log N$. This makes the argument applicable to finite sums; the remaining part is bounded by $C|\vartheta|^{C\log N}/h$.

For the design part, the sums of $b_h(X_i)$ over completed redraw segments are independent and centered, with covariance $\mathbb{E}L_b^2 \mathbb{E}(b_h b_h^{\mathsf{T}})$ and uniformly bounded fourth moments. The number of completed segments divided by N tends to $1-\rho$. The triangular-array central limit theorem gives a Gaussian limit with covariance $(1+\rho)(1-\rho)^{-1}\mathbb{E}(b_h b_h^{\mathsf{T}})$. Because the number of completed segments is itself random, we control the sum over the segments beyond the limiting count. For segment counts within δN of the limiting count, Kolmogorov's maximal inequality bounds the squared extra sum divided by N by $C\delta$. We let $N \to \infty$ and then $\delta \downarrow 0$. The two incomplete pieces are $O_p(\log N) = o_p(\sqrt{N})$.

The conditional characteristic function $\exp(-t^{\mathsf{T}} Q_N t/2)$ of the noise sum converges to a nonrandom limit. The noise and design sums then have independent Gaussian limits, with total covariance (S23). If this covariance does not converge, we apply the argument on further subsequences on which its bounded entries converge. The lower bound on the diagonal entries of Ω_h and a union bound over univariate densities bound the probability that a Gaussian maximum lies in any interval of length t by CMt. This anti-concentration bound holds for a singular covariance and makes the subsequence approximation uniform in the threshold.

S11.4 Approximation of deleted-block effects

The change after deleting a block is approximated by the sum of its population influences. We define $U_b = \sum_{i\in b} u_i$ and $M_b = \sum_{i\in b} M_i$. We write $\bar{\psi} = N^{-1}\sum_i \psi_i$ and $\widehat{M}_{-b} = (N\widehat{M} - M_b)/(N-n_b)$ for the mean influence and deleted empirical matrix. We write $\delta = \widehat{\eta} - \eta^*$ and $\delta_{-b} = \widehat{\eta}_{-b} - \eta^*$ in population coordinates. Subtracting their normal equations gives the exact identity

$$(N-n_b)(\delta - \delta_{-b}) = \widehat{M}_{-b}^{-1}\{U_b - n_b\bar{u} - (M_b - n_b\widehat{M})\delta\}. \tag{S26}$$

The matrix Bernstein inequality conditional on redraw lengths and the conditional Gaussian score bound give, simultaneously over the J blocks,

$$\|M_b - n_b I\|_{\mathrm{op}} = O_p\{\sqrt{L_N G \log N} + G\log^2 N\},$$
$$\|U_b\|_2 = O_p\{\sqrt{L_N G \log N} + \sqrt{G}\log^2 N\}.$$

The conditional Gaussian score covariance has trace at most $CL_N G$ and operator norm at most the first matrix bound plus CL_N. Its moment-generating function gives the score bound. The design score has the same bound because s is fixed and its residual is of order h^2. These bounds also give simultaneous invertibility of the deleted matrices with probability tending to one.

We substitute the block and complete-sample bounds into (S26). The first-order centering change is $-(\sum_{i\in b} z_j(X_{ij}) - n_b\bar{z}_j)^{\mathsf{T}}\eta_j^*$. The remaining products of block mean differences with δ, or complete-sample means with block coefficient differences, obey the same bounds. Their explicit decomposition is written out in Section S12, and the same algebra applies here. This gives

$$\max_b \frac{1}{\sqrt{L_N}} \Big\| D_h V_b - \sum_{i\in b}(D_h\psi_i - D_h\bar{\psi}) \Big\|_\infty = O_p\Big\{\frac{G\log^3 N}{\sqrt{N}} + \frac{G^{3/2}\log^4 N}{\sqrt{NL_N}}\Big\} = o_p(1). \tag{S27}$$

The exponents $\kappa - 1/2$ and $(3\kappa - 1 - \gamma)/2$ are negative under the stated conditions. The covariance and resampling arguments also use $\gamma > 0$, $\gamma < 1$ and $Nh \to \infty$. Centering V_b in (S6) preserves the approximation because the first-order block sums add to zero.

S11.5 Covariance consistency and studentization

We prove covariance consistency and compare the resampled and original studentized maxima. We define $S_b = \sum_{i \in b} D_h \psi_i$. These block sums satisfy $N^{-1} \sum_b \mathbb{E}(S_b S_b^{\mathsf{T}}) = \Omega_h + O(L_N^{-1})$, because the covariance sequence in (S23) decreases geometrically. To bound fluctuations, we remove $r_N = \lceil C \log N \rceil$ observations at the beginning of each block. Equation (S24) and Cauchy–Schwarz show that this removal changes the covariance average by $O_p\{(r_N/L_N)^{1/2}\} = o_p(1)$.

Gaps of length r_N give an independent coupling of the design with error at most $N\rho^{r_N}$. After r_N transitions, the Gaussian autoregression has conditional mean $\vartheta^{r_N}\varepsilon_0$ and variance $\sigma^2(1 - \vartheta^{2r_N})$. Gaussian relative entropy and Pinsker's inequality give mean total-variation distance at most $C|\vartheta|^{r_N}$ from stationarity. The Markov property extends the coupling to each subsequent block, with total error $o(1)$. For the independent portions, the fourth-moment bound gives variance

$$CN^{-2}J(L_N^2 + L_N/h) \le C\{L_N/N + 1/(Nh)\} = o(1)$$

for each entry of the covariance average. Subtracting the common sample mean changes the covariance average by $O_p(L_N/N)$. Equation (S27) now gives

$$ND_h \widehat{\Sigma} D_h - \Omega_h \longrightarrow 0 \quad \text{in probability.} \tag{S28}$$

Conditional resampling gives J independent draws from the centered empirical distribution of the deleted-block vectors. By (S24), the averaged fourth moment of $S_b/\sqrt{L_N}$ is $O_p\{1 + 1/(L_N h)\}$. After division by J, this gives the Lyapunov quantity $O_p\{L_N/N + 1/(Nh)\} = o_p(1)$. Equation (S27) adds $o_p(1)$ to this quantity and to the covariance. The conditional multivariate central limit theorem applies; the same fourth-moment bound gives consistent resampling variances. By (S28), both studentized maxima have the same Gaussian approximation. The positive diagonal bound and continuity justify the quantile comparison.

S11.6 Completion of the inference proof

Equation (S25), its bias bound, and (S28) prove known-support coverage. By Section S4, incorrect selection has probability tending to zero. On correct selection, all fitted matrices, deletions and resampling maxima equal their known-support counterparts, giving the same unconditional limit.

For a fixed number B_* of resampling draws, the interpolated quantile lies between the order statistics of ranks r_* and $r_* + 1$. The original maximum and the B_* resampled maxima are exchangeable in the continuous Gaussian limit, giving probability $r/(B_* + 1)$ that the original maximum lies below the rth order statistic. We take $r = r_*$ and $r = r_* + 1$ to prove the bounds stated after Theorem S3. This completes the proof.

S12 Proof for increasing evaluation grids

Using independent block vectors, we approximate the distribution of the largest estimation error over all reported values, whose number is allowed to increase with the sample size. The population influence bounds verify the moment and tail conditions of Chernozhukov, Chetverikov and Kato (2017, Proposition 2.1), and coupling controls dependence between the retained blocks.

S12.1 Coordinate moments uniform in the evaluation value

On known support, write $p = |\mathcal{T}_N|$ for the number of evaluation pairs and index them by $k = 1, \ldots, p$. We use D_h, ψ_i, b_h, v_h from Section S11 and put $Z_i = D_h \psi_i$, the scaled influence vector. We write $\bar{Z} = N^{-1} \sum_i Z_i$ for its sample mean. Each coordinate splits into a design part and a part linear in the error,

$$Z_{i,k} = b_{h,k}(X_i) + v_{h,k}(X_i)\varepsilon_i, \quad \mathbb{E}b_{h,k} = \mathbb{E}v_{h,k} = 0,$$

and these two parts satisfy the uniform moment bounds

$$\max_k \|b_{h,k}\|_\infty \le C, \quad \max_k \|v_{h,k}\|_\infty \le Ch^{-1/2}, \quad \max_k \mathbb{E}v_{h,k}^2 \le C, \quad \max_k \mathbb{E}v_{h,k}^4 \le C/h.$$

These coordinate bounds do not depend on the number $|\mathcal{T}_N|$ of evaluation pairs. The fourth-moment bound and the redraw transition give, on every consecutive segment I of length a,

$$\max_k \mathbb{E}\Big|\sum_{i\in I} Z_{i,k}\Big|^2 \le Ca, \qquad \max_k \mathbb{E}\Big|\sum_{i\in I} Z_{i,k}\Big|^4 \le C(a^2 + a/h). \tag{S29}$$

The proof applies the conditional Gaussian fourth-moment calculation and the bounded-summand calculation over redraw segments in Section S11 to one coordinate at a time.

Let Ω_h be the covariance matrix in (S23) for this collection of values. Its entries satisfy

$$c \le \min_k \Omega_{h,kk} \le \max_k \Omega_{h,kk} \le C, \qquad \max_{k,l} |\operatorname{Cov}(Z_{0,k}, Z_{r,l})| \le C\rho_0^r + C(\rho_0\vartheta_0)^r.$$

Coordinate second moments give the upper bounds. For the lower bound at a continuous value, the Gaussian coefficient $\sigma^2(1+\rho\vartheta)/(1-\rho\vartheta)$ is uniformly positive and the squared kernel integral is at least c/h. For categorical values, $1/P_j(\{x\}) - 1$ is bounded below because another level has probability at least c_p. Only diagonal variances need a lower bound; the covariance may be singular.

S12.2 Tail bounds for complete blocks

We define $L = L_N$, $J = \lceil N/L \rceil$, and $\bar{L} = N/J$. Then $\bar{L}/L \to 1$ and block sizes differ from $\bar{L}$ by at most one. By (S16), uniformly over segments of size $L/2$ to $2L$, the largest redraw piece is at most $C_A \log N$ and the squared lengths sum to at most $C_A L$, for every fixed $A > 0$, except with probability $o(N^{-A})$. The Bernstein exponent $cL/\log^4 N$ permits a union bound over at most N segments.

For a segment I in this range, we condition on its redraw lengths L_b and apply the scalar Bernstein inequality to the independent summands $L_b v_{h,k}(X)^2$, one for each piece with its constant covariate value. Each summand has range at most $C_A \log N/h$, their variances sum to at most $C_A L/h$, and their means sum to at most CL. A deviation of order L has probability bounded by $2\exp\{-cLh/\log N\}$. Because $\gamma > \kappa$, Lh is a positive power of N. The bound is summable over the J blocks and the $p \le N^{A_0}$ values with any fixed polynomial probability exponent. Hence,

$$\max_{b,k} \sum_{i\in I_b} v_{h,k}(X_i)^2 \le CL \tag{S30}$$

with probability $1 - o(N^{-A})$, after changing the constant for the chosen A.

Conditional on the design, each block noise sum is Gaussian and its variance is at most $v\sum_{i\in I_b} v_{h,k}(X_i)^2$. Equation (S30) and a union bound show that the maximum of the block noise sums is of order $\sqrt{L\log N}$. For the design part $b_{h,k}$, the scalar Bernstein inequality conditional on redraw lengths gives

$$\Big|\sum_{i\in I_b} b_{h,k}(X_i)\Big| \le C_A\{\sqrt{L\log N} + \log^2 N\}$$

at the same probability level. Because L is a positive power of N, the second term is smaller than the first for sufficiently large N. With $S_b = \sum_{i\in I_b} Z_i$, we obtain

$$\Pr\Big\{\max_{b,k} |S_{b,k}|/\sqrt{\bar{L}} > C_A\sqrt{\log N}\Big\} = o(N^{-A}). \tag{S31}$$

This conclusion concerns an observed process with unbounded Gaussian responses and does not assume bounded data.

S12.3 Removing short separating segments

For the probability approximation only, we remove the first $r = \lceil C_A \log N \rceil$ observations from each block, leaving I_b° and $S_b^\circ = \sum_{i\in I_b^\circ} Z_i$. The joint process is Markov, with design coupling error ρ^r

and Gaussian mean total-variation error $C|\vartheta|^r$ as in the finite-grid proof. Independent coupling of the retained blocks has total error at most $CJ(\rho^r + |\vartheta|^r) = o(N^{-A})$ for any fixed A after increasing C_A. The covariances and moments of the individual retained blocks are unchanged by this coupling.

We write $R_b = S_b - S_b^\circ$ for a removed prefix sum. These prefixes are separated by at least $L/2$ for sufficiently large N and are themselves coupled independently with error smaller than every negative power of N. Each coordinate has variance at most Cr by (S29). On the event $\max_i |\varepsilon_i| \le C_A\sqrt{\log N}$,

$$\max_{b,k} |R_{b,k}| \le C_A r h^{-1/2}\sqrt{\log N}.$$

Gaussian tails bound the probability of the complementary event by $o(N^{-A})$ without using independence of the errors. We truncate and center the coupled prefix coordinates before applying the Bernstein inequality. Cauchy–Schwarz and the moment bound make the changes in means and covariances smaller than any prescribed negative power for large A.

The scalar Bernstein inequality and a union bound over the $|\mathcal{T}_N|$ coordinates now give

$$\Big\|N^{-1/2}\sum_b R_b\Big\|_\infty = O_p\Big\{\sqrt{\frac{r\log N}{L}} + \frac{r\log^{3/2} N}{\sqrt{Nh}}\Big\}. \tag{S32}$$

The bound multiplied by $\sqrt{\log N}$ tends to zero. This bound controls the aggregate of the removed observations, not the maximum of individual omitted responses.

S12.4 Independent-vector Gaussian approximation

For the independent retained copies, we define $W_b = S_b^\circ/\sqrt{L}$. Their covariance average is

$$\overline{\Omega} = J^{-1}\sum_b \mathbb{E}(W_b W_b^\mathsf{T}) = N^{-1}\sum_b \mathbb{E}(S_b^\circ S_b^{\circ\mathsf{T}}).$$

The geometric covariance bound implies

$$\|\overline{\Omega} - \Omega_h\|_{\max} \le Cr/L, \tag{S33}$$

where $\|A\|_{\max}$ is the largest absolute matrix entry. To see this, we observe that the sum over a stationary block of length a has covariance

$$a\Omega_h - \sum_{|k|<a} |k|\,\mathrm{Cov}(Z_0, Z_k) - a\sum_{|k|\ge a}\mathrm{Cov}(Z_0, Z_k).$$

The last two matrices have uniformly bounded entries. Summing over the J retained blocks and using $\sum_b |I_b^\circ| = N - Jr$ gives (S33).

By (S29) and $Lh \to \infty$, W_b has bounded coordinate fourth moments. Equation (S31) bounds its coordinate maximum by $C_A\sqrt{\log N}$ except with probability $o(N^{-A})$. For the proof, we truncate and center each independent W_b at this level to obtain $\widetilde{W}_b$.

For a sufficiently large fixed A, truncation changes the normalized sum by $o_p(N^{-1})$ and changes every covariance entry by $o(N^{-1})$. For example,

$$|\mathbb{E}(W_{b,k} - \widetilde{W}_{b,k})| \le C\Pr\{|W_{b,k}| > C_A\sqrt{\log N}\}^{1/2} = o(N^{-A/2})$$

Here Cauchy–Schwarz bounds the mean of the discarded tail, and the fourth moment bounds its covariance. The sum over the $J \le N$ blocks remains negligible for large A.

The variables $\widetilde{W}_b$ are independent, centered, have average coordinate variances bounded below, and satisfy

$$J^{-1}\sum_b \mathbb{E}|\widetilde{W}_{b,k}|^{2+l} \le B_J^l \quad (l = 1, 2), \qquad \mathbb{E}\exp(|\widetilde{W}_{b,k}|/B_J) \le 2, \qquad B_J = C\log N.$$

The fourth-moment and truncation bounds give these inequalities after increasing C. These inequalities verify Conditions M.1, M.2 and E.1 of Chernozhukov, Chetverikov and Kato (2017, Proposition

2.1) for sample size J and dimension p. Applying the proposition to the vector and its negative gives Gaussian approximation error at most

$$C\left\{\frac{\log^9 N}{J}\right\}^{1/6} = o(1).$$

If there are fewer than three coordinates, we duplicate a coordinate to meet the dimension convention of that proposition without changing the maximum or its variance.

The Gaussian comparison bound of Chernozhukov, Chetverikov and Kato (2015, Theorem 2) and (S33) replace the Gaussian covariance by Ω_h, with error at most

$$C(r/L)^{1/3}\{1+\log(pL/r)\}^{2/3} = o(1).$$

We write $(G_1,\dots,G_p)^\mathsf{T} \sim N(0,\Omega_h)$ for the comparison vector. All coordinate variances are bounded above and below. For any $\eta > 0$, Theorem 3(i) of the same reference gives

$$\sup_t \Pr\left(\left|\max_k |G_k|/\Omega_{h,kk}^{1/2} - t\right| \le \eta\right) \le C\eta\sqrt{1+\log(2p)}. \tag{S34}$$

This inequality and (S32) restore the prefixes with negligible coupling error, giving a uniform Gaussian approximation for the standardized maximum of $N^{-1/2}\sum_i Z_i$.

S12.5 A UNIFORM APPROXIMATION FOR EXACT BLOCK DELETIONS

The algebraic deletion calculation must also hold uniformly over the evaluation grid. We use population coordinates of dimension $q = O(G)$ on the fixed support. We define $\delta = \widehat{\eta} - \eta^*$, $\bar{u} = N^{-1}\sum_i u_i$, and $\bar{z}_j = N^{-1}\sum_i z_j(X_{ij})$. We also define

$$a_N = \sqrt{G\log N/N} + G\log^2 N/N, \qquad d_N = \sqrt{G\log N/N}.$$

The matrix and score bounds in Sections S10 and S11 give $\|\widehat{M} - I\|_{\mathrm{op}} = O_p(a_N)$ and $\|\delta\|_2 = O_p(d_N)$. Simultaneously over the J blocks,

$$\|M_b - n_b I\|_{\mathrm{op}} = O_p\{\sqrt{LG\log N} + G\log^2 N\},$$
$$\|U_b\|_2 + \Big\|\sum_{i\in I_b} z_j(X_{ij})\Big\|_2 = O_p\{\sqrt{LG\log N} + \sqrt{G}\log^2 N\}.$$

The noise-score covariance has trace at most CLG because feature squared norms are at most CG. Gaussian concentration and a union bound give the noise-score bound $C\sqrt{LG\log N}$. The Bernstein inequality applied to centered redraw pieces bounds the approximation score, without a rank lower bound for individual deleted blocks.

Subtracting the block moments from the complete-sample moments yields

$$\max_b \|\widehat{M}_{-b} - I\|_{\mathrm{op}} = O_p(a_N), \qquad \max_b \|\delta_{-b}\|_2 = O_p(d_N).$$

Here $L = o(N)$, and the added block errors divided by $N - n_b$ are bounded by the stated complete-sample orders. Because $a_N \to 0$, all deleted matrices are invertible with probability tending to one.

For $k = (j,x)$, we write e_k for its population evaluation vector, including zeros on the other coefficient blocks. We write $Z^z_{b,j} = \sum_{i\in I_b} z_j(X_{ij})$. Direct subtraction of the centered reported values gives

$$V_{b,k} = e_k^\mathsf{T}(N-n_b)(\delta - \delta_{-b}) - \bar{z}_j^\mathsf{T}(N-n_b)(\delta_j - \delta_{j,-b}) - (Z^z_{b,j} - n_b\bar{z}_j)^\mathsf{T}(\eta_j^* + \delta_{j,-b}).$$

Equation (S26) supplies the coefficient difference. Its first-order part is

$$e_k^\mathsf{T}(U_b - n_b\bar{u}) - (Z^z_{b,j} - n_b\bar{z}_j)^\mathsf{T}\eta_j^* = \sum_{i\in I_b}(\psi_{i,k} - \bar{\psi}_k).$$

The remainders consist of inverse errors times block scores, block matrix errors times δ, and two centering products. Because $d_k\|e_k\|_2 \le C$ uniformly, the bounds on the coefficient vectors control

every evaluation value at once. After division by $\sqrt{L}$, the first two remainders are bounded by constant multiples of

$$a_N \left\{ \sqrt{G \log N} + \frac{\sqrt{G} \log^2 N}{\sqrt{L}} + \sqrt{L}\, d_N \right\} + d_N \left\{ \sqrt{G \log N} + \frac{G \log^2 N}{\sqrt{L}} + \sqrt{L}\, a_N \right\}.$$

The centering products have the same bound because $\|\bar{z}_j\|_2 = O_p(d_N)$ and the complete and deleted coefficient errors are $O_p(d_N)$. This gives

$$\max_{b,k} \frac{|d_k V_{b,k} - \sum_{i \in I_b} (Z_{i,k} - \bar{Z}_k)|}{\sqrt{\bar{L}}} = O_p \left\{ \frac{G \log^3 N}{\sqrt{N}} + \frac{G^{3/2} \log^4 N}{\sqrt{NL}} \right\} =: O_p(e_N). \tag{S35}$$

The two polynomial exponents are negative when $\kappa < 1/2$ and $\gamma > \kappa$. In particular, e_N times any fixed power of $\log N$ tends to zero. Recentering V_b with the actual n_b/N weights preserves this conclusion.

S12.6 Uniform covariance consistency

We independently couple and truncate the trimmed versions of $S_b/\sqrt{\bar{L}}$ as in the Gaussian-approximation argument. Coordinate products are bounded by $C \log N$; the Bernstein inequality and a union bound over p^2 entries bound the covariance fluctuation by $O_p(\log^2 N/\sqrt{J})$. Truncation changes neither this bound nor the population covariance at this order.

On the Gaussian maximum event used for the prefixes,

$$\max_{b,k} |S_{b,k} - S^{\circ}_{b,k}|/\sqrt{\bar{L}} \le C_A r \sqrt{\log N/(Lh)}.$$

Using (S31) in the difference of two products shows that restoring the prefixes changes every empirical covariance entry by

$$O_p\{r \log N/\sqrt{Lh} + r^2 \log N/(Lh)\}.$$

The covariance bias from trimming is $O(r/L)$ by (S33). The established Gaussian approximation also gives $\|\bar{Z}\|_\infty = O_p\{\sqrt{\log N/N}\}$. Subtracting the weighted global mean from every block changes its normalized vector by $O_p\{\sqrt{\log N/J}\}$, and hence changes every covariance entry by at most $O_p(\log N/\sqrt{J})$.

Finally, (S35) changes these covariance entries by $O_p(e_N \sqrt{\log N} + e_N^2)$. Combining these bounds gives

$$\|N D_h \widehat{\Sigma} D_h - \Omega_h\|_{\max} = O_p \left[\log^6 N \left\{ J^{-1/2} + (Lh)^{-1/2} + \frac{G}{\sqrt{N}} + \frac{G^{3/2}}{\sqrt{NL}} \right\} \right]. \tag{S36}$$

Every term tends to zero after multiplication by any fixed power of $\log N$, giving the covariance accuracy required for the growing maximum.

S12.7 Conditional resampling and estimated denominators

We write O_{P^*} for order in the conditional resampling probability $\Pr^*$, with the bound holding in probability over the original data. We normalize the centered deleted-block vectors by $D_h/\sqrt{\bar{L}}$. Their empirical mean is zero. By (S31) and (S35), their largest coordinate is $O_p(\sqrt{\log N})$. By (S36), their empirical coordinate variances are bounded above and below with probability tending to one.

Conditional on the data, the resampled vectors are independent. Their coordinate maximum and bounded second moments give third moments at most $C\sqrt{\log N}$ and fourth moments at most $C \log N$. Condition E.1 of Chernozhukov, Chetverikov and Kato (2017, Proposition 2.1) holds with $B_J = C \log N$. The conditional Gaussian approximation error is at most $C(\log^9 N/J)^{1/6}$ when each coordinate is divided by its conditional standard deviation.

For the standard deviations computed within each draw, we apply a scalar concentration inequality to the coordinates and their squares, which are bounded by $C \log N$. Uniformly over p coordinates, the resampled variance differs from its conditional expectation by $O_{P^*}(\log^3 N/\sqrt{J})$ in probability

over the original data. Empirical variances are bounded below, and the Gaussian approximation bounds the numerator maximum by $O_{P^*}(\sqrt{\log N})$. This implies that the change in the denominator is negligible relative to $1/\sqrt{\log N}$.

By Chernozhukov, Chetverikov and Kato (2015, Theorem 2) and (S36), we replace the empirical Gaussian covariance by the standardized Ω_h. Equation (S34) makes both denominator replacements negligible, giving agreement with the original distribution uniformly in the threshold.

S12.8 Estimation error, smoothing bias and selection

The normal equation and centering identity in (S25), with $d_k\|e_k\|_2 \le C$, give a uniform remainder $O_p(G\log^3 N/\sqrt{N})$ without a factor p. The scaled bias is at most $C\sqrt{Nh^5}$ for continuous components and zero for categorical components. Both errors times $\sqrt{\log N}$ tend to zero for $1/5 < \kappa < 1/2$.

Equation (S36) controls the estimated standard errors, extending the Gaussian comparison to centering at the true values. By Section S4, incorrect support selection has probability smaller than every fixed negative power of N. On correct selection, the grid, deletions and resampling maxima equal their known-support counterparts. This proves Theorem 3.

For finitely many independent resampling draws, the joint conditional approximation and the order-statistic argument after Theorem S3 give the same bounds for the interpolated quantile. A fixed limiting distribution of the Gaussian maxima is not required.

S12.9 Proof of Corollary S3

In this proof L denotes the curvature bound of Corollary S3, and the block length is written L_N. On simultaneous coverage of all grid values, the deterministic curvature inequality in Section S6 gives coverage between every adjacent pair of grid values. Categorical intervals remain unchanged. This proves the joint lower coverage bound from Theorem 3.

Uniform covariance consistency gives $\max_k \widehat{\Sigma}_{kk}^{1/2} = O_p\{(Nh)^{-1/2}\}$ for the continuous coordinates. A Gaussian union bound gives a joint critical value of order $O_p(\sqrt{\log N})$ because p grows at most polynomially. Linear interpolation does not increase the largest grid half-width, while its curvature correction adds at most $L\Delta_N^2/8$ to a half-width. The stated width bound follows. If $\Delta_N \le h$, then $L\Delta_N^2\sqrt{Nh} \le L\sqrt{Nh^5} \to 0$, and the interpolation correction is smaller than a pointwise standard error. This completes the proof.

S13 Proofs for correlated predictors

We derive the population coefficient of the inactive variable and show that its SCAD weight removes this variable. The inference proof separates local smoothing from finite-dimensional dependence, using matrix and score bounds from the stationary process.

S13.1 Transition probabilities and moment concentration

Independent redraws split the correlated process into independent segments. We define $\Pi_Q(x,\cdot) = Q(\cdot)$. Because $QK = Q$, one has $K\Pi_Q = \Pi_Q K = \Pi_Q$ and

$$P^r = \{1-\rho^r\}\Pi_Q + \rho^r K^r, \qquad \|P^r g\|_{L^2(Q)} \le \rho^r \|g\|_{L^2(Q)}, \qquad \mathbb{E}_Q g = 0.$$

The conditional Jensen inequality gives the norm bound for K. Independent stationary coupling after r transitions has failure probability at most ρ^r. Each segment, conditional on its geometric length L of parameter $1-\rho$, is a stationary K path starting in Q. Completed paths are independent.

For a centered matrix function F, we write $R_L = \sum_{i=1}^{L} F(X_i)$. If $\|F\|_{\mathrm{op}} \le B_F$ and both matrix second moments are bounded by V_F, the conditional Jensen inequality gives

$$\mathbb{E}(R_L \mid L) = 0, \qquad \|R_L\|_{\mathrm{op}} \le LB_F, \qquad \|\mathbb{E}(R_L R_L^{\mathsf{T}} \mid L)\|_{\mathrm{op}} \le L^2 V_F.$$

The same bound holds for $R_L^{\mathsf{T}} R_L$, and (S16) remains valid. Conditional on the lengths, the matrix Bernstein inequality for independent sums over redraw segments has variance at most CnV_F and

summand norm at most $C_A B_F \log n$. A union bound over the terminal indices adds a logarithmic factor. The two boundary paths have length at most $C_A \log n$ and obey the same summand bound.

For the population coordinates z_j, A_j defined in Section S10, bounded pairwise densities imply

$$\|z_j z_k^\mathsf{T}\|_{\mathrm{op}} \le CG, \quad \max\{\|\mathbb{E}\|z_k\|_2^2 z_j z_j^\mathsf{T}\|_{\mathrm{op}}, \|\mathbb{E}\|z_j\|_2^2 z_k z_k^\mathsf{T}\|_{\mathrm{op}}\} \le CG.$$

The same orders hold for $A_j - I$ and for centered cross blocks. Thus, outside a set of probability $o(n^{-A})$,

$$\begin{aligned}&\max_j \|\bar{A}_j - I\|_{\mathrm{op}} + \max_j \|\bar{z}_j\|_2 + \max_{j\neq k} \|\overline{z_j z_k^\mathsf{T}} - \mathbb{E}(z_j z_k^\mathsf{T})\|_{\mathrm{op}} \\ &\le C_A \left\{ \sqrt{\frac{G\log(dn)}{n}} + \frac{G \log n \log(dn)}{n} \right\} = a_n. \end{aligned} \tag{S37}$$

All pairwise densities in the selection model are bounded, including binary pairs. The bound holds over all groups before screening, and the rate conditions of Theorem S4 give $ma_n = o(1)$.

When the active set is fixed, its total regression function is bounded. For a unit vector t, the centered marginal response score $t^\mathsf{T} z_j(X_j)\{m(X) - \mu\}$ has variance bounded by a constant and absolute value at most $C\sqrt{G}$. Bernstein's inequality for its sums over redraw segments, followed by a $1/2$-net of the sphere and a union bound over the groups, bounds the marginal design-score error by

$$C_A \left\{ \sqrt{\frac{G + \log(dn)}{n}} + \frac{\sqrt{G}\log n\{G + \log(dn)\}}{n} \right\} = o(1).$$

By (S37), the Gaussian part of each marginal score has conditional covariance bounded by C/n and norm bounded by $C_A\sqrt{\{G + \log(dn)\}/n}$ at the same probability level. Thus the marginal scores converge uniformly.

The joint projection score has the same Gaussian bound. The uniform residual $O(h_L^2)$ gives a centered design score of norm at most $Ch_L^2\sqrt{G}$. The rectangular Bernstein inequality for sums over redraw segments gives

$$\max_j \|\bar{u}_j\|_2 \le C_A \sqrt{\frac{G + \log(dn)}{n}} + C_A h_L^2 \left\{ \sqrt{\frac{G\log(dn)}{n}} + \frac{\sqrt{G}\log n \log(dn)}{n} \right\} = o(\lambda_n^0). \tag{S38}$$

The equality uses $\log d \to \infty$, $\log(dn) = o(G)$ and $h_L \asymp n^{-1/5}$. The Bernstein bound for sums over redraw segments controls the response design moments; bounded covariance eigenvalues control the Gaussian quadratic term, and the cross term is conditionally Gaussian. The mean and the variance converge at fixed tolerances with errors $o(n^{-A})$, and hence the ratio of the penalty to λ_n^0 tends to one.

S13.2 Population equations for the inactive variable

We compute the population group-Lasso coefficient of the inactive binary variable.

We write $t = 2x - 1$, $u(x) = \sqrt{3}t$, and $w(x) = \sqrt{7}(5t^3 - 3t)/2$. Both functions are centered and have unit norm under the uniform marginal, and $\int uw = 0$. Summing (S9) over z gives the joint density $1 + 0.1u(x_1)u(x_2)$ of the first two covariates. Integration on either side of $x_2 = 1 - x_1$ gives

$$\mathbb{E}(Z \mid X_j = x) = 0.99\{t + 0.15t(1 - t^2)\} = r_\star u(x) + c_\star w(x), \qquad j = 1, 2.$$

Here $r_\star = 0.99(1.06)/\sqrt{3}$ and $c_\star = -0.99(0.3)/(5\sqrt{7})$. The same integration verifies the uniform marginals and the symmetric Bernoulli marginal. The density lies between 0.007 and 2.587.

We choose the same population coordinates for the two continuous groups. We define a_h and w_h as the coefficient vectors of the local-polynomial projections of u and w. Affine reproduction and orthogonal projection give

$$\|a_h\|_2 = 1, \qquad a_h^\mathsf{T} w_h = 0, \qquad \|w_h\|_2 \le 1.$$

Indeed, the local representation of $u(X)$ equals $u(X)$ for every auxiliary node. The inner product of the coefficient vector of this representation and the projection of w equals $\mathbb{E}\{u(X)w(X)\} = 0$. We

choose the centered binary coordinate to be Z itself. On these three groups, the population quadratic matrix is

$$\Gamma_h = \begin{pmatrix} I & 0.1a_h a_h^\top & r_\star a_h + c_\star w_h \\ 0.1a_h a_h^\top & I & r_\star a_h + c_\star w_h \\ (r_\star a_h + c_\star w_h)^\top & (r_\star a_h + c_\star w_h)^\top & 1 \end{pmatrix}. \tag{S39}$$

Other groups have identity diagonal blocks and zero population cross blocks. The off-diagonal block norms in (S39) are bounded by the entries of $C_0 = \begin{pmatrix} 0 & 0.1 & 0.62 \\ 0.1 & 0 & 0.62 \\ 0.62 & 0.62 & 0 \end{pmatrix}$. Its largest eigenvalue is less than 0.93, and hence the full candidate population matrix has eigenvalues between 0.07 and 1.93, independently of G and m.

The first two coefficient blocks equal $b_0 a_h$ and represent the functions of these two groups exactly. We use separate population projections for the other active groups. Their residuals are centered, independent of these three groups, and uniformly $O(h_L^2)$. Thus the normal-equation score equals the population matrix times this coefficient vector, which proves the properties in (S38).

We define

$$t_h = \frac{2r_\star/1.1 - 1}{1 - 2r_\star^2/1.1 - 2c_\star^2\|w_h\|_2^2}, \qquad 0.30 < t_h < 0.31. \tag{S40}$$

The inequalities follow by substituting $0 \le \|w_h\|_2^2 \le 1$; the denominator is between 0.3315 and 0.3327 and the numerator is positive. We define a perturbation d_h on the first three groups by

$$d_{h,3} = t_h, \qquad d_{h,1} = d_{h,2} = -\frac{1 + r_\star t_h}{1.1} a_h - c_\star t_h w_h.$$

For each of the other active groups we set $d_{h,j} = -\eta_j^*/\|\eta_j^*\|_2$, and set the perturbation to zero on the remaining groups. Direct multiplication in (S39) gives

$$(\Gamma_h d_h)_j = -\eta_j^*/\|\eta_j^*\|_2 \quad (j \in S), \qquad (\Gamma_h d_h)_3 = -1.$$

The vector $\eta^* + \lambda d_h$ satisfies the restricted group-Lasso equations on $S \cup \{3\}$ up to $O(\lambda^2)$. The active norms are bounded below, which makes the norm gradients Lipschitz; only their unit directions contribute error. Strong convexity gives distance $O(\lambda^2)$ to the restricted minimizer, whose binary coefficient equals $\lambda t_h + O(\lambda^2)$ and is positive and below λ. The eigenvalue bounds and strict inequalities in (S40) are uniform over $0 \le \|w_h\|_2^2 \le 1$. By continuity, these bounds, the density bounds, and the contraction bound below remain valid in a sufficiently small fixed neighborhood of the density parameters 0.1 and 0.99. Selection holds throughout this neighborhood.

S13.3 Proof of Theorem S4

The concentration bounds transfer the population selection calculation to the empirical criterion. Screening retains the groups with the largest marginal scores.

The population marginal score norms on the first two groups equal $1.1b_0$, and the binary score norm is $2b_0 r_\star$. Other active norms are bounded below by their component norms minus Ch_L^2; all other population scores equal zero. The preceding marginal concentration bound proves that screening retains $S \cup \{3\}$ when $m \ge s + 1$, outside a set of probability $o(n^{-A})$.

We use the simultaneous matrix and score bounds with an arbitrary fixed probability exponent. Orthogonal group rotations preserve the penalties and the fitted functions, which allows the choice $T_j = (\bar{A}_j - \bar{z}_j \bar{z}_j^\top)^{-1/2}$ with $\|T_j - I\|_{\mathrm{op}} = O(a_n)$. Empirical centering transfers constants to the intercept. We write H, b for the screened quadratic matrix and score, $\bar{u}_j$ for the empirical mean of u_{ij} from Section S10.4, and $\widehat{\theta}^{L,R}$ for the Lasso solution restricted to $S \cup \{3\}$. For the recentered population vector θ^*, we define $e = b - H\theta^*$. Equations (S37) and (S38) imply

$$\max_j \|e_j\|_2 = o(\lambda), \qquad \max_j \sum_k \|H_{jk} - (\Gamma_h)_{jk}\|_{\mathrm{op}} = o(1), \qquad \lambda/\lambda_n^0 \longrightarrow 1.$$

The constant in the middle bound does not depend on the screened set. The block of the population matrix Γ_h outside the first three groups is the identity.

We restrict the Lasso optimization to $S \cup \{3\}$. By the three preceding bounds, the $O(\lambda^2)$ population calculation and $T_j = I + O(a_n)$, the candidate $\theta^* + \lambda d_h$ has active score error $o(\lambda)$ on the $s + 1$

groups of this set. Strong convexity and monotonicity of the penalty subgradient give

$$\max_{j\in S\cup\{3\}} \|\widehat{\theta}_j^{L,R} - \theta_j^* - \lambda d_{h,j}\|_2 = o(\lambda).$$

For a group outside this set, the residual score at the restricted solution is bounded by

$$\|e_j\|_2 + \sum_{k\in S\cup\{3\}} \|H_{jk}\|_{\mathrm{op}} \|\widehat{\theta}_k^{L,R} - \theta_k^*\|_2 = o(\lambda) < \lambda.$$

These strict inequalities for the inactive groups give the unique full Lasso solution. By (S40), the support of this solution is $S \cup \{3\}$, with active norms above $a\lambda$ and binary norm below λ. Thus the SCAD weights are zero on S and equal λ on S^c.

The restricted unpenalized solution has coefficient error $H_{SS}^{-1} e_S = o(\lambda)$ in maximum group norm. For every inactive group, the residual score at this solution is

$$e_j - H_{jS} H_{SS}^{-1} e_S = o(\lambda).$$

This score has norm strictly less than λ even for the correlated binary group, which proves the one-step SCAD assertion by the convex KKT equations.

It remains to verify that the implemented coordinate descent reaches these two solutions within its sweep limit and its numerical tolerances. For the first three groups, coordinate sweeps are bounded by $(I - C_{0,L})^{-1} C_{0,U}$. Here $C_{0,L}, C_{0,U}$ are the strict lower and upper triangular parts of C_0. This bounding matrix has spectral radius below 0.862, which gives contraction by 0.865 in a positive weighted maximum group norm. The remaining cross blocks have population value zero and empirical row sums $o(1)$. The complete sweep thus eventually contracts by less than 0.87, with fixed positive weight ratios. Because $30|\log(0.87)| > 4$, $100 + \lceil 30 \log N_L \rceil$ sweeps give error $o(n^{-4})$. The inactive dual gaps are a fixed fraction of λ, which gives exactly zero inactive updates. The tolerance N_L^{-3} is $o(\lambda)$ and $o(b_0)$; eigenvalue cutoffs lie below the population bounds, including the raw order G^{-1}. All exceptional probabilities are $o(n^{-A})$. This completes the proof.

S13.4 Population approximation with finite-rank pairwise dependence

Under correlated covariates, fitting each component separately no longer solves the joint population equation. We show that the pairwise dependence representation keeps the joint matrix invertible and that solving this equation preserves the local-linear approximation order.

We define $\mathcal{A}_j$ to map a population local-polynomial coefficient vector to its averaged component in $L^2(P_j)$, and define $\mathcal{A}_j^*$ as its Hilbert-space adjoint. The conditional Jensen inequality gives $\|\mathcal{A}_j\|_{\mathrm{op}} \le 1$. We define W_j to have columns $\mathcal{A}_j^* \chi_{j,l}$, and set W as block diagonal with blocks W_j. We define $\mathcal{C} = \mathcal{R} - I$. The population quadratic matrix on the nonconstant coordinates is

$$\Gamma_h = I + W\mathcal{C}W^{\mathsf{T}} = I - WW^{\mathsf{T}} + W\mathcal{R}W^{\mathsf{T}}. \tag{S41}$$

Because the functions in each χ_j are orthonormal, $W_j^{\mathsf{T}} W_j \preceq I$ and $W_j W_j^{\mathsf{T}} \preceq I$. Thus $c_1 I \preceq \Gamma_h \preceq C_1 I$ for fixed positive constants. An intercept adds an independent population block equal to one.

We write $\eta_j^0 = \mathcal{A}_j^* f_j$ and $g_j^0 = \mathcal{A}_j \eta_j^0$ for the coefficients and averaged curve of the separate projection of f_j, and define Π_j to denote the local-polynomial orthogonal projection on the augmented space of (U_j, X_j). For a continuous group the local moment calculation in (S13) applies to both f_j and every dependence function $\chi_{j,l}$. For a categorical group the projection is exact. Thus,

$$\begin{aligned}\left| \int \chi_{j,l}(f_j - g_j^0)\, dP_j \right| &= |\mathbb{E}\{(\chi_{j,l} - \Pi_j \chi_{j,l})(f_j - \Pi_j f_j)\}| \\ &\le \|\chi_{j,l} - \Pi_j \chi_{j,l}\|_2 \|f_j - \Pi_j f_j\|_2 \le C h^4. \end{aligned} \tag{S42}$$

The equality uses orthogonality of $f_j - \Pi_j f_j$ to the projection space.

We collect the integrals in (S42) in a finite vector a. The full population score satisfies $b - \Gamma_h \eta^0 = W\mathcal{C}a$. The correct population coefficients obey

$$\eta^* - \eta^0 = \Gamma_h^{-1} W\mathcal{C}a = W\{I + \mathcal{C}W^{\mathsf{T}}W\}^{-1}\mathcal{C}a.$$

Projection error bounds give $W^{\mathsf{T}}W = I + O(h^4)$ on continuous dependence functions, with exact equality on categorical contrasts. This implies that the finite-dimensional inverse in this expression has bounded operator norm. The averaged and nodewise projections of every continuous $\chi_{j,l}$ are uniformly bounded. It follows that this correction of the population coefficients has coefficient norm and uniformly evaluated local-polynomial norm $O(h^4)$. In particular,

$$\|g_j^* - f_j\|_\infty \le Ch^2 \quad (j \text{ continuous}), \qquad \|g_j^* - f_j\|_\infty \le Ch^4 \quad (j \text{ categorical}). \tag{S43}$$

The complete local-polynomial residual is uniformly $O(h^2)$, and its population normal-equation score equals zero by projection.

The eigenvalue condition on the population parameter also identifies additive components. For centered g_j, we define $a_j = \mathbb{E}(\chi_j g_j)$ and decompose $g_j = a_j^{\mathsf{T}}\chi_j + r_j$ with $\mathbb{E}(\chi_j r_j) = 0$, and write a for the stacked vector of the a_j. Equation (S10) gives

$$\mathbb{E}_Q\Big(\sum_j g_j(X_j)\Big)^2 = a^{\mathsf{T}}\mathcal{R}a + \sum_j \|r_j\|_{P_j}^2 \ge c_1 \sum_j \|g_j\|_{P_j}^2.$$

This proves the population coercivity used in the risk calculation.

S13.5 Component-risk bound under correlation

We normalize the coordinates by $\Gamma_h^{-1/2}$, using the matrix in (S41). The normalized one-observation matrix has norm at most Cq and population mean I. The residue-class coupling following (S8) and the matrix Chernoff inequality give

$$\Pr\{\|\Gamma_h^{-1/2}\widehat{M}\Gamma_h^{-1/2} - I\|_{\mathrm{op}} > 1/2\} \le 2\ell q \exp\{-cN/(q\ell)\} + N\rho^\ell.$$

The bounded population inverse then gives the coefficient-error bound of Theorem S1: the conditional Cauchy–Schwarz inequality bounds the approximation part, and the bound vI_N on the error covariance bounds the noise part. The squared projection residual has expectation at most $Cs_cL^2h^4$, although the residuals of two different groups are no longer uncorrelated: for $j \ne k$, (S10) and (S42) give $|\mathbb{E}(r_{ij}r_{ik})| \le Ch^8$, and $\eta^* - \eta^0$ contributes $O(h^8)$. Also, $\sum_j \|g_j^*\|_{P_j}^2 \le CB^2$ follows from $\|\eta^*\|_2 \le \|\Gamma_h^{-1}\|_{\mathrm{op}}\|b\|_2$ and bounded pairwise densities. Additive coercivity transfers prediction risk to component loss.

For each centered population component g_j^*,

$$|\operatorname{Cov}\{g_j^*(X_{0j}), g_j^*(X_{rj})\}| \le \rho^r \|g_j^*\|_{P_j}^2, \qquad \sum_j \|g_j^*\|_{P_j}^2 \le CB^2.$$

The empirical center contributes at most CB^2/N. Products of empirical means and coefficient errors obey the matrix bound in Theorem S1. On the event of incorrect selection or of matrix instability, truncation bounds the loss by $4(m+s)T_N^2$, which proves (S4) with the transition parameter of (S8).

S13.6 Influence and long-run covariance under correlation

We derive influence moments and long-run covariances under coordinate dependence. In this part $h = h_I$. We use population coordinates on the true active set, including the intercept. The objects e_k, u_i, g_j^* retain their roles from Section S11. The population influence is now

$$\psi_{i,k} = e_k^{\mathsf{T}}\Gamma_h^{-1}u_i - g_j^*(X_{ij}). \tag{S44}$$

The centering term in (S44) is unchanged by dependence between coordinates. With the earlier scaling matrix D_h, we write $Z_i = D_h\psi_i = b_h(X_i) + v_h(X_i)\varepsilon_i$. Both functions have mean zero. The identity

$$\Gamma_h^{-1} = I - W\{I + \mathcal{C}W^{\mathsf{T}}W\}^{-1}\mathcal{C}W^{\mathsf{T}}$$

expresses the evaluation row of the inverse as its local row plus a fixed finite number of bounded smooth rows. For a continuous evaluation value the correction is multiplied by $\sqrt{h}$; for a categorical evaluation value it is bounded. The bounds in (S43) and the local kernel bounds give, uniformly in the reported value,

$$\|b_{h,k}\|_\infty \le C, \quad \|v_{h,k}\|_\infty \le Ch^{-1/2}, \quad \mathbb{E}v_{h,k}^2 \le C, \quad \mathbb{E}v_{h,k}^4 \le C/h. \tag{S45}$$

The design residual is uniformly $O(h^2)$. The local kernel of size C/h gives $O(h)$ before continuous scaling; finite-rank corrections have the same or smaller order. Centering remains bounded.

For a continuous group the local noise row has squared norm at least c after scaling and its finite-rank correction is $O(\sqrt{h})$ in $L^2(Q)$. This gives $\mathbb{E}v_{h,k}^2 \geq c$ for all sufficiently small h. For a categorical group we define $a = \Gamma_h^{-1} e_k$ and $v(X) = \mathbb{E}\{\phi(U, X) \mid X\}^\mathsf{T} a$. Because this group has no auxiliary-node variance, its block in $\Gamma_h a = e_k$ is $\mathbb{E}\{z_j(X_j)v(X)\} = e_{k,j}$. Because $\mathbb{E}z_j z_j^\mathsf{T} = I$, the Cauchy–Schwarz inequality gives $\mathbb{E}v^2 \geq \|e_{k,j}\|_2^2 = 1/P_j(\{x\}) - 1 \geq c$.

We define the stationary long-run covariance

$$\Omega_h = \sum_{r \in \mathbb{Z}} \mathrm{Cov}(Z_0, Z_r). \tag{S46}$$

The L^2 contraction of P makes the design covariances decrease geometrically in the lag r, and the independence of the errors from the design contributes the factor $\vartheta^{|r|}$ to the noise covariances. The series converges absolutely entrywise and uniformly, and the cross covariance of the two terms is zero. The design part of Ω_h is positive semidefinite. For the noise part, Gaussian autoregressive covariances satisfy

$$\sigma^2 \frac{1 - |\vartheta|}{1 + |\vartheta|} I \preceq \mathrm{Cov}(\varepsilon_1, \ldots, \varepsilon_N) \preceq vI.$$

Conditioning on the design and then taking expectations shows $c \leq \Omega_{h,kk} \leq C$ uniformly in k and h.

S13.7 Block approximations and completion of the proof of Theorem S5

The Gaussian approximation requires moment and tail bounds for block sums. For a segment I of length a, conditional Gaussian fourth moments give

$$\mathbb{E}\Big(\sum_{i \in I} v_{h,k}(X_i)\varepsilon_i\Big)^4 \leq C\mathbb{E}\Big(\sum_{i \in I} v_{h,k}(X_i)^2\Big)^2 \leq C(a^2 + a/h).$$

The last bound uses the L^2 contraction for the centered squared functions and (S45). For the bounded design sum, independent centered sums over redraw segments have bounded fourth moments because their lengths have geometric tails. Truncation at each segment boundary changes at most two such rewards. Thus

$$\max_k \mathbb{E}\Big|\sum_{i \in I} Z_{i,k}\Big|^2 \leq Ca, \qquad \max_k \mathbb{E}\Big|\sum_{i \in I} Z_{i,k}\Big|^4 \leq C(a^2 + a/h).$$

The covariance of two coordinates at lag r is bounded by $C\rho^r + C(\rho|\vartheta|)^r$.

Here $L = L_N$ denotes block length, not curvature, and $J = \lceil N/L \rceil$. Bernstein's inequality for centered sums of $v_{h,k}^2$ over redraw segments has variance at most CL/h and largest summand at most $C \log N/h$ under the length bounds. A deviation of order L has probability at most $2\exp\{-cLh/\log N\}$. Because $\gamma > \kappa$, this probability is summable over the blocks and over the polynomial number of evaluation values. The conditional Gaussian bound and the bounded design rewards then give

$$\Pr\left\{\max_{b,k} \left|\sum_{i \in I_b} Z_{i,k}\right| / \sqrt{L} > C_A \sqrt{\log N}\right\} = o(N^{-A}).$$

These are the moment and tail bounds in (S29) and (S31), now proved under correlation.

For $\widehat{M} - \Gamma_h$ and complete or deleted scores, bounds for sums over redraw segments give the orders in Section S12. The population inverse and scaled evaluation rows are bounded. Inversion and recentering give

$$\Big\|\sqrt{N} D_h(\widehat{f} - g^*) - N^{-1/2} \sum_i Z_i\Big\|_\infty = O_p\{G \log^3 N/\sqrt{N}\}.$$

The exact deletion identity (S26) is algebraic and uses no population independence. Replacing I there by Γ_h and including its inverse in the first-order score yields

$$\max_b L^{-1/2}\Big\|D_h V_b - \sum_{i\in I_b}(Z_i - \bar{Z})\Big\|_\infty = O_p\Big\{\frac{G\log^3 N}{\sqrt{N}} + \frac{G^{3/2}\log^4 N}{\sqrt{NL}}\Big\}.$$

To verify this order, we define $a_N = \sqrt{G\log N/N} + G\log^2 N/N$ and $d_N = \sqrt{G\log N/N}$. The inverse error times a block score and the block matrix error times a coefficient error are bounded by

$$Ca_N\{\sqrt{G\log N} + \sqrt{G}\log^2 N/\sqrt{L} + \sqrt{L}\,d_N\} + Cd_N\{\sqrt{G\log N} + G\log^2 N/\sqrt{L} + \sqrt{L}\,a_N\}.$$

The two centering products have the same bound. Substitution gives the stated order, exactly as in (S35). Replacing the identity by the population inverse changes only fixed constants.

We remove $\lceil C_A \log N\rceil$ observations per block. By (S8), independent design coupling has error at most JN^{-A-2} for sufficiently large C_A; Gaussian autoregressive coupling covers the joint segments. The segment moments proved above give (S32). The geometric covariance series bounds the difference between the covariance of the retained blocks and (S46) by $C\log N/L$ entrywise.

We then truncate and center the independent normalized block coordinates at $C_A\sqrt{\log N}$. Because $Lh \to \infty$, their fourth moments are bounded, and the average variances are bounded below. These bounds verify conditions M.1, M.2 and E.1 of Chernozhukov, Chetverikov and Kato (2017, Proposition 2.1) with $B_J = C\log N$, which gives the error $C\{\log^9 N/J\}^{1/6} = o(1)$. The positive diagonal variances allow us to apply Chernozhukov, Chetverikov and Kato (2015, Theorems 2 and 3) to restore the prefixes and replace the covariance by (S46).

For the empirical block covariances and the studentization, the independent-block calculation in Section S12 uses only the fourth moments $C(L^2 + L/h)$, polynomially summable block tails, geometric coupling, and the exact deletion error above. This calculation gives

$$\|ND_h\widehat{\Sigma}D_h - \Omega_h\|_{\max}\log^2 N = o_p(1)$$

and the same conditional Gaussian approximation for the resampled maximum with a separately estimated denominator in each draw. Each remainder in the two preceding linearizations, after the required logarithmic factors, tends to zero under $1/5 < \kappa < 1/2$ and $\gamma > \kappa$, as in the independent-block calculation. The bias bounds (S43) satisfy

$$\sqrt{Nh}\,h^2\sqrt{\log N} \to 0, \qquad \sqrt{N}\,h^4\sqrt{\log N} \to 0.$$

Thus the Gaussian approximation concerns the true component values. The studentized quantile comparison proves coverage $1 - \alpha + o(1)$ for the known active set.

On the event $\widehat{S} = S$, the fitted matrices, deletions and resampling maxima equal the known-support quantities. Coverage differs by at most $\Pr(\widehat{S} \neq S)$, also when selected inactive groups are included. The finite-resampling and interpolation arguments remain valid. This completes the proof.

S14 Proofs for general population selection

Correct selection requires the sample fit to retain the population separation between active coefficients and inactive residual scores. We bound the sampling error and show that it preserves the strict inequalities of Theorem 1. In the Gaussian-copula example, the Hermite decomposition reduces the selection calculation to three dimensions.

The following condition collects the premises of Theorem 1 beyond Condition 1(iv). The active part $\bar{\theta}_S$ of $\bar{\theta}$ is its restriction to S. The projection residual equals the true additive response minus its population local-line approximation before averaging over nodes. We fix a constant C.

Condition S1. (i) The marginal density and level-probability bounds in Condition 1(i) and the first and third bounds below hold:

$$\max_j \sup_x \|z_j(x)\|_2 \le C\sqrt{G}, \qquad \max_{j,k}\big\|\mathbb{E}\{\|z_j\|_2^2 z_k z_k^{\mathsf{T}}\}\big\|_{\mathrm{op}} \le CG, \qquad \|\theta^*\|_2 + \|\theta^L\|_2 \le C.$$

(ii) The additive projection residual is bounded by Ch_L^2, $\|\bar{\theta}_S\|_2 \le C$ and $\lambda_n = O(1)$. (iii) The stationary transition is $P(x,\cdot) = (1-\rho)Q(\cdot) + \rho K(x,\cdot)$. Here Q is the stationary covariate distribution, K is a Markov transition with $QK = Q$, and $0 \le \rho \le \rho_0 < 1$.

The middle bound follows by normalization, after enlarging C if needed. Condition S1(iii) permits correlated covariates through Q and temporal dependence through K.

S14.1 Proof of Theorem 1

Within this proof, we write $\lambda = \lambda_n$. We first work in population coordinates. We write $\widetilde{H}$ and $\widetilde{b}$ for the empirically centered quadratic matrix and score in these coordinates, and define T as the block-diagonal empirical normalization. We choose aligned group coordinates; this choice changes neither group norms nor support. Then $H = T^{\mathsf{T}}\widetilde{H}T$ and $b = T^{\mathsf{T}}\widetilde{b}$. The marginal and grid bounds give $0 \preceq A_j(x) \preceq CGI$ and $\mathbb{E}A_j = I$ by the local moment calculation in Section S10. These bounds control the diagonal blocks, and (S37) gives

$$\|T - I\|_{\mathrm{op}} \le C_A r_n, \qquad \|\widetilde{H} - \Gamma_M\|_{\mathrm{op}} \le C_A m r_n, \qquad \|\Gamma_M\|_{\mathrm{op}} \le m.$$

The last inequality follows from positive semidefiniteness and the identity diagonal blocks.

For a centered bounded scalar or vector g, we define W_L as its sum along a complete segment between independent redraws. Conditional stationarity and the geometric length bound give, for every integer $\nu \ge 2$,

$$\mathbb{E}W_L = 0, \qquad \mathbb{E}\|W_L\|^\nu \le \mathbb{E}L^\nu\, \mathbb{E}\|g\|^\nu \le \nu! C^\nu \|g\|_\infty^{\nu-2} \mathbb{E}\|g\|^2.$$

The moment Bernstein inequality of Tropp (2012, Theorem 6.2), with the rectangular dilation used for a vector, applies to the independent complete paths. A union bound over the possible terminal path indices and the geometric bound for the two boundary pieces give the following bound. We set $x_n = \log n$ for a single scalar and $x_n = \log(dn)$ for the score-block union.

$$\Big\|\frac{1}{n}\sum_i g(X_i)\Big\| \le C_A\Big\{\sqrt{\frac{\mathbb{E}\|g\|^2\, x_n}{n}} + \frac{\|g\|_\infty x_n}{n}\Big\}.$$

The exceptional probability is $o(n^{-A})$. We write $\widehat{\lambda}$ for the penalty computed from the screening response scale.

We write $F(X) = \mathbb{E}(Y - \mu \mid X)$ and define $r(U, X) = F(X) - \sum_{j\in M} \phi_j(U_j, X_j)^{\mathsf{T}}\theta_j^*$. Here $\phi_j = (R_j^0)^{\mathsf{T}} v_j$. The function r is the local-polynomial projection residual of F on M, before averaging over the auxiliary nodes. Its supremum norm is at most Ch_L^2, and $\mathbb{E}(\phi_j r) = 0$. The preceding vector bound and the conditional Gaussian score bound give

$$\max_{j\in M} \|\widetilde{b}_j - (\widetilde{H}\theta^*)_j\|_2 \le C_A r_n.$$

The empirical-intercept correction multiplies the empirical feature mean by the average of the projection residual and regression error; this product has a smaller order. Using $\|\theta^*\|_2 \le C$ and transferring the normalization gives

$$\|H - \Gamma_M\|_{\mathrm{op}} + \|b - \Gamma_M\theta^*\|_2 \le C_A m r_n.$$

The projection gives $\|F\|_\infty \le C(\sqrt{mG} + h_L^2)$ and $\mathbb{E}F^2 \le C(m + h_L^4)$. Marginal score errors, including centering and normalization, are at most $C_A\sqrt{m}\, r_n$. On M, population norms are at most $\{\theta^{*\mathsf{T}}\Gamma_M\theta^*\}^{1/2} \le C\sqrt{m}$; this bound holds for all groups because M contains the largest scores. The errors in squared scores are at most $C_A m r_n = o(1)$, and the strict population margin gives the screened set M.

For the response scale, we apply the scalar path bound to $(F^2 - \mathbb{E}F^2)/\sigma_Y^2$. Its supremum norm and variance are bounded by $C(mG + h_L^4)$ because $\sigma_Y \ge \sigma_-$ and $\mathbb{E}F^4 \le \|F\|_\infty^2 \mathbb{E}F^2$. The Gaussian quadratic-form bound and the conditional Gaussian bound for $n^{-1}\sum_i F(X_i)\varepsilon_i$ give the same or a smaller order. The sample-mean square and the factor $n/(n-1)$ contribute corrections of smaller order as well. Hence

$$\Big|\frac{\widehat{\sigma}_Y}{\sigma_Y} - 1\Big| \le C_A\{\sqrt{m}\, r_n + m r_n^2\}, \qquad \sqrt{m}\,|\widehat{\lambda} - \lambda| \le C_A \lambda m r_n(1 + \sqrt{m}\, r_n) = o(\lambda).$$

Here $m r_n = o(\lambda)$ and $\lambda = O(1)$ imply $m r_n \to 0$. All preceding bounds hold simultaneously with probability $1 - o(n^{-A})$.

We now compare the initial group-Lasso fit with its population counterpart. Strong convexity and monotonicity of the group-norm subgradient imply

$$\|\widehat{\theta}^L - \theta^L\|_2 \le \frac{C}{\kappa_\Gamma}\left\{\|b - \Gamma_M\theta^*\|_2 + \|H - \Gamma_M\|_{\mathrm{op}}(\|\theta^L\|_2 + \|\theta^*\|_2) + |\widehat{\lambda} - \lambda|\sqrt{m}\right\}.$$

The preceding rates make the right-hand side $o(\lambda)$. We define $\widehat{w}_j = p'_{\widehat{\lambda},a}(\|\widehat{\theta}^L_j\|_2)$. The SCAD derivative is jointly Lipschitz in its norm and penalty arguments. Summing the blockwise bounds over the groups gives

$$\|\widehat{w} - w\|_2 = o(\lambda), \qquad \max_{j\in M}|\widehat{w}_j - w_j| = o(\lambda).$$

It remains to show that the weighted fit selects S. We restrict the empirical weighted problem to S and write $\widehat{\theta}^R$ for its minimizer, the empirical counterpart of $\bar{\theta}$. Applying the preceding strong convexity argument to the empirical and population restricted problems gives

$$\max_{j\in S}\|\widehat{\theta}^R_j - \bar{\theta}_j\|_2 = o(\lambda).$$

The active margin implies $\widehat{\theta}^R_j \neq 0$ for every $j \in S$. We define $\delta = \widehat{\theta}^R_S - \bar{\theta}_S$. The restricted variational inequalities also give $\delta^\mathsf{T} H_{SS}\delta = o(\lambda^2)$. Because H is positive semidefinite and $H_{jj} = I$, the block-matrix inequality gives $\|H_{jS}\delta\|_2^2 \le \delta^\mathsf{T} H_{SS}\delta$. For an inactive candidate group $j \in M \setminus S$, the score at the restricted solution satisfies

$$\begin{aligned}&\|H_{jS}\widehat{\theta}^R_S - b_j\|_2\\ &\quad\le \|\{\Gamma_M(\bar{\theta} - \theta^*)\}_j\|_2 + \|(H_{jS} - \Gamma_{jS})\bar{\theta}_S\|_2 + \|b_j - (\Gamma_M\theta^*)_j\|_2 + \{\delta^\mathsf{T} H_{SS}\delta\}^{1/2}\\ &\quad\le w_j - c\lambda + o(\lambda) < \widehat{w}_j.\end{aligned}$$

These strict inequalities extend the restricted solution to the unique full weighted solution, whose support is S. This completes the proof.

S14.2 Proof of Proposition S4

We use $\mathcal{A}_j$, Π_j and W from (S41) on the full candidate set M. The local moment bounds in Section S10 give a raw minimum eigenvalue of order G^{-1} and a bounded raw evaluation norm. Centering and normalization give $\sup_x \|z_j(x)\|_2 \le C\sqrt{G}$. The contraction property of the averaging operator gives $\mathbb{E}(z_k z_k^\mathsf{T}) \preceq I$. Hence

$$\mathbb{E}\{\|z_j\|_2^2 z_k z_k^\mathsf{T}\} \preceq CG\,\mathbb{E}(z_k z_k^\mathsf{T}) \preceq CGI.$$

This calculation does not require independence of the covariates X_j and X_k.

We define $\mathcal{C} = \mathcal{R} - I$, with zero columns for empty dependence vectors. The identity (S41) gives

$$\Gamma_M = I - WW^\mathsf{T} + W\mathcal{R}W^\mathsf{T}, \qquad \min(1, c_0)I \preceq \Gamma_M \preceq \max\{1, \|\mathcal{R}\|_{\mathrm{op}}\}I.$$

Here $W^\mathsf{T}W \preceq I$, and the upper constant is fixed because $\mathcal{R}$ has fixed dimension and identity diagonal blocks.

We define $\eta^0_j = \mathcal{A}_j^* f_j$ and $g^0_j = \mathcal{A}_j\eta^0_j$, with $f_j = 0$ outside S. The local-polynomial approximation bound is uniform for both f_j and each nonempty dependence function. For categorical groups the approximation is exact. Orthogonality of the projection Π_j yields

$$a_{j,l} := \int \chi_{j,l}(f_j - g^0_j)\,dP_j = \mathbb{E}\{(\chi_{j,l} - \Pi_j\chi_{j,l})(f_j - \Pi_j f_j)\}, \qquad |a_{j,l}| \le Ch_L^4.$$

There are only a fixed number of these entries. Also $W^\mathsf{T}W = I + O(h_L^4)$. The normal equations give

$$\theta^* - \eta^0 = W\{I + \mathcal{C}W^\mathsf{T}W\}^{-1}\mathcal{C}a.$$

The matrix in braces converges to $\mathcal{R}$, giving a bounded inverse. Each column of W projects one dependence function and has bounded nodewise evaluation. Thus the correction has evaluation and

coefficient norms $O(h_L^4)$ on finitely many groups. The s local residuals of the active components sum to $O(h_L^2)$, proving $\|r_M\|_\infty \le Ch_L^2$. Empty dependence vectors give zero correction.

Finally, $\|\eta^0\|_2^2 \le \sum_{j\in S}\|f_j\|_{P_j}^2 \le C$, which gives $\|\theta^*\|_2 \le C$. Either penalized minimizer has an objective value at most the objective value at zero. Because all penalty weights are nonnegative,

$$\frac{c}{2}\|\theta\|_2^2 \le \theta^{\mathsf{T}}\Gamma_M\theta^* \le C\|\theta\|_2\|\theta^*\|_2, \qquad \theta \in \{\theta^L, \bar{\theta}\}.$$

This proves the remaining coefficient bounds, including the bound for the restricted weighted minimizer. No property of an empirical residual has been assumed. This completes the proof.

S14.3 Proof of Proposition S5

Sylvester's criterion applied to $\Gamma_0 - 0.13I$ gives the leading principal minors

$$\frac{87}{100}, \qquad \frac{2669}{10000}, \qquad \frac{5353}{1000000}.$$

Thus $\lambda_{\min}(\Gamma_0) > 0.13$. For $\varsigma = (-1,-1,1)^{\mathsf{T}}$, direct inversion gives

$$\Gamma_0^{-1}\varsigma = \left(-\frac{2070}{487}, -\frac{2310}{487}, -\frac{680}{487}\right)^{\mathsf{T}}.$$

Hence $\theta^L = \beta^* - \lambda\Gamma_0^{-1}\varsigma$ equals the vector in the proposition and satisfies the Lasso KKT equations. In particular, $680/487 > 101/100$. A set selected by a common cutoff on the absolute initial coefficients that contains the first active coordinate must contain the inactive third coordinate. No hard threshold selects S.

The SCAD weights divided by λ are

$$\left(\frac{269}{270}, 0, \frac{11219}{13149}\right).$$

Solving the restricted weighted equations on S gives

$$\frac{\bar{\theta}_S}{\lambda} = \left(-\frac{221766499}{67059900}, -\frac{5616844}{670599}\right)^{\mathsf{T}}.$$

Both coordinates are negative. The absolute inactive score is

$$\left|\Gamma_{0,3S}(\bar{\theta}_S - \beta_S^*)\right|/\lambda = \frac{18023}{27540} < \frac{11219}{13149}.$$

The strict inactive KKT inequality and positive definiteness prove that the SCAD minimizer has support S.

For the adaptive weights, the values divided by λ are

$$\left(\frac{100}{101}, \frac{1}{5}, \frac{487}{680}\right).$$

The corresponding restricted solution has inactive score $43903/51510$, and

$$\frac{43903}{51510} > \frac{487}{680}.$$

The restricted solution does not extend to a weighted solution with the third coordinate equal to zero. This completes the proof.

S14.4 Proof of Corollary S4

We transfer the population inequalities to the Gaussian-copula model and control the sampling error. We define He_k as the normalized Hermite polynomial of degree k. For Gaussian scores,

$$\mathbb{E}\{\mathrm{He}_k(Z_j) \mid Z_l\} = (\Gamma_0)_{jl}^k \mathrm{He}_k(Z_l).$$

We define $X_j = \Phi(Z_j)$ and $q = \Phi^{-1} \in L^2(0,1)$. We fix $\epsilon > 0$ and choose $q_\epsilon \in C^2[0,1]$ equal to q on $[\epsilon, 1-\epsilon]$. The local coefficients $q_\epsilon(u_g)$ and $h_L q'_\epsilon(u_g)$ have uniform Taylor error $O(h_L^2)$ by $\sum_g \pi_g(x) = 1$; centering subtracts the mean error. We let $h_L \to 0$ and then $\epsilon \downarrow 0$. The Gaussian tails have arbitrarily small L^2 norm. This gives the approximation of q.

In the additive L^2 space, the degree-k Gram matrix is $\Gamma_0^{\circ k}$, whose (j,l) entry is $(\Gamma_0)_{jl}^k$. It is positive definite, uniformly in k, because $\Gamma_0^{\circ k} \to I$. The response scores are zero for $k \geq 2$, and hence removing those coefficients decreases the loss and the group norms. The unique Lasso and weighted solutions lie in the first Hermite degree and equal those in Proposition S5. The local-polynomial criterion dominates the additive criterion at the averaged components: node-conditional variance contributes nonnegative terms to loss and group norms. Approximating the L^2 minimizer attains the limiting minimum. After rescaling coefficients by λ_n and the criterion by λ_n^2, uniform coercivity and strong convexity give convergence in product L^2. The KKT margins, initial norm ordering and adaptive inactive violation are strict multiples of λ_n and persist in the local-polynomial criterion. For the adaptive violation, this persistence uses the convergence of dual norms proved in Section S14.5, which holds for every fixed residual in $L^2(0,1)$.

It remains to pass from this population calculation to the sample problem. Screening retains all three groups. The coordinates are bounded by $C\sqrt{G}$, and $q(X_j)$ and the errors are Gaussian. Truncation at $C_A\sqrt{\log n}$, the Bernstein inequality and Gaussian tails give matrix error r_n outside an event of probability $o(n^{-A})$. The Gaussian score error is at most $C_A\{(G + \log n)/n\}^{1/2}$. We write $m_i = \lambda_n U_i$. Here U_i is centered Gaussian with fixed variance. Then $\mathbb{E}\|z_j(X_{ij})m_i\|_2^2 \leq C\lambda_n^2 G$. Truncating $|U_i|$ at $C_A\sqrt{\log n}$ bounds the centered summands by $C_A\lambda_n\sqrt{G\log n}$ and changes the means by less than any fixed negative power of n for large C_A. The vector Bernstein inequality gives error $C_A\lambda_n\{\sqrt{G\log n/n} + \sqrt{G}(\log n)^{3/2}/n\} = O(\lambda_n r_n)$. Both score errors are $o(\lambda_n)$ by $mr_n = o(\lambda_n)$ and $\lambda_n = O(1)$. The third unpenalized block is $o(\lambda_n)$ by L^2 convergence, and the screening response-scale ratio is $1 + o_p(1)$. The optimality argument in the proof of Theorem 1 gives the SCAD conclusion. The comparison matrix with off-diagonal entries 0.70, 0.05, 0.55 has Gauss–Seidel radius $0.837 < 0.84$; the inequality $30|\log 0.84| > 4$ makes the sweep limit sufficient. The strict sample ordering $\|\theta_3^L\|_2 > \|\theta_1^L\|_2$ excludes every upper-level-set threshold, and the adaptive inactive violation persists. This completes the proof.

S14.5 A BOUNDED SMOOTH SELECTION EXAMPLE

Replacing the Gaussian score by a fixed smooth bounded function preserves the three strict inequalities in Proposition S5.

We write $\beta^\sharp = \beta^*/\lambda$, with the population coefficient β^* defined before that proposition. For $0 < \epsilon < 1/4$, we choose a smooth function χ_ϵ on $[0,1]$ that equals one on $[\epsilon, 1-\epsilon]$, equals zero in neighborhoods of both endpoints, satisfies $0 \leq \chi_\epsilon \leq 1$, and is symmetric about $1/2$. We set $q_\epsilon = \chi_\epsilon q$. Here $q = \Phi^{-1}$. We extend q_ϵ by zero at the boundaries. Then

$$q_\epsilon \in C^\infty[0,1], \qquad \int_0^1 q_\epsilon(x)\,dx = 0, \qquad \delta_\epsilon = \|q_\epsilon - q\|_{L^2(0,1)} \longrightarrow 0.$$

The mean is zero by antisymmetry. Dominated convergence proves the norm limit because $q \in L^2(0,1)$. We choose ϵ below and then keep it fixed as n increases.

We define $\mathcal{H}$ as the direct sum of three centered $L^2(0,1)$ spaces, with $\|g\|_{\mathcal{H}}^2 = \sum_j \|g_j\|_2^2$. We define the additive map $Tg = \sum_j g_j(X_j)$ and the population Gram operator $\Gamma = T^*T$ under the Gaussian copula with correlation matrix Γ_0. The Hermite calculation in the preceding proof diagonalizes Γ into the matrices $\Gamma_0^{\circ k}$, $k \geq 1$. For $k \geq 2$, every off-diagonal absolute row sum is at most $0.70^2 + 0.55^2 = 0.7925$. The bound for $k = 1$ was proved above by Sylvester's criterion. Hence

$$0.13\|g\|_{\mathcal{H}}^2 \leq \|Tg\|_2^2 \leq 3\|g\|_{\mathcal{H}}^2.$$

These inequalities give a common strong-convexity constant on the entire additive space, not only on its first Hermite degree.

We divide the response signal and every candidate by λ_n, and divide the objective by λ_n^2. We define $f = (\beta_1^\sharp q, \beta_2^\sharp q, 0)$ and $f^\epsilon = (\beta_1^\sharp q_\epsilon, \beta_2^\sharp q_\epsilon, 0)$. We write g^L and $g^{L,\epsilon}$ for the corresponding

population minimizers with group penalty $\sum_j \|g_j\|_2$. Subtracting their optimality equations and using monotonicity of the group-norm subgradient gives

$$0.13\|g^{L,\epsilon} - g^L\|_{\mathcal{H}}^2 \le \langle g^{L,\epsilon} - g^L, \Gamma(f^\epsilon - f)\rangle \le C\delta_\epsilon \|g^{L,\epsilon} - g^L\|_{\mathcal{H}}.$$

The initial norms change by at most $C\delta_\epsilon$. The SCAD derivative at penalty one is Lipschitz; $1/\|g_j^L\|_2$ is locally Lipschitz because all original norms are positive. The same restricted optimality calculation gives error $C\delta_\epsilon$ for the weighted minimizer on S. Because $\|\Gamma\|_{\mathrm{op}} \le 3$, the inactive scores have the same error.

The original SCAD inactive margin, reversed initial-norm gap and adaptive inactive violation, respectively, are

$$\frac{2666179}{13411980} > 0, \qquad \frac{18813}{48700} > 0, \qquad \frac{28051}{206040} > 0. \tag{S47}$$

We choose ϵ such that both restricted active norms stay positive and all gaps in (S47) retain at least half their values. The smoothed SCAD solution then has support S; no initial-norm threshold selects S, and the adaptive solution retains its inactive violation.

We next transfer these strict inequalities to the local-polynomial criterion. For a centered local polynomial $t_j(U_j, X_j)$, we write $g_j(X_j) = \mathbb{E}\{t_j(U_j, X_j) \mid X_j\}$ and $v_j = \mathbb{E}\,\mathrm{Var}\{t_j(U_j, X_j) \mid X_j\}$. Conditional independence of the auxiliary indices gives

$$\mathbb{E}\Big\{Tf^\epsilon - \sum_j t_j\Big\}^2 = \|T(f^\epsilon - g)\|_2^2 + \sum_j v_j, \qquad \mathbb{E}t_j^2 = \|g_j\|_2^2 + v_j.$$

For fixed nonnegative weights, the augmented objective exceeds the full-space objective at g by at least $\sum_j v_j/2$, with strong-convexity constant at least 0.13 in the augmented norm. Smooth centered functions are dense in centered L^2 and have convergent centered local-linear approximants by Taylor's formula. Successive approximation of the full-space minimizer attains the limiting minimum. The lower bound and strong convexity give convergence of averaged minimizers in $\mathcal{H}$, $\sum_j v_j \to 0$, and convergence of augmented group norms. This argument applies first to the initial Lasso and then to the restricted SCAD and adaptive objectives with convergent weights.

We also verify the convergence of the dual norms of the inactive scores. We define $A_{h,j}$ to map a centered local polynomial to its conditional average in centered $L^2(0,1)$. This map is a contraction. For every fixed centered $r \in L^2(0,1)$,

$$\|A_{h,j}^* r\| = \sup_{\|t\| \le 1} \langle r, A_{h,j} t\rangle \longrightarrow \|r\|_2.$$

The contraction property gives the upper bound and extends the limit to $r_h \to r$ in L^2. For $r \neq 0$, smooth and local-polynomial approximation of $r/\|r\|_2$ gives the lower bound; for $r = 0$, the limit is immediate. We apply this limit to the conditional expectation of the restricted residual given the inactive covariate. All three inequalities in (S47) persist for small h_L, without a uniform bound on the correlated projection residual.

It remains to pass to the sample problem. Screening retains all three candidates. Uniform marginals give population-normalized coordinate bounds $\|z_j(x)\|_2 \le C\sqrt{G}$ and $\mathbb{E}(z_j z_j^{\mathsf{T}}) \preceq I$. In particular,

$$\big\|\mathbb{E}\{\|z_j\|_2^2 z_k z_k^{\mathsf{T}}\}\big\|_{\mathrm{op}} \le CG.$$

The bound does not require a bounded copula density. By the matrix Bernstein inequality, as in the proof of Theorem 1, the matrix error is $O_p(r_n)$. For fixed ϵ, Tf^ϵ is bounded, giving signal-score error $O_p(\lambda_n r_n)$. Conditional Gaussian concentration gives noise error $O_p(r_n)$, and the screening response-scale ratio tends to one. Because $r_n = o(\lambda_n)$ and $\lambda_n = O(1)$, strong convexity and subgradient monotonicity give scaled coefficient and weight errors $o_p(1)$. All three strict sample inequalities hold with probability tending to one.

The Hermite bounds 0.70, 0.05, 0.55 on conditional-expectation norms are preserved by the contraction property of $A_{h,j}$. The empirical errors are $o_p(1)$, giving a Gauss–Seidel radius below 0.84 with probability tending to one. By the inequality $30|\log 0.84| > 4$, the sweep error is $o(n^{-3})$; earlier stopping gives error $O(n^{-3})$ by strong convexity. Once this error is below the strict margin, a complete sweep sets the inactive SCAD group to zero while active norms stay positive. A zero

third adaptive group would leave a violation of order λ_n, contradicting the stopping rule. The initial ordering excludes every threshold, proving the three limits for the implemented calculation.

We take $f_{j,n}(x) = \lambda_n \beta_j^\sharp q_\epsilon(x)$ for $j = 1, 2$ and $f_{3,n} = 0$. These functions are centered, bounded and twice continuously differentiable uniformly in n, proving Theorem 2 for a smooth bounded model sequence. This completes the proof. The following refinement removes the condition $r_n = o(\lambda_n)$ for three groups.

S14.6 Selection with a fixed penalty multiplier

The preceding proof controls every coordinate through a common concentration bound. For three groups, a direct bound on the Gaussian score norm permits a fixed penalty multiplier for the same selector. We keep the Gaussian-copula design, the vector $\beta^\sharp = \beta^*/\lambda$, and a fixed q_ϵ from Section S14.5. We retain all three candidate groups. The covariate vectors are independent over time. The centered stationary Gaussian errors are independent of the complete design process, have variance $\sigma^2 > 0$, and satisfy $2\pi\|g_\varepsilon\|_\infty \le v$ with the spectral convention preceding Proposition S2. The screening and selection blocks each contain a fixed positive fraction of the epoch observations. We write $N_L \asymp n$ for the selection-block size, and set

$$\lambda_n = c_\lambda \sigma \left\{ \frac{\log 3}{N_L h_L} \right\}^{1/2}, \qquad f_{j,n} = \lambda_n \beta_j^\sharp q_\epsilon \quad (j = 1, 2), \qquad f_{3,n} = 0.$$

The algorithm uses $\widehat{\lambda} = c_\lambda \widehat{\sigma}_Y \{\log 3/(N_L h_L)\}^{1/2}$. Here $\widehat{\sigma}_Y$ is the screening-block response standard deviation. We use this same penalty in the initial group Lasso, the SCAD derivative and the adaptive weights $\widehat{\lambda}^2/\|\widehat{\theta}_j^L\|_2$. The numerical tolerances are those in Section S20. The following corollary gives simultaneous conclusions for these selection rules.

Corollary S5. *Assume that the preceding model holds with sufficiently small fixed $\epsilon > 0$. There exists a fixed $C < \infty$ such that, if*

$$c_\lambda \ge C, \qquad h_L \asymp n^{-1/5}, \qquad G \asymp h_L^{-1}, \qquad 4 \le (G-1)h_L \le 8,$$

then, for every fixed $A > 0$,

$$\Pr(\widehat{S}^{\mathrm{SCAD}} = S) = 1 - o(n^{-A}), \qquad \Pr(\widehat{S}^{\mathrm{adap}} = S) = o(n^{-A}),$$
$$\Pr\{\exists t \ge 0 : \{j : \|\widehat{\theta}_j^L\|_2 > t\} = S\} = o(n^{-A}).$$

In Corollary S5, $\lambda_n \asymp n^{-2/5}$ and the minimum active component norm is of the same order. The multiplier c_λ remains fixed as n increases. The conclusion also permits an error spectrum that equals zero at some frequencies. The corollary is a comparison on the constructed smooth model, not a uniform ordering of the three selectors over all correlated additive models.

Proof. We bound the error score without the restriction $r_n = o(\lambda_n)$. We rescale coefficients by λ_n and the objective by λ_n^2, giving the fixed signal Tf^ϵ. For sufficiently small fixed ϵ and large n, Section S14.5 bounds the initial norms, reversed norm gap, restricted active norms, SCAD inactive margin and adaptive violation from below by a common $d_0 > 0$. This bound is independent of c_λ, n and A. The augmented Gram eigenvalues lie in $[0.13, 3]$.

We use the aligned population coordinates of Section S14. We define b^{sig} as the empirical score with the error omitted and b_0^{sig} as its population counterpart after the same coordinate alignment. Because Tf^ϵ is bounded, the matrix Bernstein inequality and the corresponding vector score bound give, with probability $1 - o(n^{-A})$,

$$\|H - \Gamma_M\|_{\mathrm{op}} + \lambda_n^{-1}\|b^{\mathrm{sig}} - b_0^{\mathrm{sig}}\|_2 \le C_A \left\{ \sqrt{\frac{G \log n}{n}} + \frac{G \log^2 n}{n} \right\} = o(1).$$

Empirical centering and within-group normalization are included in this comparison. Indeed, $\|z_j(x)\|_2 \le C\sqrt{G}$, $\mathbb{E} z_j z_j^\mathsf{T} \preceq I$, and $\|Tf^\epsilon\|_\infty \le C$ bound the signal-score covariance by a fixed multiple of I and its summands by $C\sqrt{G}$. There are only three groups. Thus the normalized sample Gram matrix has smallest eigenvalue at least $0.13/2$ on this event.

The error score has a smaller bound than the preceding joint uniform bound. In the empirical coordinates we define $\widetilde{a}_{ij} = R_j^{\mathsf{T}} a_j(X_{ij})$ and $e_j = N_L^{-1} \sum_{i \in I_L} \widetilde{a}_{ij} \varepsilon_i$. The conditional Jensen inequality, the empirical normalization and the spectral bound proved in Proposition S2 give

$$N_L^{-1} \sum_{i \in I_L} \widetilde{a}_{ij} \widetilde{a}_{ij}^{\mathsf{T}} \preceq I, \qquad \mathrm{Cov}(e_j \mid X) \preceq vI/N_L.$$

Each group has at most $2G$ coordinates. Diagonalizing this covariance and applying Markov's inequality at $N_L/(4v)$ yields

$$\Pr\{\|e_j\|_2^2 > C_1 vG/N_L \mid X\} \le 2^G \exp(-C_1 G/4) \le \exp(-cG)$$

for a sufficiently large fixed C_1. We stack the three vectors e_j into e. A union bound over the three groups gives

$$\frac{\|e\|_2}{\lambda_n} \le \frac{\sqrt{3C_1 vGh_L}}{c_\lambda \sigma \sqrt{\log 3}} \le \frac{C_2}{c_\lambda}$$

outside an event of probability at most $3\exp(-cG)$. The constants C_1, C_2 do not depend on A.

The Gaussian quadratic-form bound, with covariance norm at most v, controls the second moment and the squared mean of the error at the same probability level. The bounded signal is $O(\lambda_n)$, with empirical square $O(\lambda_n^2)$. On the event of bounded empirical error variance, the Cauchy–Schwarz inequality bounds the cross term by $O(\lambda_n)$, giving

$$\widehat{\sigma}_Y/\sigma = 1 + o(1), \qquad \widehat{\lambda}/\lambda_n = 1 + o(1)$$

with probability $1 - o(n^{-A})$. No independence between the screening and selection error vectors is used.

We fix $\eta > 0$ and choose fixed c_λ with $C_2/c_\lambda \le \eta$. Subtracting rescaled optimality equations and using subgradient monotonicity gives initial coefficient and norm errors $C\eta + o(1)$. The population norms are bounded from below, and both rescaled weight maps are locally Lipschitz. The same argument on S bounds restricted weighted errors by $C\eta + o(1)$; bounded Gram operators give the same error for the inactive scores. We choose $\eta > 0$ such that all errors are below $d_0/2$. The SCAD restricted solution retains both active groups and satisfies the inactive inequality strictly, and hence minimizes the objective over all three groups. The reversed initial ordering excludes every threshold, and the adaptive inactive violation persists. The exceptional probability is $o(n^{-A}) + 3\exp(-cG) = o(n^{-A})$.

It remains to show that the computed solution has the same support as the exact minimizer. The design contraction bounds in Section S14.5 give a Gauss–Seidel radius below 0.84 with probability $1 - o(n^{-A})$ and sweep error $o(N_L^{-3})$. At sweep completion, each block residual is the sum of the later coefficient changes multiplied by cross blocks of H, giving optimality error $o(N_L^{-3})$. An earlier stop gives coefficient error $O(N_L^{-3})$ by strong convexity. Lipschitz weight bounds apply to the computed initial fit. Because $N_L^{-3} = o(\lambda_n)$, the ordering and SCAD margins persist; Proposition S8 gives the support. If the computed third adaptive norm were at most N_L^{-3}, active optimality and strong convexity would give distance $O(N_L^{-3})$ from the restricted solution on S. The inactive violation of order λ_n would then contradict the tolerance. Thus the reported adaptive support differs from S. Raw eigenvalues of order G^{-1} and positive normalized eigenvalues exceed the numerical cutoffs. This completes the proof. □

S15 Proofs for continuing estimation

Support stabilization permits estimation on an asymptotically full fraction of the stream. The pooled quadratic equation gives the component-risk bound and the derivative needed for exact deleted-block inference. Both calculations use the current-bandwidth average of the fitted local polynomials and their pooled empirical centers.

S15.1 Independent alternative in Theorem 4

We first bound the probability of any incorrect selection update after an early time. Put $t_0 = n/\log n$ and $k_0 = \lceil t_0^{3/4} \rceil$. Apply the assumed selection corollary to the indices in (S12). The gaps between

consecutive selection indices are at least one, which gives redraw probabilities at least $1-\rho_0$ and retained Gaussian covariance at most vI. These redraw and covariance bounds are the bounds used in Theorem S2 and both corollaries. Each transition admits an independent redraw indicator with success probability $1-\rho_0$; additional redraws split the segments of geometric length. For $r=2,4$, splitting decreases the sum of rth powers of lengths. Redraw values remain independent of these lengths. The matrix and score bounds are uniform over groups and candidate subsets, allowing screening, scale estimation and selection on the same observations.

For retained size $k \geq k_0$, the support error probability is uniformly $o(k^{-A})$ for every fixed $A>0$. Summing over at most $n^{3/4}$ updates gives total error probability $n^{3/4}o(k_0^{-A}) = o(n^{-D})$ for every fixed $D>0$ by taking $A > 1+4D/3$. For large n, every block after time t_0 contains more than 75 selection indices, and the numerical rule updates at every boundary. The first update with k_0 observations occurs at $t_0\{1+o(1)\} + O(t_0^{3/5}) = o(n/\log\log n)$. Outside this error event, later updates select S, and all earlier resets precede the activation of the current reporting grids.

Let u_l be the first reporting time of the current grid. Nesting and monotonicity of ℓ_t give $cn \leq u_l \leq n$ for fixed $c>0$, with activation of order $n/\log\log n$. The first complete block adds $O(n^{3/5})$ indices. On support stabilization, $\mathcal{I}_n$ is the deterministic set excluding this prefix and the selection subsequence. Thus

$$N = |\mathcal{I}_n| = n - o(n), \qquad \min \mathcal{I}_n \geq cn/\log\log n.$$

The same argument applies to the inference grid and proves the two observation-count conclusions.

Write $h = h_n = c_E n^{-1/5}$ for the point-estimation bandwidth at the reporting time, and let h_i be the bandwidth of the block that contains observation $i \in \mathcal{I}_n$. The block lengths are $O(i^{3/5})$, and hence $h_i = c_E i^{-1/5}\{1+o(1)\}$ uniformly on this set. Riemann summation over the retained indices gives

$$\frac{1}{N}\sum_{i\in\mathcal{I}_n} h_i^2 \leq Ch^2, \qquad \frac{1}{N}\sum_{i\in\mathcal{I}_n} h_i^4 \leq Ch^4, \qquad 1 \leq h_i/h \leq C(\log\log n)^{1/5}.$$

An incomplete terminal block evaluated with its reporting-time bandwidth satisfies the same bounds. The omitted prefix and selection indices have contributions smaller than the corresponding full sums.

The component risk follows from the normal equation of the pooled quadratic, and we first bound the matrix of that equation. Use the population coordinates of Section S10 for the current grid and reference bandwidth h, including the intercept. The powers $(1,(x-u)/h)$ are a nonsingular rescaling of the algorithm's coordinates. Let ϕ_i be the common feature vector evaluated at auxiliary nodes whose probabilities use h_i, and put

$$M_i = \mathbb{E}(\phi_i\phi_i^{\mathsf{T}} \mid X_i), \qquad \Gamma_n = \frac{1}{N}\sum_{i\in\mathcal{I}_n}\mathbb{E}M_i, \qquad \widehat{M}_n = \frac{1}{N}\sum_{i\in\mathcal{I}_n} M_i.$$

Here Γ_n and $\widehat{M}_n$ are the population and empirical averages of the conditional moment matrices M_i. On the last half of the sample, h_i/h is bounded and $\pi_{h_i,g}(x) \geq c\pi_{h,g}(x)$ on the reference support by the kernel and denominator bounds. Because the active set is fixed and node probabilities multiply, the population quadratic of these observations is bounded below by a multiple of the reference quadratic, which equals the identity. Node moments and the average of h_i^2/h^2 give the upper bound. Thus

$$cI \preceq \Gamma_n \preceq CI, \qquad \|M_i\|_{\mathrm{op}} \leq CG\{1+(h_i/h)^2\}.$$

These bounds are consequences of the population model.

Separate the retained indices into $\ell = \lceil C_D \log n\rceil$ residue classes in physical time and couple within each class to independent stationary covariates. The coupling error is at most $n\rho^{\ell}$. Each class has a positive asymptotic fraction of its indices in the last half of the sample. Matrix Chernoff, applied to the nonidentically distributed positive matrices of each class, then gives

$$\Pr\{cI \not\preceq \widehat{M}_n \text{ or } \widehat{M}_n \not\preceq CI\} \leq C\ell q \exp\left\{-\frac{cN}{G(\log\log n)^{2/5}\ell}\right\} + n\rho^{\ell} = o(n^{-D})$$

for every fixed $D>0$, with possibly different fixed constants c, C. Here $q = O(G)$ because S is fixed, and $G = O(\ell_n n^{1/5})$. Within this proof, let $\mathcal{G}_n$ denote the complementary event, on which these matrix bounds hold.

Let $z_{j,h}(x)$ be the averaged evaluation vector whose inner product with the coefficient block gives the local-polynomial curve at x before centering. With the grid spacing $\Delta = (G-1)^{-1}$, each nonzero raw coordinate has size $O(\Delta/h)$, and at most $O(h/\Delta)$ node blocks are nonzero. The eigenvalues of a reference node moment are of order Δ. Conditional Jensen also bounds the integrated evaluation quadratic of this coordinate by the identity. Thus,

$$\sup_x \|z_{j,h}(x)\|_2^2 \le C/h, \qquad \mathbb{E}\{z_{j,h}(X_j) z_{j,h}(X_j)^\mathsf{T}\} \preceq I.$$

This estimate uses the averaging in the reported component and does not replace h^{-1} by the larger number G.

Let $\bar{z}_j$ be the pooled empirical mean of the same reference coordinate, using each observation's bandwidth h_i. The actual reported component subtracts $\bar{z}_j^\mathsf{T}\widehat{\eta}_j$. On $\mathcal{G}_n$, the intercept row of $\widehat{M}_n$ gives $\sum_j \|\bar{z}_j\|_2^2 \le C$. Let $e_j(x)$ contain $z_{j,h}(x) - \bar{z}_j$ in block j and zeros elsewhere. Then

$$\left\| \int e_j(x) e_j(x)^\mathsf{T} \, dP_j(x) \right\|_{\mathrm{op}} \le C, \qquad \int \|e_j(x)\|_2^2 \, dP_j(x) \le C/h.$$

We now compare the pooled fit with a deterministic coefficient vector built from local Taylor polynomials. At node u, take the Taylor polynomial $f_j(u) + f_j'(u)(x-u)$. Transfer its reference population mean to the intercept, and write the resulting full coefficient vector as η^T. For observation i, the complete residual r_i of this vector satisfies $|r_i| \le C \sum_{j \in S} h_i^2 \le C h_i^2$ because S is fixed. The known-support empirical equation is

$$\widehat{\eta} - \eta^T = \widehat{M}_n^{-1}(b_r + b_\varepsilon), \qquad b_r = \frac{1}{N}\sum_{i \in \mathcal{I}_n} \mathbb{E}(\phi_i r_i \mid X_i), \qquad b_\varepsilon = \frac{1}{N}\sum_{i \in \mathcal{I}_n} \mathbb{E}(\phi_i \mid X_i)\varepsilon_i.$$

Here b_r is the score of the approximation residual and b_ε is the score of the errors. Conditional Cauchy–Schwarz gives

$$b_r^\mathsf{T} \widehat{M}_n^{-1} b_r \le \frac{1}{N}\sum_{i \in \mathcal{I}_n} \mathbb{E}(r_i^2 \mid X_i) \le C h^4.$$

The integrated evaluation operator bound makes the approximation contribution to the component risk at most Ch^4 on $\mathcal{G}_n$.

Independence of the errors and the design, the error covariance bound vI, and conditional Jensen give

$$\mathrm{Cov}(b_\varepsilon \mid X) \preceq \frac{v}{N}\widehat{M}_n.$$

The integrated conditional variance of the reported component is thus at most

$$\frac{v}{N}\int e_j(x)^\mathsf{T} \widehat{M}_n^{-1} e_j(x) \, dP_j(x) \le \frac{C}{Nh}$$

on $\mathcal{G}_n$. This bound retains the effect of empirical centering.

It remains to bound the smoothing bias and the error of the estimated component center. The current-bandwidth average of the Taylor polynomial differs from f_j by at most Ch^2. Its pooled empirical center equals $N^{-1}\sum_{i \in \mathcal{I}_n} f_j(X_{ij}) + O(h^2)$. For deterministic retained indices, the redraw covariance identity of the population model gives

$$\mathbb{E}\left\{ \frac{1}{N}\sum_{i \in \mathcal{I}_n} f_j(X_{ij}) \right\}^2 \le \frac{1+\rho_0}{1-\rho_0}\frac{\|f_j\|_{P_j}^2}{N}.$$

Combining these bounds, we obtain the bound $C\{h^4 + (Nh)^{-1} + N^{-1}\}$ for the expected component loss on $\mathcal{G}_n$ under correct selection.

Raw node eigenvalues have orders Δ and Δh^2, exceeding N^{-3}; the normalized lower bound on $\mathcal{G}_n$ exceeds N^{-1}. Thus the zero decisions are inactive on this event. Clipping does not increase the error for $\log N \ge B$. A zero decision has component loss at most B^2. Every other loss on the complementary event is at most $4\log^2 n$ for sufficiently large n. The probability of the complementary event is $o(n^{-D})$ for every fixed D. Multiplying by the polynomial number of candidates also gives a negligible aggregate loss. With $N/n \to 1$ and $h = c_E n^{-1/5}$, the expected-risk rate follows. The moment bounds use only the density bounds, and the Taylor approximation uses the derivatives of f_j. This completes the proof.

S15.2 Bias of the reported component

Blocks are fitted at different bandwidths. Pooling moments at different bandwidths changes the fitted local slopes as well as the local intercepts, both of which enter the reported component average. A univariate population calculation identifies the resulting bias without an expansion for a random selected fit.

We assume the covariate is uniform on $[0,1]$, we define f to have four bounded derivatives and mean zero, and we fix an interior interval. We use the deterministic oracle index set $\mathcal{I}_n$ from the preceding proof. We define $h = h_n = c_E n^{-1/5}$, $N = |\mathcal{I}_n|$, $\overline{h^r} = N^{-1}\sum_{i\in\mathcal{I}_n} h_i^r$ and $\mu_r = \int v^r K(v)\,dv$. The grid spacing satisfies $\Delta/h \to 0$. At an interior node u, the pooled population local intercept and slope, denoted by $\beta_{0,n}(u)$ and $\beta_{1,n}(u)$, satisfy

$$\beta_{0,n}(u) = f(u) + \frac{\mu_2}{2}\overline{h^2} f''(u) + o(h^2), \qquad \beta_{1,n}(u) = f'(u) + O(h^2) + o(h).$$

The zeroth and second nodal moments are $\Delta\{1+o(1)\}$ and $\Delta\mu_2\overline{h^2}\{1+o(1)\}$; symmetry gives a first moment equal to zero. We expand $f(u+h_i v)$ in the normal equations, using $\overline{h^4} = O(h^4)$ and $\overline{h^2} \asymp h^2$. We write $\overline{D} = N^{-1}\sum_i \mathbb{E}D_i$ and $\overline{c} = N^{-1}\sum_i \mathbb{E}a_i$ for the pooled population moments of the univariate basis, using bandwidth h_i for observation i. For the constant-polynomial vector e, the identities $\overline{D}e = \overline{c}$ and $\overline{c}^{\dagger}e = 1$ give a pooled fitted mean equal to zero under the corresponding centering measure.

We define $g_n^*(x)$ as the average of these local polynomials with the reporting weights at bandwidth h. A second Taylor expansion about x gives

$$g_n^*(x) - f(x) = \frac{\mu_2}{2}\{\overline{h^2} - h^2\} f''(x) + o(h^2), \qquad x \text{ in the fixed interior interval.}$$

For $h_i = c_E i^{-1/5}\{1+o(1)\}$ on the oracle set, $\overline{h^2}/h^2 \to a_2 = 5/3$. The Epanechnikov kernel has $\mu_2 = 1/5$, and the leading bias is $h^2 f''(x)/15$. If the fit and report use a common bandwidth, the term of order h^2 in the reported average equals zero, although each fitted local intercept has a term of order h^2.

The same calculation is expressed in terms of the population smoothing kernel. At relative observation time $t \in (0,1]$, we define

$$L_t(v) = t^{1/5}\int K(w)K\{t^{1/5}(v-w)\}\left\{1 - \frac{w(v-w)}{\mu_2 a_2}\right\} dw. \tag{S48}$$

The first term in braces in (S48) comes from the fitted local intercept, and the second comes from the fitted local slope evaluated at the reporting bandwidth. We use $z = t^{1/5}(v-w)$; symmetry of K gives

$$\int L_t(v)\,dv = 1, \qquad \int vL_t(v)\,dv = 0, \qquad \int v^2 L_t(v)\,dv = \mu_2\{1 + t^{-2/5}(1-2/a_2)\}.$$

Integrating the last identity over $0 < t \le 1$ gives $\mu_2(a_2 - 1) = 2/15$. This calculation yields the same bias coefficient.

For independent errors of variance σ^2, we define the oracle noise term $Z_n(x) = N^{-1}\sum_{i\in\mathcal{I}_n} h^{-1}L_{i/n}\{(X_i - x)/h\}\varepsilon_i$. A change of variables gives

$$\mathrm{Var}\{Z_n(x)\} = \frac{\sigma^2}{nh}\int_0^1\int L_t(v)^2\,dv\,dt\{1+o(1)\}.$$

These interior population identities explain why averaging the bias and variance constants of ordinary local-intercept fits does not give the constants for the reported component. Theorem 4 controls the actual finite-grid selected estimator, including boundaries and empirical centering, without requiring a leading integrated-risk expansion.

S15.3 Independent alternative in Theorem 5

The inference proof uses the pooled population equation to obtain influence sums for both the complete fit and the deleted-block fits. On the support stabilization event of Theorem 4, the inference

state equals its oracle state on the deterministic index set $\mathcal{I}_n$. This set excludes the activation prefix, the selection indices and the blocks below the minimum size. For the known-support inference state, we define $N = |\mathcal{I}_n|$, $h = c_I n^{-1/4}$ and $G = G_{I,n}$. Here $G_{I,n}$ is the number of nodes in the inference reporting grid at time n. The bandwidth h differs from the point-estimation bandwidth $h_n = c_E n^{-1/5}$ of the preceding proof. The activation and block bounds give

$$N/n \to 1, \qquad G = O(\ell_n n^{1/4}), \qquad h \le h_i \le Ch(\log\log n)^{1/4},$$
$$\frac{1}{N}\sum_{i\in\mathcal{I}_n} h_i^r \le Ch^r \quad (r = 1,2,3), \qquad \frac{1}{N}\sum_{i\in\mathcal{I}_n} h_i^4 \le Ch^4\{1 + \log\log\log n\}.$$

Integral comparison proves these bounds on the actual retained indices. An incomplete reporting block has $O(n^{3/5})$ observations, all evaluated at bandwidth h, and obeys the same bounds.

Pooled population equation. The accumulated fit uses observations processed at different bandwidths. We normalize its coordinates using their average population moments; this normalization yields the population equation for the actual accumulated criterion. For each group, we center and normalize the averaged population moments $N^{-1}\sum_i \mathbb{E}D_{ij}$ and $N^{-1}\sum_i \mathbb{E}a_{ij}$, using the common local powers $(1, (X_{ij} - u)/h)$. We denote the resulting deterministic coordinate matrix by $R_{n,j}$. We include an intercept, and define ϕ_{ni} as the full normalized feature vector with auxiliary nodes drawn at bandwidth h_i. We define

$$M_{ni} = \mathbb{E}_U(\phi_{ni}\phi_{ni}^{\mathsf{T}} \mid X_i), \qquad \Gamma_n = \frac{1}{N}\sum_i \mathbb{E}M_{ni}, \qquad \eta_n^* = \Gamma_n^{-1}\frac{1}{N}\sum_i \mathbb{E}(\phi_{ni}Y_i).$$

The pooled population normal equation defines η_n^* from the average moment matrix Γ_n. All sums in this proof are over $\mathcal{I}_n$ unless stated otherwise. We define $z_{ni,j} = \mathbb{E}_U(\phi_{ni,j} \mid X_i)$. The pooled centering gives $N^{-1}\sum_i \mathbb{E}z_{ni,j} = 0$. For a reported pair $k = (j, x)$, we construct $e_{n,k}$ to contain the group-j average at x with bandwidth h and zeros in the other blocks, including the intercept. The population reported value of this pair is $g_{n,k}^* = e_{n,k}^{\mathsf{T}}\eta_n^*$.

We first bound the distance between the population reported value $g_{n,k}^*$ and the component value $f_j(x)$. We write $h_{\max} = \max_i h_i$ and $\Delta = (G-1)^{-1}$. The last half of the observations gives a lower eigenvalue bound $c\Delta$ for each scaled nodal moment. The average second moment gives the upper bound $C\Delta$. The score of a local Taylor residual is bounded at each node by

$$C\Delta\left\{\frac{1}{N}\sum_i h_i^2 + \frac{1}{Nh}\sum_i h_i^3\right\} \le C\Delta h^2.$$

Thus the separate one-group projection has scaled nodal coefficient error $O(h^2)$ and, at observation i, polynomial residual bounded by Ch_i^2.

The cross blocks depend on feature means at each h_i. On an infinite equispaced grid, partitioning the integral into intervals of length Δ makes the numerator sum equal the periodic denominator, giving $\int \pi_{h_i,g} = \Delta$. Reflection about u_g gives a first centered moment equal to zero. These identities hold on the finite grid at distance above $2h_{\max}$ from the boundary. A Taylor expansion of the density gives means $\Delta p_j(u_g) + O(\Delta h_i^2)$ and $O(\Delta h_i^2/h)$. The boundary probability is $O(h_{\max})$, and the nodal inverse bounds give

$$\max_{i,j}\|\mathbb{E}z_{ni,j}\|_2 \le C\sqrt{h_{\max}}\log^C n, \qquad \|\Gamma_n - I\|_{\mathrm{op}} \le Ch_{\max}\log^C n = o(1).$$

For the second bound, independence of the covariate coordinates gives $(\Gamma_n)_{jk} = N^{-1}\sum_i(\mathbb{E}z_{ni,j})(\mathbb{E}z_{ni,k})^{\mathsf{T}}$ for $j \ne k$; the diagonal blocks are identity matrices and the intercept cross blocks are zero. Only the ratio $h_{\max}/h$ and the fixed number of active groups enter the logarithmic factors.

The mean residual of each separate projection is $O(h_i^2)$. By the preceding mean bound, the cross score of this residual in another group has norm at most $C\sqrt{h_{\max}}h^2\log^C n$. Applying Γ_n^{-1} and the evaluation-row bound $\|e_{n,k}\|_2 \le Ch^{-1/2}$ gives

$$\max_{k\ \text{continuous}} |g_{n,k}^* - f_j(x)| \le Ch^2\log^C n, \qquad \max_i \operatorname*{ess\,sup}_{U,X} |m(X) - \phi_{ni}^{\mathsf{T}}\eta_n^*| \le Ch^2\log^C n.$$

The essential supremum is over the conditional support of the auxiliary nodes at bandwidth h_i. The second inequality follows from the nodal coefficient bound and $Gh_{\max} \le C\ell_n(\log\log n)^{1/4}$. A categorical feature does not depend on h_i and has mean zero at every index. Its population cross blocks are zero and its component is represented exactly. Thus $g^*_{n,k} = f_j(x)$ for categorical values.

Linearization and Gaussian approximation. We linearize the reported values and approximate their maximum error by a Gaussian maximum. We define the one-observation score

$$u_{ni} = \mathbb{E}_U\{\phi_{ni}(Y_i - \phi_{ni}^{\mathsf{T}}\eta_n^*) \mid X_i, Y_i\}, \qquad \widehat{\Gamma}_n = \frac{1}{N}\sum_i M_{ni}, \qquad \bar{u}_n = \frac{1}{N}\sum_i u_{ni}.$$

Then $\widehat{\eta} - \eta_n^* = \widehat{\Gamma}_n^{-1}\bar{u}_n$. In population coordinates the empirical component center is $\bar{z}_{n,j}^{\mathsf{T}}\widehat{\eta}_j$. Here $\bar{z}_{n,j} = N^{-1}\sum_i z_{ni,j}$. The derivative of the reported component with respect to the weight of observation i is

$$\psi_{ni,k} = e_{n,k}^{\mathsf{T}}\Gamma_n^{-1}u_{ni} - z_{ni,j}^{\mathsf{T}}\eta_{n,j}^*, \qquad \psi_{ni}^0 = \psi_{ni} - \mathbb{E}\psi_{ni}. \tag{S49}$$

The individual means need not be zero, but their sum is zero by the pooled normal equation and centering identities. We construct D_n to have diagonal entries $\sqrt{h}$ for continuous reported values and one for categorical reported values. We write $\widehat{f}^{\mathrm{C}}$ and g_n^* for the vectors of these values. Matrix concentration on physical-time residue classes and the conditional Gaussian score bound give

$$\|\widehat{\Gamma}_n - \Gamma_n\|_{\mathrm{op}} + \|\bar{u}_n\|_2 = O_p\{\sqrt{G/n}\log^C n\}.$$

The same bound applies to the empirical feature means. We expand the exact normal equation and empirical center to give

$$\Big\|\sqrt{N}D_n(\widehat{f}^{\mathrm{C}} - g_n^*) - \frac{1}{\sqrt{N}}\sum_i D_n\psi_{ni}^0\Big\|_\infty = O_p\{G\log^C n/\sqrt{n}\} = o_p(1). \tag{S50}$$

Here each remainder term is a product of two of the preceding estimation errors, and the scaled evaluation row satisfies $(D_n)_{kk}\|e_{n,k}\|_2 \le C$. The continuous approximation bias is negligible because $\sqrt{Nh}\,h^2\log^C n = O(n^{-1/8}\log^C n)$.

We identify the design and noise parts of these observation terms. We write $D_n\psi_{ni}^0 = b_{ni}(X_i) + v_{ni}(X_i)\varepsilon_i$, with $\mathbb{E}b_{ni} = 0$. Uniformly in i, b_{ni} has mean squared distance $o(1)$ from the vector with continuous entries zero and categorical entries $-f_j(X_{ij})$. For continuous entries, the residual bound is $C\sqrt{G}h^2\log^C n = o(1)$ and centering is scaled by $\sqrt{h}$. For categorical value c, we define $p_c = P_j(\{c\})$ and $r_c(C) = \mathbf{1}\{C = c\}/p_c - 1$. The remaining design term is $r_c(X_{ij})\sum_{l\in S\setminus S_d}\{f_l(X_{il}) - g_{ni,l}(X_{il})\}$. Here $g_{ni,l}(X_{il}) = z_{ni,l}^{\mathsf{T}}\eta_{n,l}^*$. Coordinate independence centers this term, and the residual bound makes its second moment tend to zero. Redraw contraction makes the normalized sum of these remainders $o_p(1)$. The central limit argument for sums over independent redraw segments of Section S11 applies to the remaining categorical vector.

Conditional on the design, the noise sum is Gaussian. Its covariance is

$$Q_n = \frac{\sigma^2}{N}\sum_{i,l\in\mathcal{I}_n}\vartheta^{|i-l|}v_{ni}(X_i)v_{nl}(X_l)^{\mathsf{T}}.$$

The nodal and evaluation bounds give $\max_i\|v_{ni,k}\|_\infty \le Ch^{-1/2}\log^C n$ and $\max_i\mathbb{E}v_{ni,k}^4 \le Ch^{-1}\log^C n$. For a lag r, overlapping segments contribute at most $r+1$ covariance terms; the remaining terms decrease geometrically with their separation. We truncate the noise covariance sum at lag $C\log n$ and sum these bounds to obtain

$$\max_{k,l}|(Q_n - \mathbb{E}Q_n)_{kl}| = O_p\{\log^C n/\sqrt{nh}\} = o_p(1).$$

This calculation does not require $\mathbb{E}v_{ni} = 0$ at each index.

We define $\Omega_n = \mathrm{Cov}\{N^{-1/2}\sum_i D_n\psi_{ni}^0\}$. Conditional Jensen, the bounded eigenvalues of Γ_n and the upper autoregressive covariance bound imply $\max_k(\Omega_n)_{kk} \le C$. For a continuous group, we use $a_{n,j}(x)$ and $a_{i,j}(z)$ to denote the raw averaged features at bandwidths h and h_i. We define

$\overline{D}_{n,j} = N^{-1}\sum_i \mathbb{E}D_{ij}$ and $c_{ni,j} = \mathbb{E}a_{i,j}(X_{ij})$. Before centering, the local noise row of this group is

$$\ell_{ni,j}(x,z) = a_{n,j}(x)^{\mathsf{T}}\overline{D}_{n,j}^{-1}a_{i,j}(z), \qquad \ell_{ni,j}(x,z) = 0 \quad \text{if } |x-z| > h + h_i.$$

For retained $i \geq n/2$, the nodal mean calculation above gives

$$\int \ell_{ni,j}(x,z)\, dP_j(z) = a_{n,j}(x)^{\mathsf{T}}\overline{D}_{n,j}^{-1}c_{ni,j} = 1 + O(h).$$

Indeed, the averaged centering vector equals $\overline{D}_{n,j}e_j$. Here e_j represents the constant polynomial one. At interior nodes the individual mean differs from this average by $O(\Delta h^2)$ in the intercept coordinate and $O(\Delta h)$ in the scaled slope coordinate. The inverse nodal bound gives the remainder term of this identity. Cauchy–Schwarz and the upper density bound then give

$$\int \ell_{ni,j}(x,z)^2\, dP_j(z) \geq \frac{\{\int \ell_{ni,j}(x,z)\, dP_j(z)\}^2}{P_j\{|X_j - x| \leq Ch\}} \geq c/h.$$

Centering subtracts one from this one-group noise row. The subtracted term has a bounded squared integral, and the population cross-block correction is $o(h^{-1/2})$ in L^2. The lower autoregressive covariance bound hence gives a positive scaled variance.

For a categorical value, $\mathbb{E}r_c(X_{ij})^2 = p_c^{-1} - 1 > 0$ gives the same conclusion. Thus

$$0 < c \leq \min_k(\Omega_n)_{kk} \leq \max_k(\Omega_n)_{kk} \leq C.$$

The conditional Gaussian characteristic function and the design central limit theorem give a Gaussian approximation with covariance Ω_n. We apply the argument along further subsequences on which this bounded matrix converges. A union bound over the finitely many standardized Gaussian densities gives uniform anti-concentration, also for singular limits.

Exact block deletion. The exact block deletions are approximated by the same observation terms. We write n_b for the retained size of block b and J for the number of these blocks, and we define $L_{\min} = \min_b n_b$ and $L_{\max} = \max_b n_b$. Here $J = O(n^{2/5})$, $L_{\min} \geq 1$ and $L_{\max} = O(n^{3/5})$. We write $M_{n,b} = \sum_{i\in b} M_{ni}$ and $U_{n,b} = \sum_{i\in b} u_{ni}$. The exact normal-equation identity (S26) holds in these pooled population coordinates. The mean matrices are not identical across blocks:

$$\|\mathbb{E}M_{n,b} - n_b\Gamma_n\|_{\mathrm{op}} \leq Cn_b \log^C n.$$

For the centered parts of the block matrices, the residue-class matrix bound gives $\|M_{n,b} - \mathbb{E}M_{n,b}\|_{\mathrm{op}} \leq O_p\{(\sqrt{n_bG}+G)\log^C n\}$ uniformly in b. The conditional Gaussian score and bounded residual give $\|U_{n,b} - \mathbb{E}U_{n,b}\|_2 \leq O_p\{(\sqrt{n_bG}+\sqrt{G})\log^C n\}$. The deterministic block score and center means are at most $Cn_b\sqrt{G}h^2\log^C n$ in the normalized reported coordinates. All deleted matrices are invertible with probability tending to one because $L_{\max}/N \to 0$ and the complete matrix is stable.

We define $V_b = (N - n_b)(\widehat{f}^{\mathrm{C}} - \widehat{f}^{\mathrm{C}}_{-b})$ and $\bar{\psi}^0_n = N^{-1}\sum_i \psi^0_{ni}$. We substitute the preceding bounds and (S49) into the exact deletion identity, including the changed empirical centers, to give

$$\max_b \frac{1}{\sqrt{n_b}}\Big\|D_nV_b - \sum_{i\in b} D_n(\psi^0_{ni} - \bar{\psi}^0_n)\Big\|_\infty = O_p\Big[\log^C n\Big\{\frac{G}{\sqrt{n}} + \frac{G^{3/2}}{\sqrt{nL_{\min}}} + \sqrt{\frac{L_{\max}G}{n}}\Big\}\Big] = o_p(1).$$

The last term includes the products of the deterministic matrix differences and the complete coefficient error, and the deterministic means of the block scores, using $h^2 \asymp n^{-1/2}$. The corresponding exponents are at most $-1/4$, $-1/8$ and $-3/40$, and these exponents also cover partial blocks of smaller retained size. The same bound holds for $\widetilde{V}_b = V_b - (n_b/N)\sum_a V_a$ in (S6).

Covariance and resampling. We establish covariance consistency and the conditional resampling approximation. We define $S_b = \sum_{i\in b} D_n\psi^0_{ni}$. For a retained consecutive segment I of size a, conditional Gaussian fourth moments and the redraw covariance contraction give

$$\max_k \mathbb{E}\Big|\sum_{i\in I}(D_n\psi^0_{ni})_k\Big|^2 \leq Ca\log^C n, \qquad \max_k \mathbb{E}\Big|\sum_{i\in I}(D_n\psi^0_{ni})_k\Big|^4 \leq C\log^C n(a^2 + a/h).$$

Equation (S24) applies to these deterministic index-dependent rows. The omitted cross-block covariances are at most $CJ\log^C n/N = o(1)$. We remove $C\log n$ observations per complete block, and the second-moment bound and $J\log^C n/N \to 0$ show that this removal changes the covariance by a negligible amount. We couple the remaining joint segments independently as in Section S11. A short reporting block is omitted from this sum, and the second-moment bound restores its contribution. We write S_b° for the coupled retained sum. For each covariance entry,

$$\mathbb{E}\Big|\frac{1}{N}\sum_b\{S_{b,k}^\circ S_{b,l}^\circ - \mathbb{E}(S_{b,k}^\circ S_{b,l}^\circ)\}\Big|^2 \le \frac{C\log^C n}{N^2}\sum_b(n_b^2 + n_b/h)$$
$$\le C\log^C n\left\{\frac{L_{\max}}{N} + \frac{1}{Nh}\right\} = o(1).$$

Restoring the removed observations and subtracting the size-weighted sample mean preserve convergence. The exact-deletion approximation gives

$$ND_n\widehat{\Sigma}D_n - \Omega_n \longrightarrow_p 0.$$

For the resampling step, the centered vectors $\sqrt{J/N}\,D_n\widetilde{V}_b$ are sampled independently and uniformly over $1 \le b \le J$. Their conditional covariance converges to Ω_n. Their conditional fourth moment divided by J is bounded by

$$\frac{1}{N^2}\sum_b\|D_n\widetilde{V}_b\|_2^4 = O_p\left[\log^C n\left\{\frac{L_{\max}}{N} + \frac{1}{Nh}\right\}\right] + o_p(1) = o_p(1).$$

The conditional central limit theorem and consistent resampling standard deviations give the same Gaussian approximation for both studentized maxima. Equation (S50), the bias bound and the anti-concentration bound give the oracle coverage $1 - \alpha + o(1)$. The support error probability after stabilization tends to zero and bounds the coverage difference for the selected procedure. This completes the proof.

S15.4 Proof of Proposition S6

We count the scalars kept at time n: the retained selection observations, the moments of each active resolution and the deleted-block inference states. By Theorem 4, all active states have support S with probability tending to one. The $n^{3/4}\{1 + o(1)\}$ selection observations use $O(dn^{3/4})$ scalars; the incomplete block adds $O(dn^{3/5})$. Active resolutions have first reporting time at most $Cn\log\log n$. This yields largest grid sizes of orders

$$\ell_{n\log\log n}(n\log\log n)^{1/5}, \qquad \ell_{n\log\log n}(n\log\log n)^{1/4}.$$

The squared dimensions of the nested grids sum to at most a constant times the largest squared dimension. Point estimation gives the second term of Proposition S6; one inference state per $O(n^{2/5})$ blocks gives the third. Remaining summaries have smaller orders. This completes the proof.

S16 Lower bounds under bounded total signal

The minimum component norm and the number of active effects cannot vary independently when total signal is bounded. We construct alternatives that satisfy both the minimum-norm and the total-signal restrictions and use the population component loss in Theorem S1.

We fix $L, B, \sigma > 0$ and use $b_N > 0$ to bound nonzero component norms from below. We define $\mathcal{F}_U$ to consist of additive functions with at most s active components, each centered on $[0,1]$ and satisfying $\max_{0\le r\le 2}\|f_j^{(r)}\|_\infty \le L$. The observations are independent, the design coordinates are independent and uniform on $[0,1]$, and the errors are independent $N(0,\sigma^2)$ variables independent of the design. We write $S(f) = \{j : f_j \not\equiv 0\}$ for the active set of f and define

$$\mathcal{F}_U^{b_N,B} = \Big\{f \in \mathcal{F}_U : |S(f)| = s, \quad \min_{j\in S(f)}\|f_j\|_{L^2[0,1]} \ge b_N, \quad \sum_{j=1}^d\|f_j\|_\infty^2 \le B^2,$$
$$f_j(0) = f_j(1) = f_j'(0) = f_j'(1) = 0 \quad \text{for every } j\Big\}.$$

The boundary restrictions specify a comparison class for the lower bound; they are not imposed on the estimation model. We define $\varrho_N = \{\log(esN^{1/5})/N\}^{2/5}$. We write $\mathbb{P}_f$ and $\mathbb{E}_f$ for probability and expectation under the observations generated by f.

The following theorem establishes an integrated component-risk lower bound when the minimum signal and sparsity satisfy the bounded-signal restriction.

Theorem S6. *There exist $N_\star < \infty$ and $c_0, c > 0$, depending only on L, B, and σ, such that, if*

$$N \geq N_\star, \qquad 1 \leq s \leq d, \qquad \sqrt{s}(b_N + \varrho_N) \leq c_0,$$

then

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_U^{b_N, B}} \mathbb{P}_f \Big\{ \sum_{j=1}^{d} \left\| \widehat{f}_j - f_j \right\|_{L^2[0,1]}^2 \geq csN^{-4/5} \Big\} \geq \frac{1}{8}$$

Theorem S6 shows that the oracle rate remains a lower bound on component risk when total signal is bounded and sparsity satisfies the stated restriction. Multiplying the loss threshold by its probability lower bound gives the corresponding expected-risk lower bound. The comparison concerns component estimation and not optimality of confidence-interval length. Corollary 1 supplies the matching upper bound on the intersection of the comparison class with the selection class of that corollary.

Proof. We construct alternatives in $\mathcal{F}_U^{b_N, B}$ by adding narrow bumps to a common base component and bound the probability of distinguishing them. We choose fixed nonzero smooth functions g and φ with supports contained in $(0, 1/4)$ and $(-1/2, 1/2)$, respectively, such that

$$\int g = \int \varphi = 0, \qquad \max_{0 \leq r \leq 2} \left\| g^{(r)} \right\|_\infty \leq 1, \qquad \max_{0 \leq r \leq 2} \left\| \varphi^{(r)} \right\|_\infty \leq 1.$$

Such functions are obtained by differentiating smooth functions of compact support and multiplying by fixed constants. We define $\nu_0 = \|g\|_{L^2[0,1]} > 0$ and $\bar{f}_N = (b_N/\nu_0)g$. For $h \leq 1/16$, we set $m = \lfloor 1/(8h) \rfloor$, $x_\ell = 1/2 + (2\ell - 1)h$ for $1 \leq \ell \leq m$, and

$$\psi_\ell(x) = ah^2 \varphi\{(x - x_\ell)/h\}.$$

Here $a > 0$ is a fixed sufficiently small constant. The supports of the ψ_ℓ are pairwise disjoint, lie in $[1/2, 3/4]$, and do not intersect the support of g.

For $S_\star = \{1, \ldots, s\}$ and $\theta \in \{0, 1\}^{sm}$, we define

$$f_{\theta,j} = \bar{f}_N + \sum_{\ell=1}^{m} \theta_{j\ell} \psi_\ell \quad (j \in S_\star), \qquad f_{\theta,j} = 0 \quad (j \notin S_\star).$$

Disjoint supports and centering give

$$\int_0^1 f_{\theta,j} = 0, \qquad \|f_{\theta,j}\|_{L^2[0,1]}^2 = b_N^2 + \sum_{\ell=1}^{m} \theta_{j\ell} \|\psi_\ell\|_{L^2[0,1]}^2 \geq b_N^2 \quad (j \in S_\star).$$

For $h = \kappa_0 N^{-1/5}$ with a sufficiently small fixed $\kappa_0 > 0$, we have $h^2 \leq \varrho_N$. The conditions $\sqrt{s}(b_N + \varrho_N) \leq c_0$ and $h \leq 1$ imply

$$\max_{j,\, 0 \leq r \leq 2} \left\| f_{\theta,j}^{(r)} \right\|_\infty \leq b_N/\nu_0 + a \leq L, \qquad \sum_j \|f_{\theta,j}\|_\infty^2 \leq Cs(b_N^2 + h^4) \leq B^2,$$

provided that a and c_0 are chosen sufficiently small. Every component and its first derivative equal zero at the two boundaries. Thus, all the alternatives belong to $\mathcal{F}_U^{b_N, B}$.

We use P_θ to denote the joint distribution of the observations under f_θ and define $D = sm$. We write $D_{\mathrm{KL}}(P, Q) = \int \log(dP/dQ)\, dP$ for the Kullback–Leibler divergence when P is absolutely continuous with respect to Q. If θ and θ' differ in one coordinate, the common design distribution and the uniform marginal of X_{ij} give

$$D_{\mathrm{KL}}(P_\theta, P_{\theta'}) = \frac{N}{2\sigma^2} \|\psi_\ell\|_{L^2[0,1]}^2 = \frac{Na^2h^5 \|\varphi\|_2^2}{2\sigma^2} \leq \frac{1}{2}.$$

Pinsker's inequality bounds the total variation distance of this pair by $1/2$; see Tsybakov (2009, Chapter 2) for the testing inequalities used here.

For any estimator, we classify the presence of each bump and count the incorrect decisions. For an estimator $\widehat{f}$, we define

$$\widehat{\theta}_{j\ell} = \mathbb{1}\left\{\int_0^1 (\widehat{f}_j - \bar{f}_N)\psi_\ell \geq \tfrac{1}{2}\,\|\psi_\ell\|^2_{L^2[0,1]}\right\}, \qquad H = \sum_{j=1}^{s}\sum_{\ell=1}^{m} \mathbb{1}\{\widehat{\theta}_{j\ell} \neq \theta_{j\ell}\}.$$

Pairing the alternatives that differ in one coordinate yields

$$2^{-D}\sum_{\theta} \mathbb{E}_\theta H \geq D/4.$$

Because $0 \leq H \leq D$, the average on the left-hand side of this inequality is at most $D/8 + D\,2^{-D}\sum_\theta \mathbb{P}_\theta(H \geq D/8)$. Hence $2^{-D}\sum_\theta \mathbb{P}_\theta(H \geq D/8) \geq 1/8$.

A wrong decision costs a fixed fraction of the squared norm of the corresponding bump. On a coordinate for which $\widehat{\theta}_{j\ell} \neq \theta_{j\ell}$, the Cauchy–Schwarz inequality gives

$$\int_{\mathrm{supp}(\psi_\ell)} (\widehat{f}_j - f_{\theta,j})^2 \geq \tfrac{1}{4}\,\|\psi_\ell\|^2_{L^2[0,1]}\,.$$

Summing over the disjoint bump supports gives

$$\sum_j \left\|\widehat{f}_j - f_{\theta,j}\right\|^2_{L^2[0,1]} \geq \tfrac{1}{4}a^2h^5\,\|\varphi\|_2^2\,H.$$

On $\{H \geq D/8\}$, the right-hand side is at least csh^4 because $m \geq 1/(16h)$. Taking the supremum over the alternatives and then the infimum over estimators proves the probability bound.

It remains to check that these alternatives satisfy the total-signal, smoothness, and stability restrictions relevant to the comparison. For independent uniform coordinates, centering and Hoeffding's inequality give, for every real t,

$$\mathbb{E}\exp\Big\{t\sum_j f_j(X_{ij})\Big\} \leq \exp(t^2B^2/2),$$

$$\mathbb{E}\Big[\exp\Big\{t\sum_k f_k(X_{ik})\Big\} \mid X_{ij}\Big] \leq \exp(|t|B + t^2B^2/2).$$

These bounds give the sub-Gaussian signal control for selection. Coordinate independence makes the population quadratic the identity, and all continuous densities equal one. These alternatives satisfy the conditions of Theorem S1. To verify the minimax comparison on the common class, we repeat the construction with $\widetilde{b}_N = 2c_0/\sqrt{s}$ in place of b_N in the common base function. Because $\sqrt{s}b_N \leq c_0$, every alternative still has component norms at least b_N. We choose $a \leq L/2$ and c_0 sufficiently small that $2c_0/\nu_0 + a \leq L$ and $(4/\nu_0^2 + a^2)c_0^2 \leq B^2$. The disjoint supports and $sh^4 \leq c_0^2$ preserve every restriction in $\mathcal{F}_U^{b_N,B}$. With $N \asymp n$, we define $\overline{\lambda}_n = c_\lambda(\sigma^2 + B^2)^{1/2}\{\log d/(N_L h_L)\}^{1/2}$. Coordinate independence gives $\sigma_Y^2 \leq \sigma^2 + B^2$ uniformly and hence $\lambda_n^0 \leq \overline{\lambda}_n$. We write m_{sel} for the number of screening candidates in the selection corollary. Under its growing-dimensional conditions,

$$s(h_L^2 + \overline{\lambda}_n)^2 \leq C m_{\mathrm{sel}} n^{-4/5}\log(dn) \leq C\frac{m_{\mathrm{sel}}^2 G_L \log^3(dn)}{n} = o(1).$$

The minimum component norm is at least $2c_0/\sqrt{s}$, which gives the required uniform signal-to-penalty separation. For fixed d, this minimum norm is bounded away from zero. The fixed-separation condition holds for sufficiently large n when the penalty multiplier satisfies Corollary S2. The common base cancels from every difference of two alternatives. The Kullback–Leibler and testing bounds are unchanged, proving the matching lower bound on the intersection with the applicable selection class. This completes the proof. □

S17 Proofs of the selection identity and component derivative

Both proofs use the empirical quadratic criterion: the selection identity controls the inactive scores, and the component derivative includes the centering constraints.

S17.1 Proof of Proposition S1

Proof. We define $\delta_S = \widehat{\theta}^L_S - \theta^*_S$ for the Lasso solution restricted to S. The optimality equations of this restricted solution give

$$H_{SS}\delta_S = e_S - \lambda u_S, \qquad \|u_S\|_{2,\infty} \le 1, \qquad \|\delta_S\|_{2,\infty} \le \frac{\lambda/4 + \lambda}{1 - 1/4} = \frac{5}{3}\lambda.$$

For $j \notin S$, the restricted score norm is at most $\lambda/4 + (1/4)(5\lambda/3) = 2\lambda/3 < \lambda$. All Lasso equations then hold for the restricted solution, and the active norms exceed $a\lambda$. This active margin gives the stated SCAD weights. For the restricted unpenalized fit, $\delta^o_S = H^{-1}_{SS} e_S$ gives $\|\delta^o_S\|_{2,\infty} \le \lambda/3$ and inactive score at most $\lambda/4 + \lambda/12 = \lambda/3$. These bounds show that the restricted unpenalized fit satisfies the weighted equations. The row bound gives $\theta^\mathsf{T} H \theta \ge 3\|\theta\|_2^2/4$, proving uniqueness. This completes the proof. □

S17.2 Proof of Proposition S3

Proof. We differentiate the normal equation and the empirical centering constraint. We use c_j to represent the constant one in the estimable group space. Then $\widehat{D}_j c_j = \widehat{c}_j$ and $a_{ij}^\mathsf{T} c_j = 1$. Multiplying the group equations by c_j^T gives the intercept equation; the centering multipliers are zero. We stack a_{ij} in a_i; we construct M_i to have diagonal blocks D_{ij} and cross blocks $a_{ij}a_{ik}^\mathsf{T}$, and we define $\widehat{M} = N^{-1}\sum_i M_i$. Differentiating the raw normal equation gives

$$\widehat{M}\dot{\beta}_i = a_i(Y_i - \bar{Y}) - M_i\widehat{\beta} - \widehat{c}(Y_i - \bar{Y}), \qquad \widehat{c} = N^{-1}\sum_i a_i.$$

Differentiating the centering constraint gives

$$\widehat{c}_j^\mathsf{T}\dot{\beta}_{j,i} = -\widehat{f}_{ij}.$$

We write $\dot{\beta}_{j,i} = R_j v_{j,i} - c_j\widehat{f}_{ij}$. The raw matrix maps group constants to stacked group means, which are removed by premultiplication with R_j^T. Substitution into the differentiated normal equation gives $Hv_i = s_i$. Multiplication by $a_j(x)^\mathsf{T}$ proves the first equality in (S5). The score equations $N^{-1}\sum_i s_i = 0$ and the zero empirical component means give the second equality. This completes the proof. □

S18 Variation from Empirical Centering

The following exact-fit example isolates component error from empirical centering.

S18.1 An Exact Linear-Component Calculation

We assume X_i has density $p(x) = 1/2 + x$ and follows the independent-redraw transition. We consider the noiseless response $Y_i = a(X_i - \mu_X)$, where $\mu_X = 7/12$ and $a \neq 0$ is fixed. To isolate uncertainty about the center, we first suppose the affine shape is known and only its center is estimated. We write P_X for the marginal distribution and $\bar{X} = N^{-1}\sum_{i=1}^N X_i$ for the sample mean. We define the intercept and component estimates for every design sample by

$$\widehat{\alpha} = a(\bar{X} - \mu_X), \qquad \widehat{f}(x) = a(x - \bar{X}).$$

On a full-rank local-linear representation, the unpenalized criterion reproduces this affine response and gives these same values. The known-shape definition also specifies these estimates on singular samples, which occur with positive probability under the redraw model. The fitted regression function equals the true conditional mean at every x, but the component error is the constant $a(\mu_X - \bar{X})$. The population integrated squared error of the component equals $a^2(\bar{X} - \mu_X)^2$.

The second moment of X is $5/12$, and hence $\mathrm{Var}(X) = 11/144$. The redraw process has covariance $\rho^k \mathrm{Var}(X)$ at lag k. The exact expected component loss is

$$\mathbb{E}\|\widehat{f} - f\|^2_{P_X} = \frac{11a^2}{144N}\Big\{1 + 2\sum_{k=1}^{N-1}(1 - k/N)\rho^k\Big\}. \tag{S51}$$

The ratio of this expected loss to $11a^2(1+\rho)/\{144N(1-\rho)\}$ tends to one as N increases. The conditional prediction error is zero in the same experiment. These two facts do not conflict because the component and the fitted conditional mean are different statistical targets.

In the case-weight derivative, the residual and integration-variation scores are zero for an exactly reproduced noiseless affine response. Proposition S3 then gives

$$\dot{f}_i(x) = -a(X_i - \bar{X}).$$

The only contribution to this derivative is empirical centering, which also gives the nonzero component risk in (S51). Holding the component center fixed during an uncertainty calculation would discard all variation in this example.

The same distinction occurs for a categorical component. If C_i has marginal distribution P_C, fixed level probabilities p_c and a known centered shape $g(c)$, its empirically recentered version is $\widehat{g}(c) = g(c) - N^{-1}\sum_i g(C_i)$. With the redraw transition,

$$\mathbb{E}\|\widehat{g} - g\|^2_{P_C} = \frac{\mathrm{Var}\{g(C)\}}{N}\Big\{1 + 2\sum_{k=1}^{N-1}(1 - k/N)\rho^k\Big\}.$$

Every contrast $\widehat{g}(c) - \widehat{g}(c')$ is exact in this known-shape experiment, although the centered levels themselves are not. The numerical centering comparison in Section S26 uses noisy versions of these models and estimates the support without providing it to the procedure.

The case-weight derivative in Proposition S3 includes variation of the coefficients, their empirical centers and integration over local grid points. Exact block deletion recomputes all three. The inference fit is untruncated; the bounded point estimate used for expected risk is defined in Section S4 and need not remain empirically centered after truncation.

S19 Categorical levels and contrasts

We derive uncertainty for categorical levels and their contrasts. A level includes its estimated center; a difference of levels removes it.

S19.1 A centered level and a difference of levels

We take j to be a categorical active group with probabilities $p_c = P_j(\{c\})$, $c = 0, \ldots, K_j - 1$. The number of active groups and K_j are fixed. The final block contains N observations. Continuous groups use the inference bandwidth and population model from Theorem S3. We write ε_i for the structural regression error and define

$$r_c(C) = \frac{\mathbb{1}(C = c)}{p_c} - 1.$$

The quantity r_c is a population contrast used in the proof; the algorithm does not require the probabilities p_c.

The following proposition gives the influence function of a centered categorical level and of a difference of two levels.

Proposition S7. *Suppose the conditions of Theorem S3 hold and j is a categorical active group. Then*

$$\sqrt{N}\{\widehat{f}_j(c) - f_j(c)\} = N^{-1/2}\sum_{i=1}^{N}\{r_c(X_{ij})\varepsilon_i - f_j(X_{ij})\} + o_p(1).$$

For $c \neq c'$,

$$\sqrt{N}\{[\widehat{f}_j(c) - \widehat{f}_j(c')] - [f_j(c) - f_j(c')]\} = N^{-1/2}\sum_{i=1}^{N}\{r_c(X_{ij}) - r_{c'}(X_{ij})\}\varepsilon_i + o_p(1).$$

In Proposition S7, omitting the center term discards uncertainty for a level. Its cancellation from a level difference is exact.

Proof. For a categorical group, $D_j(x) = a_j(x)a_j(x)^\mathsf{T}$, and its local integration-variation score equals zero exactly. In population coordinates the evaluation kernel is $k_j(c, X_j) = \mathbb{1}(X_j = c)/p_c$. The population projected categorical function equals f_j. Substituting these identities into (S20) gives the population influence

$$r_c(X_{ij})\varepsilon_i - f_j(X_{ij}) + r_c(X_{ij}) \sum_{\substack{k \in S \\ k \text{ continuous}}} \{f_k(X_{ik}) - g_k^*(X_{ik})\}.$$

The last summand has mean zero. Population independence across coordinates, bounded r_c, and the approximation bound give its variance at most Csh^4. Its lag covariance decreases as ρ^l under the redraw transition. Its normalized sum is thus $O_p(\sqrt{s}\, h^2) = o_p(1)$ for fixed s.

By (S25), the categorical inverse and centering remainders are $o_p(N^{-1/2})$, proving the first assertion. Subtracting the two expansions cancels $-f_j(X_{ij})$, giving the second assertion. Selection error tends to zero and changes neither limit. This completes the proof. □

S19.2 Long-run variance under the redraw model

We define $\tau_j^2 = \mathbb{E} f_j(X_j)^2$. Independence between the errors and the design, and the identities

$$\mathbb{E}\{r_c(X_j)r_{c'}(X_j)\} = \frac{\mathbb{1}(c = c')}{p_c} - 1, \qquad \mathbb{E}\{r_c(X_j) - r_{c'}(X_j)\}^2 = p_c^{-1} + p_{c'}^{-1}$$

give the limiting covariance of two centered levels:

$$\Omega_{cc'} = \sigma^2 \frac{1 + \rho\vartheta}{1 - \rho\vartheta} \left\{ \frac{\mathbb{1}(c = c')}{p_c} - 1 \right\} + \frac{1 + \rho}{1 - \rho} \tau_j^2. \tag{S52}$$

The signal term is common to every pair of levels because each level estimate includes the same random centering shift. This term is not diagonal measurement noise. For a difference between two distinct levels, the variance is

$$\sigma^2 \frac{1 + \rho\vartheta}{1 - \rho\vartheta} (p_c^{-1} + p_{c'}^{-1}). \tag{S53}$$

Equations (S52) and (S53) follow by summing the geometric covariances of the influence in Proposition S7. The deleted-block procedure does not require estimating either temporal parameter in these formulas.

If the errors are independent over time, that is, $\vartheta = 0$, persistence of the covariate still increases the centering variance by the factor $(1 + \rho)/(1 - \rho)$, while the error contribution is unchanged. This implies that adjusting only a residual error autocorrelation does not account for all component uncertainty in this experiment.

S19.3 Joint contrasts from the same block summaries

We define $F = (\widehat{f}_j(0), \ldots, \widehat{f}_j(K_j - 1))^\mathsf{T}$ and take A to be any fixed matrix of level contrasts, with each row summing to zero. We apply A to F and to every deleted-block effect V_b. The resulting covariance estimate of the contrasts is exactly $A\widehat{\Sigma}A^\mathsf{T}$. The resampling maximum is computed for these contrasts with their own standard errors; subtracting two marginal interval limits is not the same procedure.

For all pairwise hour differences, A has $K_j(K_j - 1)/2$ rows, which for $K_j = 24$ gives 276 contrasts. A single maximum over these contrasts accounts for selecting a pair after viewing the hour curve. The same moment summaries and the same deleted fits suffice. The application remains nominal for the calendar series, as stated in Section S23; forming a contrast does not make its deterministic clock sequence satisfy the dependent population model.

S20 Sequential computation and identifiable evaluations

We specify moment reconstruction, numerical tolerances and identifiable evaluations, including the separate convention for unsupported predictions.

S20.1 Raw moments and their updates

For a fixed set of groups and resolution, the following moments determine the criterion. For the observations in a block I, we write a_i for the vector that stacks over groups the kernel-weighted average of the local-line basis at X_{ij}. We write $D_j(X_{ij})$ for the corresponding weighted second-moment matrix. We store

$$n_I = |I|, \quad y_I = \sum_{i\in I} Y_i, \quad y_{I,2} = \sum_{i\in I} Y_i^2, \quad a_I = \sum_{i\in I} a_i, \quad b_I = \sum_{i\in I} a_i Y_i,$$

$$Q_I = \sum_{i\in I} a_i a_i^{\mathsf{T}}, \qquad D_{I,j} = \sum_{i\in I} D_j(X_{ij}).$$

Each moment is additive over disjoint blocks, giving the same criterion for every division into chunks. The matrices $D_{I,j}$ are block diagonal in grid nodes and require storage linear in the node count at fixed polynomial degree. They retain the variation of local predictions across grid nodes. The within-group blocks of Q_I use only the averaged predictions and would lose this variation if substituted for $D_{I,j}$.

To reconstruct the criterion, we set $\bar{Y} = y_I/n_I$, $\widehat{c} = a_I/n_I$, and form the raw quadratic matrix with diagonal blocks $D_{I,j}/n_I$ and off-diagonal blocks from Q_I/n_I. The raw linear term after eliminating the intercept is $b_I/n_I - \widehat{c}\bar{Y}$. The equality $\widehat{c}_j^{\mathsf{T}}\beta_j = 0$ makes the derivative with respect to the intercept equal to $\alpha - \bar{Y}$. No separate estimate of a marginal density is substituted into a cross block.

Marginal screening fits one group at a time and hence needs $D_{I,j}, a_{I,j}, b_{I,j}$ but not Q_I across different groups. If R_j is the centered normalization of the moment matrix $D_{I,j}/n_I$ of that group, the quadratic objective of marginal screening for that group is

$$\tfrac{1}{2}\|\theta_j\|_2^2 - \theta_j^{\mathsf{T}} R_j^{\mathsf{T}}\{b_{I,j}/n_I - \widehat{c}_j\bar{Y}\}.$$

The minimum of this objective is minus one half of the squared norm of the normalized linear term. This proves the marginal-score formula used by the screening state. The response standard deviation for the penalty is computed from

$$\widehat{\sigma}_Y^2 = \frac{y_{I,2} - y_I^2/n_I}{n_I - 1}.$$

For numerical stability, the implementation subtracts a fixed response origin before accumulating response moments and products with the design. When blocks use different origins, their moments are expressed at a common origin before addition or subtraction. This leaves the criterion and centering unchanged in exact arithmetic. The intercept adds the response origin back. The screening stage fixes this penalty scale before the selection block is processed.

S20.2 Centering in a positive quadratic form

We construct the centering and normalization matrix R_j from the stored group moments. We write D for a group moment matrix and c for its empirical mean vector. We write $D = U\Lambda U^{\mathsf{T}}$ using its positive eigenvalues and their eigenvectors. Directions with zero eigenvalues do not change fitted values at the observed covariates. We set $W = U\Lambda^{-1/2}$ to normalize the quadratic form. We impose empirical centering by restricting the remaining coordinates to the nullspace of $c^{\mathsf{T}}W$. If B has orthonormal columns spanning this nullspace, then

$$R = WB, \qquad R^{\mathsf{T}}DR = I, \qquad c^{\mathsf{T}}R = 0.$$

Thus the norm in the penalized criterion becomes an ordinary Euclidean group norm. The basis B is not unique. Replacing R by RO with an orthogonal matrix O rotates the coefficients to $O^{\mathsf{T}}\theta$ and leaves the fitted component, criterion and group norm unchanged.

For the continuous representation, the raw constant vector e has entry one in each local-intercept position and zero in every slope position. For a categorical representation it is the vector of ones. The one-observation identities are

$$D_j(x)e = a_j(x), \qquad a_j(x)^{\mathsf{T}}e = 1.$$

After averaging, $De = c$ and $c^{\mathsf{T}}e = 1$. These identities justify transferring a component constant to the intercept. On an exact zero-norm quotient they remain valid for the projected representative of e.

The numerical implementation of this construction discards eigenvalues below its stated decreasing tolerance. If a direction with a positive eigenvalue below the tolerance is discarded, the numerical problem is the quadratic criterion restricted to the retained subspace. The population matrix bounds prove that this restriction removes no positive population direction on the high-probability samples used in the theorems. Derivative formulas for an exact fixed-rank quotient are not a justification for differentiating through an arbitrary rank threshold. The deleted-block implementation solves each restricted problem directly and checks the resulting evaluations.

S20.3 Coordinate minimization and a stopping criterion

With the diagonal normalization $H_{jj} = I$, fixing the other groups reduces the objective for group j to

$$\tfrac{1}{2}\|\theta_j - z_j\|_2^2 + \omega_j\|\theta_j\|_2, \qquad z_j = b_j - \sum_{k\neq j} H_{jk}\theta_k.$$

The unique minimizer is zero if $\|z_j\|_2 \le \omega_j$, and equals $(1 - \omega_j/\|z_j\|_2)z_j$ when $\|z_j\|_2 > \omega_j$. This follows by writing a nonzero candidate as ru with u a unit vector, minimizing over the direction, which gives $u = z_j/\|z_j\|_2$, and then minimizing $r^2/2 - r\|z_j\|_2 + \omega_j r$ over $r \ge 0$.

The algorithm maintains the residual vector $b - H\theta$. If a coordinate update changes block j by Δ_j, we update this residual vector by subtracting $H_{\cdot j}\Delta_j$. This is an exact matrix identity and avoids repeated multiplication by the full coefficient vector.

The stopping rule measures, group by group, how far the current coefficients are from satisfying the optimality condition of this criterion. For a nonzero group, the stationarity residual is

$$(H\theta - b)_j + \omega_j\theta_j/\|\theta_j\|_2.$$

For a zero group the stationarity violation is $\{\|(H\theta - b)_j\|_2 - \omega_j\}_+$. The maximum of these violations is the reported optimality error. The primary selection calculation stops when this error is at most N_L^{-3} and allows at most $100 + \lceil 30 \log N_L \rceil$ sweeps. A nonconverged calculation is recorded as a failure; it is not replaced by a different simulated sample.

Let the initial maximum block error be E and off-diagonal row sums be at most $\gamma < 1$. Because group thresholding is nonexpansive, the first updated block error is at most γE. The same bound then propagates through the sweep: the blocks already updated have errors at most γE and the blocks not yet updated have errors at most E, which bounds each new error by γE. A complete sweep contracts the maximum block error by the factor γ. For $\gamma \le 1/4$, the prescribed sweep limit attains the tolerances in Theorem S2.

For the final unpenalized equation, the implementation first checks the smallest eigenvalue and then solves the positive definite system. It does not compute a matrix inverse to multiply the response score. The inverse appearing in theoretical formulas denotes the corresponding linear map.

A small reported optimality error also gives a deterministic support guarantee. For the normalized quadratic criterion on m groups, let θ° be its unique exact minimizer and put $S^\circ = \mathrm{supp}(\theta^\circ)$. Define $\beta = \min_{j\in S^\circ}\|\theta_j^\circ\|_2$ and $\delta = \min_{j\notin S^\circ}\{\omega_j - \|(H\theta^\circ - b)_j\|_2\}$. A minimum over an empty set is $+\infty$ here. Let $\widetilde{\theta}$ have maximum group optimality error at most ϵ according to the preceding criterion. The following proposition specifies the required accuracy and does not assume that a fixed number of sweeps attains this accuracy.

Proposition S8. *If $H \succeq \kappa I$, $H_{jj} = I$, $\kappa > 0$, and*

$$\frac{\sqrt{m}\,\epsilon}{\kappa} < \beta, \qquad \epsilon\{1 + \sqrt{m/\kappa}\} < \delta,$$

then

$$\mathrm{supp}(\widetilde{\theta}) = S^\circ, \qquad \|\widetilde{\theta} - \theta^\circ\|_2 \le \frac{\sqrt{m}\,\epsilon}{\kappa}.$$

Proposition S8 shows that the numerical solution and the exact minimizer select the same groups when the optimization error is small relative to the active norms and inactive score margins. The selection proof establishes these quantities; they are not additional assumptions on a random sample. The proposition applies once the computed optimality error satisfies the two stated bounds.

Proof. By the definition of the optimality error, we choose a subgradient v of the criterion at $\widetilde{\theta}$ with $\|v_j\|_2 \le \epsilon$ for all groups, including the zero groups. Put $e = \widetilde{\theta} - \theta^\circ$. Monotonicity of the penalty subgradient gives

$$\kappa\|e\|_2^2 \le e^\mathsf{T} H e \le v^\mathsf{T} e \le \sqrt{m}\,\epsilon\|e\|_2.$$

It follows that $\|e\|_2 \le \sqrt{m}\,\epsilon/\kappa$ and $(e^\mathsf{T} H e)^{1/2} \le \sqrt{m/\kappa}\,\epsilon$. Because $H_{jj} = I$, the Cauchy–Schwarz inequality in the quadratic form gives $\|(He)_j\|_2 \le (e^\mathsf{T} H e)^{1/2}$. The first strict inequality in the proposition implies that every group in S° is retained. If $j \notin S^\circ$ but $\widetilde{\theta}_j \ne 0$, then

$$\epsilon \ge \|v_j\|_2 \ge \omega_j - \|(H\widetilde{\theta} - b)_j\|_2 \ge \delta - \sqrt{m/\kappa}\,\epsilon > \epsilon,$$

a contradiction. Thus no inactive group is retained. This completes the proof. □

In Theorem 1, the margins are of order λ and the sample eigenvalue is bounded below. For $N_L \asymp n$, $\epsilon = N_L^{-3}$ satisfies $\sqrt{m}\,\epsilon = o(\lambda)$ under $mr_n = o(\lambda)$. The initial coefficient error is $O(\sqrt{m}\,\epsilon)$, and the Lipschitz constant $1/(a-1)$ of the SCAD derivative gives a Euclidean weight error $o(\lambda)$. The restricted argument of Section S14 preserves the margins. The numerical SCAD solution has the stated support once the group-Lasso minimization and the weighted minimization both attain the tolerance. This argument does not establish that the sweep limit suffices for every matrix in that theorem.

Indeed, positive definiteness alone is insufficient for this fixed sweep limit. Consider two scalar groups with equal weights $\omega > 0$, matrix $H = \left(\begin{smallmatrix} 1 & r \\ r & 1 \end{smallmatrix}\right)$, $0 < r < 1$, and score $b = H(1,1)^\mathsf{T} + \omega(1,1)^\mathsf{T}$. Starting at zero, cyclic coordinate minimization gives positive coordinates after every sweep, and $\theta^{(K)}$ denotes the coefficient pair after K sweeps. The corresponding optimality residual is $v^{(K)} = H\theta^{(K)} - b + \omega(1,1)^\mathsf{T}$. After K sweeps,

$$\theta_1^{(K)} - 1 = r^{2K-1}, \qquad \theta_2^{(K)} - 1 = -r^{2K}, \qquad \|v^{(K)}\|_\infty = (1-r^2)r^{2K-1}.$$

For $K = 100 + \lceil 30 \log N_L \rceil$, this error has order $N_L^{60 \log r}$. If $r > \exp(-1/20)$, this error is not $O(N_L^{-3})$, although $\lambda_{\min}(H) = 1 - r > 0$ is fixed. Thus an unqualified sweep-limit guarantee cannot follow from the general eigenvalue condition. The stronger contraction bounds proved for the product model and for the specified three-group models remain valid.

S20.4 One pass through the three phases

We write n for the planned epoch size and set $g = \lceil 4 \log n \rceil$. We define the phase sizes by

$$n_0 = \lfloor 0.25(n - 2g) \rfloor, \quad n_L = \lfloor 0.35(n - 2g) \rfloor, \quad N = n - 2g - n_0 - n_L.$$

The screening observations occupy the first n_0 positions, followed by g unused observations. Selection uses the next n_L positions, followed by the second unused gap, and final estimation uses the last N observations. All boundaries are determined before the epoch begins.

The numerical implementation evaluates the fractions in these definitions in binary floating-point arithmetic before truncating to integers. At $n = 1{,}500$ this gives $(n_0, n_L, N) = (360, 503, 577)$, transferring one observation from selection to final estimation relative to exact decimal arithmetic. The reported simulations use these implemented boundaries; this one-observation transfer leaves the proportional block sizes in the theory unchanged.

At the screening boundary, we compute every marginal score and retain the $m = \min\{d, \lceil 4 n_0^{1/5} \rceil\}$ groups with the largest scores. We resolve equal scores by the original covariate order. We discard the screening moments after recording the candidate set, scores and response scale. On the selection block, we accumulate only the candidate cross moments and perform the two convex minimizations. We discard the candidate cross moments after determining the final support. The final observations update a point-estimation state and a separate inference state on that support.

The resolutions are

$$h_L = 0.85 n_L^{-1/5}, \qquad h_E = 0.85 N^{-1/5}, \qquad h_I = 0.8 N^{-1/4}, \qquad G_\bullet = 25 \vee \{\lceil 4/h_\bullet \rceil + 1\},$$

where the subscript $\bullet$ denotes the selection, estimation or inference resolution. The quadrature weights are trapezoidal. The constants are fixed in the primary numerical procedure. Except where a design

states other values, the SCAD parameter is $a = 2.4$ and the penalty is $\lambda = 0.4\,\widehat{\sigma}_Y\{\log d/(n_L h_L)\}^{1/2}$, where $\widehat{\sigma}_Y$ is the screening-block standard deviation.

This procedure never needs to reconstruct an earlier observation at a different resolution. The stored state does not provide retrospective estimates at arbitrary bandwidths. Such retrospective estimates would require storing additional resolutions in advance or retaining the observations. The planned-epoch formulation makes this computational distinction explicit.

S20.5 Deleting a block from moments

We write I for the complete final block and I_b for one of its resampling blocks. We compute the moments of $I \setminus I_b$ by subtraction. For example,

$$\widehat{c}_{-b} = \frac{a_I - a_{I_b}}{N - n_b}, \qquad \widehat{D}_{j,-b} = \frac{D_{I,j} - D_{I_b,j}}{N - n_b}, \qquad \bar{Y}_{-b} = \frac{y_I - y_{I_b}}{N - n_b}.$$

The centered matrices $R_{j,-b}$ and the normalized cross matrix are then recomputed from these moments. Holding R_j or $\bar{Y}$ fixed would produce a different deleted-block estimate and would omit part of the center variation.

The centered vectors $\widetilde{V}_b$ of (S6) sum to zero because $\sum_b n_b = N$. Conditional resampling draws from this centered empirical distribution and re-estimates the denominator from each resample. If its coordinate variance is $\widehat{v}_k = J^{-1}\sum_b \widetilde{V}_{b,k}^2$, the mean of J independent draws has variance $\widehat{v}_k/J$. For equal block sizes, substitution into (S6) gives the usual factor $(J-1)/J$ multiplying the sum of centered deleted-fit outer products.

S20.6 Finite-sample estimability and a prediction extension

For one group, we write D for its raw moment matrix and $\mathcal{N}$ for its nullspace. If $u \in \mathcal{N}$, then

$$0 = u^{\mathsf{T}} D u = N^{-1}\sum_i \mathbb{E}\big[\{v_j(U_j, X_{ij})^{\mathsf{T}} u\}^2 \mid X_i\big].$$

Because every summand is nonnegative, the local-polynomial values are zero at all observations and nodes of positive probability. Hence $a_j(X_{ij})^{\mathsf{T}} u = 0$, $c_j^{\mathsf{T}} u = 0$, and the products of the cross moments with u are zero. Adding this direction to the coefficients preserves the criterion, the component norm and the centering constraint.

An evaluation $a_j(x)^{\mathsf{T}}\beta_j$ is identified if it agrees for all coefficient vectors that the observations do not distinguish. This holds precisely when

$$a_j(x)^{\mathsf{T}} u = 0 \quad \text{for all } u \in \mathcal{N}.$$

Necessity follows by comparing β_j and $\beta_j + tu$ for arbitrary real t. Sufficiency follows because all representatives of the same quotient differ by a null vector. This test is a property of the observed design and does not prove a population density lower bound.

The implementation tests this relation after every deletion at every requested inference value and reports a violation as an inference failure. For prediction, we use a separate convention. At an unsupported continuous value, we interpolate the fitted values at the identified fixed-grid points and at the observed minimum and maximum, and use constant continuation beyond the first or last such point. This convention depends only on the fitted moment state and the two observed extremes per group. It does not estimate an unidentified coefficient by assigning it a statistical interpretation.

This extension has a deterministic error bound. Suppose the identified knot values have errors bounded by e, the largest adjacent knot gap is Δ, and $|f_j''| \le L$. Linear interpolation within the knot range has error at most

$$e + L\Delta^2/8.$$

This bound follows from the interpolation remainder in Section S6. Outside the knot interval $\mathcal{I}$, the bound $|f_j'| \le M_1$ gives a continuation error of at most $e + M_1\,\mathrm{dist}(x, \mathcal{I})$. Under $\|f_j\|_\infty \le B$ and $\|f_j''\|_\infty \le L$ on $[0,1]$, we take $M_1 = 4B + L/4$, because a Taylor expansion on an interval of length $1/2$ extending from each x gives this bound. Coverage at data-dependent knots still requires a probability argument.

S20.7 Storage as the planned sample size increases

The phase-specific counts are given in Section S6. For fixed s, $G_I \asymp N^{\kappa}$ and $L_N \asymp N^{\gamma}$, inference storage is sublinear when $\gamma > 2\kappa$. The choices $\kappa = 1/4$, $\gamma = 3/5$ satisfy this inequality and the inference theorem. For $s \asymp N^{a}$, storage requires $\gamma > 2\kappa + 2a$; this does not extend fixed-support inference to growing s. The counts concern exact criterion reconstruction, not statistical sufficiency.

S20.8 Exact reduction of stored cross moments

The cross-product representation of the stored moments is reduced without changing a coefficient or a deleted-block estimate. We write p_j for the raw coefficient dimension of selected group j and define $p = \sum_j p_j$ and $H_p = \sum_{j<k} p_j p_k$. The criterion uses the between-group blocks of Q_I and the separate diagonal matrices $D_{I,j}$ but does not use the within-group blocks of Q_I. Retaining each between-group block only once requires H_p numbers; the other moments require $O(p)$ numbers for fixed polynomial degree. The transpose of each retained block supplies the opposite block.

For a block with n_b observations, we stack the row vectors a_i^{T} in a matrix A_b. If $pn_b < H_p$, we retain a rectangular QR factor R_b with at most n_b rows. In exact arithmetic,

$$A_b^{\mathsf{T}} A_b = \mathsf{R}_b^{\mathsf{T}} \mathsf{R}_b, \qquad (Q_{I_b})_{jk} = \mathsf{R}_{b,\cdot j}^{\mathsf{T}} \mathsf{R}_{b,\cdot k}, \quad j \neq k,$$

where $\mathsf{R}_{b,\cdot j}$ consists of the columns for group j. The diagonal matrices $D_{I_b,j}$ cannot be recovered from this factor and are retained separately. At a new chunk with row matrix A_{new}, the implementation computes a QR factor of the vertical concatenation of R_b and A_{new}. The new triangular factor satisfies

$$\mathsf{R}_{\mathrm{new}}^{\mathsf{T}} \mathsf{R}_{\mathrm{new}} = \mathsf{R}_b^{\mathsf{T}} \mathsf{R}_b + A_{\mathrm{new}}^{\mathsf{T}} A_{\mathrm{new}}.$$

This update uses no rank truncation and is valid for singular matrices and for a chunk shorter than the coefficient dimension. Nonnegative observation weights are handled by multiplying each new row by the square root of its weight. The linear and diagonal moments use the same weights.

The planned block sizes determine which representation is smaller before their observations arrive. If a block size is unknown in advance and later makes pn_b exceed H_p, the retained factor is converted to the between-group blocks by the first identity. The complete inference block uses $H_p + O(p)$ numbers and the deleted blocks use

$$\sum_{b=1}^{J} \{\min(H_p, pn_b) + O(p)\}.$$

Because $H_p \leq p^2/2$, this proves (S7). Subtracting a reconstructed deleted block from the complete moments gives exactly the earlier normal equation. No approximation is introduced into the centering, group norms or standard errors. Floating-point QR and summation affect only numerical precision.

For fixed s and the stated bandwidth and block-length powers, the earlier $O(N^{9/10})$ bound remains a valid upper bound. The reduction is most substantial when a deleted block is short relative to its coefficient dimension. The reduction is not a claim of a smaller asymptotic exponent for every sample-size sequence. The numerical eigen-decomposition that produces the centered normalization still allocates a dense normalized matrix temporarily. This allocation and the resampling vectors are included in the peak process memory but are distinct from the persistent moment storage.

S21 Component risk under dependent covariates

A population comparison between the squared additive sum and the sum of squared component norms permits dependent coordinates in the point-risk calculation. The selection-error probability remains explicit.

S21.1 Dependence between coordinates in the point-risk calculation

We relax independence in the comparison of component loss with the fitting criterion. Selection consistency under correlation remains a separate question.

We replace the design transition in Theorem S1 by $P(x,\cdot) = (1-\rho)Q(\cdot)+\rho K(x,\cdot)$, where $QK = Q$ and $0 \le \rho \le \rho_0 < 1$. The chain has initial distribution Q. The transition K need not retain the preceding observation or be reversible. The coordinate marginals of Q are P_j, and Q_S is the marginal distribution on S. For every collection of centered functions $g_j \in L^2(P_j)$, we assume

$$c_Q \sum_{j\in S} \|g_j\|_{P_j}^2 \le \mathbb{E}_Q\Big\{\sum_{j\in S} g_j(X_j)\Big\}^2 \le C_Q \sum_{j\in S} \|g_j\|_{P_j}^2, \qquad 0 < c_Q \le 1 \le C_Q < \infty. \tag{S54}$$

The constants do not depend on s or N. The marginal, signal, error, estimator and tuning conditions of Theorem S1 are unchanged. The lower bound ensures that cancellation between additive effects does not prevent their identification. The upper bound controls the variance of their sum. A stronger sufficient condition is

$$0 < c_Q \le \frac{dQ_S}{d\bigotimes_{j\in S} P_j} \le C_Q < \infty.$$

This condition bounds the density of the joint distribution on the active set relative to the product of its coordinate marginals, and it implies the comparison for every collection of centered functions.

The comparison in (S54) also holds when $X_j = T_j(Z_j)$ for measurable maps T_j, where the coordinates of $Z = (Z_j)_{j\in S}$ are standard normal, every pair is jointly Gaussian, and the correlation matrix R_Z satisfies $c_Q I \preceq R_Z \preceq C_Q I$. This case includes continuous Gaussian copulas and categorical transformations, provided that the unchanged marginal conditions hold. This Gaussian comparison is a consequence of Guo and Zhang (2022, Corollaries 1–2). We apply the comparison to the augmented local-polynomial criterion and retain the selection and centering errors.

Proposition S9. *Under* (S54)*, the bound* (S4) *holds with constants also depending on* c_Q *and* C_Q*.*

Proposition S9 retains the explicit selection term $\Pr(\widehat{S} \ne S)$ and permits any Q-invariant transition between redraws. Its Gaussian sufficient condition extends risk, not selection or simultaneous inference.

Proof. We first extend (S54) to the augmented coordinates. Let $t_j(U_j, X_j)$ be centered and square integrable, and put $g_j(X_j) = \mathbb{E}\{t_j(U_j, X_j) \mid X_j\}$. Conditional independence of the auxiliary indices gives

$$\mathbb{E}_Q\Big(\sum_j t_j\Big)^2 = \sum_j \mathbb{E}\,\mathrm{Var}(t_j \mid X_j) + \mathbb{E}_Q\Big(\sum_j g_j\Big)^2,$$
$$\sum_j \mathbb{E}t_j^2 = \sum_j \mathbb{E}\,\mathrm{Var}(t_j \mid X_j) + \sum_j \mathbb{E}g_j^2.$$

Thus the same comparison holds for t_j, with the same constants. An intercept adds its square to both sides because every component is centered.

Use the marginal population coordinates R_j^0 and include an intercept. Write ϕ for the resulting vector and $\Gamma = \mathbb{E}_Q(\phi\phi^{\mathsf{T}})$. Under the product of the marginals, the second moment matrix of this vector is I_q. The augmented additive comparison implies

$$c_Q I_q \preceq \Gamma \preceq C_Q I_q.$$

The conditional distributions of the auxiliary indices given each coordinate are unchanged. No comparison is needed for nonadditive functions of X and U.

Choose the local-linear Taylor polynomial for each active component. Subtract the population mean of each polynomial and transfer the resulting constants to the intercept. If $\bar{e}_j$ is the approximation error after this centering, then $\mathbb{E}_{P_j,U_j}\bar{e}_j = 0$ and

$$\mathbb{E}_Q\Big(\sum_j \bar{e}_j\Big)^2 \le C_Q \mathbb{E}_{\otimes_j P_j}\Big[\mathbb{E}\Big\{\Big(\sum_j \bar{e}_j\Big)^2 \mid X\Big\}\Big] \le C_Q s_c L^2 h^4/4.$$

The population projection η^* for Γ has residual second moment bounded by the same quantity. Let $g_j^*(x)$ be the averaged component of this projection. The g_j^* are centered in P_j. Jensen's inequality and the lower bound in (S54) give

$$\sum_j \|g_j^* - f_j\|_{P_j}^2 = \int \Big\{\sum_j (g_j^* - f_j)\Big\}^2 \prod_j dP_j \le c_Q^{-1}\mathbb{E}_Q\Big\{\sum_j (g_j^* - f_j)(X_j)\Big\}^2 \le C s_c L^2 h^4.$$

The intercept is still μ because all groups have marginal population mean zero. The nonconstant coefficients satisfy

$$\sum_j \|\eta_j^*\|_2^2 \le c_Q^{-1}\mathbb{E}_Q\Big\{\sum_j f_j(X_j)\Big\}^2 \le (C_Q/c_Q)B^2.$$

Normalize the entire population coordinate vector by $\Gamma^{-1/2}$ to give identity population second moments. The resulting conditional one-observation matrix is positive semidefinite and has norm at most Cq/c_Q. In an interval of length ℓ, the probability of no independent redraw is ρ^ℓ. After a redraw, $QK = Q$ makes the terminal observation stationary and independent of observations before that interval. Thus the residue-class coupling of (S15) applies to this transition. The matrix Chernoff inequality and this coupling give

$$\Pr\{\|\Gamma^{-1/2}\widehat{M}\Gamma^{-1/2} - I\|_{\mathrm{op}} > 1/2\} \le 2\ell q e^{-cN/(q\ell)} + N\rho^\ell.$$

On the complementary event, $\|\widehat{M}^{-1}\|_{\mathrm{op}} \le 2/c_Q$. The normal-equation residual consists of the approximation part and the error part. Their coefficient second moments are bounded by $Cs_cL^2h^4$ and Cvq/N, respectively, using the conditional Cauchy–Schwarz and covariance calculations of the component-risk proof.

The empirical center is $N^{-1}\sum_i g_j^*(X_{ij})$ plus coefficient-error terms. For a centered $g \in L^2(Q)$, write $(Kg)(x) = \int g(y)K(x,dy)$. Invariance and Jensen's inequality give the first bound below. For $k \ge 1$, $P^k g = \rho^k K^k g$, and the Cauchy–Schwarz inequality gives the second bound:

$$\|Kg\|_{L^2(Q)}^2 \le \int K(g^2)\,dQ = \|g\|_{L^2(Q)}^2,$$
$$|\operatorname{Cov}\{g(X_0), g(X_k)\}| = \rho^k\left|\int gK^k g\,dQ\right| \le \rho^k\|g\|_{L^2(Q)}^2.$$

Apply this bound to $g(X) = g_j^*(X_j)$. Summing the geometric covariance bound and using the coefficient bound shows that the summed center variances are at most CB^2/N. On the matrix event, the bounded empirical mean row of $\widehat{M}$ bounds the centering products by the coefficient error. The truncation and selection steps follow the proof of Theorem S1, which gives (S4).

For the Gaussian sufficient condition, apply Guo and Zhang (2022, Corollary 1, equation (18), and Corollary 2) to the centered functions $g_j\{T_j(Z_j)\}$. Their summed second moments equal $\sum_j \|g_j\|_{P_j}^2$, and the extreme eigenvalues of R_Z bound the second moment of their sum. This gives (S54), including measurable categorical transformations. Zero-variance functions contribute zero to both sides.

For a correlation matrix with entries $r^{|j-k|}$ and $0 \le r < 1$, the Fourier series $\sum_{\ell\in\mathbb{Z}} r^{|\ell|}e^{i\ell t} = (1-r^2)/(1-2r\cos t + r^2)$ lies between $(1-r)/(1+r)$ and $(1+r)/(1-r)$. Integrating this series against $|\sum_j v_j e^{ijt}|^2$ for an arbitrary vector v and using Parseval's identity gives the same bounds for the correlation eigenvalues. The choice $r = 1/2$ in Design D1 gives $c_Q = 1/3$ and $C_Q = 3$.

For two coordinates with correlation $0 < r < 1$, the Gaussian density ratio relative to independent standard marginals equals $(1-r^2)^{-1/2}\exp\{rt^2/(1+r)\}$ at (t,t) and $(1-r^2)^{-1/2}\exp\{-rt^2/(1-r)\}$ at $(t,-t)$. The first is unbounded and the second tends to zero as $t \to \infty$. Negative r exchanges the two diagonals. The same ratio is obtained after coordinatewise strictly monotone continuous transformations. Thus the Gaussian sufficient condition does not require the stronger density-ratio bounds. This completes the proof. □

S21.2 What the bounded-signal class permits

The condition $\sum_j \|f_j\|_\infty^2 \le B^2$ implies $sb_N^2 \le B^2$, where b_N is the smallest norm among the active components. The condition permits many weak components but not an increasing number of components whose norms are all bounded below by the same positive constant. This restriction is a property of the parameter class and not of smooth backfitting.

Table S1: Data dimensions, distributions and temporal parameters. In D1–D6, ρ is the latent temporal autoregression coefficient; in D7–D8, it is the covariate retention probability.

Design	n	d	ρ	ϑ	Covariates
D1	1500	200	0.0	0.0	Gaussian copula
D2	1500	200	0.3	0.0	Gaussian copula
D3	1500	200	0.6	0.3	Gaussian copula
D4	1500	500	0.3	0.0	Gaussian copula
D5	3000	500	0.3	0.0	Gaussian copula
D6	6000	500	0.3	0.0	Gaussian copula
D7	1500	200	0.3	0.3	Nonuniform marginals
D8	3000	200	0.3	0.3	Mixed groups

For the continuous comparison, we take $d = n^2$, $m \asymp n^{1/5}$ and $h_L \asymp n^{-1/5}$. If $s \le cn^{1/5}$ and $b_n = n^{-1/5}$, then

$$\frac{m^2 G_L \log^3(dn)}{n} \asymp n^{-2/5}\log^3 n \to 0, \qquad sh_L^2 \lesssim n^{-1/5} \to 0,$$

$$\frac{h_L^2 + \lambda_n^0}{b_n} \lesssim n^{-1/5}(1+\sqrt{\log n}) \to 0, \qquad sG_E \log^2 n/n \lesssim n^{-3/5}\log^2 n \to 0.$$

The lower-bound construction also satisfies

$$\sqrt{s}\, b_n \lesssim n^{-1/10} \to 0, \qquad \sqrt{s}\{\log(esn^{1/5})/n\}^{2/5} \lesssim n^{-3/10}\log^{2/5} n \to 0.$$

Thus the upper- and lower-bound parameter classes have a nonempty intersection with increasing support size and decreasing minimum signal. This example concerns estimation; the inference proofs allow a growing evaluation grid but keep s fixed.

S22 Numerical studies

The eight initial designs use 200 independent replications each; every additional design uses at least 100. Methods share observations within each replication. All runs enter the summaries, and an omitted active effect counts as noncoverage. The procedures and their constants were fixed before the production runs, and every pilot run used seeds disjoint from the production seeds.

S22.1 Designs and comparisons

The first five covariates have nonzero effects. The continuous designs have four nonlinear effects and one linear effect; D8 replaces the fifth by a categorical effect. Continuous effects are centered in their population marginals. All errors have population standard deviation 1.5 and follow a stationary Gaussian autoregression independent of the covariates. Table S1 gives the sample sizes, dimensions and temporal parameters of the eight designs. The detailed regression functions and stationary initialization are given in Section S25.

D1–D6 have correlated Gaussian-copula covariates. The latent spatial correlation between coordinates j and k is $0.5^{|j-k|}$, and ρ is their temporal autoregression coefficient. D1 has independent observations and satisfies the design comparison in Proposition S9. The proposition gives a component-risk bound with an explicit selection-error term, not a simultaneous-coverage guarantee. The dependent designs D2–D6 do not have the independent-redraw transition of the design classes for which the guarantees are proved.

D7 has marginal density $p(x) = 1/2 + x$, and D8 includes one 24-level categorical predictor. In D7 and D8, coordinates are independent at each time, and the complete vector is independently redrawn with probability $1 - \rho$. These two designs follow the dependence structure of the original population class. We do not claim that every asymptotic rate inequality is already small at their finite sample sizes.

SCAD-refit is the proposed procedure. Lasso-refit applies the same final unpenalized fit to the earlier group-Lasso support; Lasso penalizes the screened candidates on the final block. Oracle uses the

Table S2: Integrated component error of the additive fits.

Design	SCAD-refit	Lasso-refit	Lasso	Oracle
D1	0.116 (0.034)	0.119 (0.035)	0.578 (0.108)	0.115 (0.033)
D2	0.140 (0.183)	0.142 (0.183)	0.603 (0.167)	0.120 (0.033)
D3	0.186 (0.209)	0.191 (0.208)	0.626 (0.206)	0.157 (0.048)
D4	0.137 (0.129)	0.138 (0.129)	0.684 (0.156)	0.125 (0.040)
D5	0.064 (0.017)	0.064 (0.018)	0.367 (0.068)	0.064 (0.017)
D6	0.034 (0.009)	0.034 (0.009)	0.206 (0.029)	0.034 (0.009)
D7	0.149 (0.139)	0.153 (0.141)	0.492 (0.160)	0.125 (0.034)
D8	0.118 (0.025)	0.119 (0.027)	0.316 (0.060)	0.117 (0.025)

Numbers in parentheses represent standard deviations. An omitted active effect contributes its squared norm to the loss.

Table S3: Selection of the active components by SCAD, with the exact recovery of Lasso-refit.

Design	TPR	FDP	Exact recovery	Lasso-refit exact	Selected size
D1	1.000	0.014	0.915	0.790	5.08 (0.28)
D2	0.998	0.016	0.910	0.770	5.08 (0.28)
D3	0.997	0.033	0.810	0.600	5.19 (0.46)
D4	0.998	0.004	0.970	0.920	5.01 (0.16)
D5	1.000	0.006	0.965	0.915	5.04 (0.18)
D6	1.000	0.004	0.975	0.935	5.03 (0.16)
D7	0.995	0.026	0.875	0.710	5.14 (0.47)
D8	1.000	0.013	0.930	0.850	5.08 (0.34)

Numbers in parentheses represent standard deviations. TPR is the true-positive rate and FDP is the false discovery proportion.

known five active variables only for comparison. All methods use the same component target, point bandwidth and pointwise bound.

The loss is $\sum_{j=1}^{d} \|\widehat{f}_j - f_j\|_{P_j}^2$. This loss includes both omitted active components and estimated inactive components. Numerical integration uses 161-point Gaussian quadrature for continuous covariates and the exact probability-weighted sum for a categorical component. The true-positive rate, denoted by TPR, is the fraction of the five active components selected. The false discovery proportion, denoted by FDP, is the fraction of selected components that are inactive, defined to be zero for an empty model.

S22.2 Component estimation and variable selection

SCAD-refit has lower mean error than Lasso in all eight designs in Table S2. Its proximity to Lasso-refit indicates that removing shrinkage explains much of the gain. SCAD-refit also retains fewer inactive effects than Lasso-refit in every design, and Table S3 gives the exact recovery of both.

At dimension 500, component error decreases across D4–D6 and approaches the Oracle error. Table S3 shows that exact recovery is less frequent than retaining all active effects: one extra null group already makes the selected support incorrect.

The sampling standard deviations in Table S2 include occasional support errors. In D2, D3, D4 and D7, these errors contribute to larger variability than for the known-support fit. No pointwise truncation occurred at the 501 continuous check points per selected component or at any categorical level.

S22.3 Joint uncertainty at the reported values

For each selected continuous component, inference uses 29 points from 0.15 to 0.85; for a selected categorical component, it uses all 24 levels. Each replication uses 999 studentized resampling draws from exact deleted-block fits. The block length is $\lceil 0.4N^{3/5} \rceil$, and the inference bandwidth is $0.8N^{-1/4}$. Here N is the final block size. These choices also determine the inference storage

Table S4: Finite-grid confidence coverage and interval width.

Design	All selected	True active	Pointwise	Mean width
D1	0.950	0.950	0.948	1.106 (0.051)
D2	0.950	0.950	0.951	1.134 (0.053)
D3	0.950	0.950	0.946	1.232 (0.067)
D4	0.935	0.940	0.952	1.126 (0.051)
D5	0.960	0.960	0.956	0.819 (0.026)
D6	0.965	0.965	0.964	0.613 (0.015)
D7	0.940	0.940	0.951	1.262 (0.083)
D8	0.980	0.980	0.948	1.092 (0.040)

Numbers in parentheses represent standard deviations. All selected requires coverage of all selected values and the presence of every active component.

Table S5: Component error and exact recovery with correlated predictors.

Design	Component error		Exact recovery	
	SCAD-refit	Lasso-refit	SCAD	Lasso
R1	0.226 (0.089)	0.250 (0.101)	0.92	0.28
R2	0.121 (0.027)	0.131 (0.033)	0.97	0.21
R3	0.061 (0.013)	0.066 (0.015)	0.99	0.25

Numbers in parentheses represent standard deviations.

bound. Section S29.5 varies both constants. All selected components enter the resampling maximum, including any selected null component.

Table S4 gives two joint coverage summaries. True-active coverage requires every true active effect to be selected and covered at every reported value. All-selected coverage also requires coverage of every selected inactive effect. Pointwise coverage is the average over five central values and uses the pointwise Student quantile with $J - 1$ degrees of freedom, with an omitted active component contributing zero to that average. Mean width is the average over the reported values of the selected true active components.

All-selected coverage ranges from 0.935 to 0.980. The empirical Monte Carlo standard error at coverage 0.95 with 200 replications is about 0.015. Mean width decreases from 1.1257 to 0.8190 and 0.6132 as the sample size increases across D4–D6. These results concern a finite evaluation grid; they do not measure coverage between its values.

The largest within-replication average width across the 1,600 primary replications is 1.601.

Section S26 isolates centering variation, and Section S27 examines increasing grids and curvature-based interpolation.

S22.4 The effect of SCAD selection under correlation

Refitting removes shrinkage but preserves an incorrectly selected variable. Designs R1–R3 examine this distinction under the population model of Theorem S4. The two correlated continuous effects are affine, and an inactive binary variable is associated with both. Three further active variables have two nonlinear continuous effects and one 24-level categorical effect. The stationary Markov transition changes covariates between independent draws, as allowed by (S8). Section S28 specifies the distribution and its exact simulation.

The epoch sizes are $1{,}500$, $3{,}000$ and $6{,}000$, and the corresponding dimensions are 200, 500 and 500. The reset probabilities are 0.5, 0.25 and 0.25; the error autoregression coefficient is 0.3 in all three designs. We retain the procedure, penalty rule, bandwidths and loss definition used in D1–D8. Each design uses 100 independent replications.

In Table S5, SCAD improves recovery by removing the inactive binary group. Both procedures have the same true-positive rates. The difference does not trade missed effects for fewer false positives. Table S13 gives the selection frequencies.

Table S6: Time and memory for identical estimates and intervals.

Design	Representation	Time	Peak RSS	Moments
B1	Batch	3.07 (0.26)	234.1 (9.7)	–
B1	Dense online	2.91 (0.23)	230.0 (9.0)	14.74 (1.91)
B1	Reduced online	2.80 (0.26)	204.3 (7.1)	1.66 (0.13)
B2	Batch	10.02 (0.84)	329.3 (10.9)	–
B2	Dense online	8.99 (0.94)	297.9 (9.6)	26.28 (2.40)
B2	Reduced online	8.77 (0.76)	246.0 (6.4)	3.32 (0.17)
B3	Batch	15.83 (2.51)	466.4 (12.5)	–
B3	Dense online	14.37 (2.23)	387.9 (13.0)	45.82 (3.26)
B3	Reduced online	13.99 (2.11)	292.9 (7.1)	6.99 (0.25)

Numbers in parentheses represent standard deviations. Time is in seconds. Memory is in MB, with $1\,\mathrm{MB} = 10^6$ bytes. Peak RSS is process memory; Moments measures retained numerical moment summaries.

SCAD-refit reduces mean error by 9.5%, 7.9% and 7.7% relative to Lasso-refit. The shared final unpenalized fit isolates the selection step. Section S28 gives paired differences and their Monte Carlo uncertainty.

The same selected models give all-selected simultaneous coverage 0.96, 0.93 and 0.93, with mean widths 1.5027, 1.1254 and 0.8065. Each calculation uses 999 studentized resamples and the same reported values as in the eight initial designs. The Monte Carlo intervals for these three coverage estimates, given in Section S28, all contain the nominal level 0.95. These results connect the correlated selection experiment to the inferential calculation in Theorem S5.

S22.5 Time and storage for the same statistical calculation

The moment reduction changes what is stored, not which estimator is computed. We compare three implementations, labeled Batch, Dense online and Reduced online: a batch calculation retaining the observations, a sequential calculation with dense cross moments, and the reduced sequential representation in Section S20.8. Each implementation computes support selection, both final fits and all deleted-block intervals on the same data, with identical tuning and resampling draws. The sequential implementations receive chunks of 128 observations.

Designs B1–B3 use the distribution of the correlated selection experiment, with epoch sizes $1{,}500$, $3{,}000$ and $6{,}000$ and dimensions 200, $1{,}000$ and $1{,}000$. Each design uses 100 independent datasets, each processed by all three implementations in separate processes. We measure elapsed time and peak resident set size, denoted by RSS, for the complete calculation. The size of retained numerical moment summaries is recorded separately. Batch stores the response and covariate arrays, Dense online stores the full cross matrix, and Reduced online uses between-group blocks or the exact QR representation for short blocks. Each computes both final fits, all exact deleted fits and 999 resamples, with the same evaluation values and resampling seeds. The B1–B3 datasets are independent of those used in R1–R3.

Each implementation runs in a new process, with its order rotated across replications. Elapsed time includes data generation, selection, both fits and intervals, but excludes process startup and library imports. Peak RSS includes libraries and temporary allocations. The moment count is the maximum over the screening, selection and final phases, including retained arrays and three scalars per state but excluding Python object overhead. Memory uses 10^6 bytes per MB. Computation uses an Apple M4 Max with 128 GiB, macOS 26.2 and x86-64 Python 3.11.7 with NumPy 1.26.4 and SciPy 1.17.1. Six paired replications run concurrently with one numerical-library thread per process; the timing results concern this shared-machine configuration.

Table S6 shows smaller mean moment storage, peak RSS and elapsed time for the reduced representation in all three designs. In B3, moment storage decreases from 45.82 to 6.99 MB relative to dense sequential computation, and peak RSS decreases from 387.9 to 292.9 MB. These finite-sample reductions support the memory advantage in (S7); the relative time savings are smaller. Batch retains $8n(d+1)$ bytes of float64 observations, which equal 2.41, 24.02 and 48.05 MB in B1–B3. The reduced moments occupy 69, 14 and 15 percent of these sizes.

Table S7: Support recovery and integrated component error in the Gaussian-copula separation designs. Numbers in parentheses represent standard deviations.

		Exact support			Integrated component error		
Design	n	SCAD	Threshold	Adaptive	SCAD	Threshold	Adaptive
G1	12,000	0.860	0.000	0.000	0.0205 (0.0015)	0.8738 (0.3590)	0.0340 (0.1000)
G2	15,000	0.940	0.000	0.010	0.0150 (0.0010)	0.7225 (0.2360)	0.0177 (0.0011)
G3	25,000	0.990	0.000	0.000	0.0079 (0.0002)	0.4555 (0.1232)	0.0093 (0.0005)

Every selected support agrees across the three implementations in all 300 datasets. The largest absolute difference among the evaluated point estimates, interval centers, standard errors and interval limits is less than 3×10^{-14}. Thus the memory reduction does not alter selection or the reported uncertainty in these comparisons.

S22.6 SCAD SELECTION BEYOND TWO-STAGE REFITTING

The following experiment uses the population calculation in Proposition S5. The first three latent Gaussian scores have correlation matrix Γ_0, and $X_j = \Phi(Z_j)$. The response has the two active probit-score components in that proposition. The remaining covariates are independent uniform variables. The Gaussian errors have standard deviation 0.2 and autoregression coefficient 0.25.

Designs G1–G3 use $(n, d) = (12{,}000, 10)$, $(15{,}000, 10)$ and $(25{,}000, 10)$. Their penalties are 0.13, 0.115 and 0.09. SCAD-refit, threshold-refit and adaptive-refit use the same screening set, local-polynomial quadratic, penalty and final block. Threshold-refit retains the groups with initial group norms above the penalty. Adaptive-refit uses weight $\lambda^2/\|\widehat{\theta}_j^L\|_2$. Each design has 100 independent replications. The SCAD parameter is $a = 3.7$. The initial and SCAD solves use optimality tolerance 10^{-8} and at most $2{,}500$ sweeps; the adaptive solve uses n_L^{-3} and $100 + \lceil 30 \log n_L \rceil$ sweeps. The initial and SCAD supports use norm cutoff 10^{-7}; the adaptive support uses n_L^{-3}.

Table S7 compares selected supports with the same final unpenalized fit. SCAD error is close to the Oracle error in Table S14. G1–G3 use the different solver tolerances above; the next comparison uses common tolerances.

The initial group norm of the inactive variable exceeds the smaller of the two active norms in 1.000, 0.990 and 1.000 of the replications. This is the ordering that no threshold of these norms corrects. A search over all cutoffs applied in common to every initial norm recovers the exact support in no replication of G1 or G3 and in one replication of G2. These finite-sample orderings agree with the strict population inequalities in Proposition S5.

S22.7 SELECTION WITH BOUNDED SMOOTH EFFECTS

The probit-score effects of the preceding designs are unbounded near the ends of the covariate range. The comparison in Section S14.5 separates the selection methods without unbounded component functions. We examine this distinction with three Gaussian-copula covariates having correlation matrix Γ_0. For a real number z, we set $\chi(z) = 1$ when $|z| \le 3$ and $\chi(z) = 0$ when $|z| \ge 4$. We define $t = |z| - 3$ on $3 < |z| < 4$. The transition and transformed function are

$$\chi(z) = \frac{\exp(-1/(1-t))}{\exp(-1/t) + \exp(-1/(1-t))}, \qquad q_0(x) = \chi\{\Phi^{-1}(x)\}\Phi^{-1}(x).$$

The function q_0 is the probit score multiplied by this transition, is set to zero near the boundaries of $[0, 1]$ and is bounded, smooth and centered. The two active effects are $\lambda\beta_1^\sharp q_0$ and $\lambda\beta_2^\sharp q_0$. Here $\beta^\sharp = \beta^*/\lambda$ is defined in Section S14.5; q_0 is the function q_ϵ of that section with $\epsilon = \Phi(-3)$. The third effect is zero; this makes the third covariate the inactive variable of the comparison. The error standard deviation is 0.2 and its autoregression coefficient is 0.25.

Designs H1–H3 use sample sizes $12{,}000$, $15{,}000$ and $25{,}000$ and penalties 0.13, 0.115 and 0.09, respectively. All three covariates enter the candidate set. Every selection rule uses the same local-polynomial criterion, selection observations and unpenalized final fit. Initial Lasso, SCAD with $a = 3.7$ and adaptive group Lasso use the common optimality tolerance n_L^{-3} and the sweep limit

Table S8: Support recovery and component error for bounded smooth effects. Numbers in parentheses represent standard deviations.

Design	Method	Exact support	Component error	Active error	Inactive error
H1	SCAD-refit	0.99	0.0182 (0.0007)	0.0182	0.0000
H1	Threshold-refit	0.00	0.8388 (0.3722)	0.7509	0.0879
H1	Adaptive-refit	0.02	0.0209 (0.0011)	0.0202	0.0007
H1	Lasso-refit	0.00	0.0210 (0.0011)	0.0203	0.0007
H1	Oracle	1.00	0.0182 (0.0006)	0.0182	0.0000
H2	SCAD-refit	1.00	0.0137 (0.0005)	0.0137	0.0000
H2	Threshold-refit	0.00	0.6270 (0.3170)	0.5612	0.0658
H2	Adaptive-refit	0.00	0.0155 (0.0007)	0.0151	0.0004
H2	Lasso-refit	0.00	0.0155 (0.0007)	0.0151	0.0004
H2	Oracle	1.00	0.0137 (0.0005)	0.0137	0.0000
H3	SCAD-refit	1.00	0.0074 (0.0002)	0.0074	0.0000
H3	Threshold-refit	0.00	0.3926 (0.1871)	0.3515	0.0411
H3	Adaptive-refit	0.00	0.0083 (0.0004)	0.0080	0.0002
H3	Lasso-refit	0.00	0.0083 (0.0004)	0.0080	0.0002
H3	Oracle	1.00	0.0074 (0.0002)	0.0074	0.0000

$100 + \lceil 30 \log n_L \rceil$. Lasso, SCAD and adaptive supports use the same norm tolerance. Component error is integrated by a 20-point composite Gauss–Legendre rule split at kernel boundaries and at the smooth-transition boundaries. Each design has 100 independent replications. All 300 calculations reached their stated tolerances.

Table S8 gives the smallest mean error among estimated-support methods for SCAD. Its paired reductions relative to adaptive group Lasso are 0.00267, 0.00176 and 0.00088, with Monte Carlo standard errors 0.000093, 0.000061 and 0.000031. The inactive initial norm exceeds the smaller active norm in 99, 100 and 100 runs. Thus the selection distinction persists for bounded smooth effects, without claiming an ordering for every design or penalty.

S22.8 Continuing estimation

Designs A1–A3 use the five continuous effects in the initial numerical study with independent uniform coordinates. A1 has $(n, d) = (12{,}000, 20)$ and independent covariates and errors over time. A2 has $(n, d) = (18{,}000, 50)$ and retains the preceding covariate vector with probability 0.3. A3 has $(n, d) = (30{,}000, 50)$, the same covariate transition and error autoregression coefficient 0.3. These designs follow the population process in Section S4.

The continuing estimator reserves the $n^{3/4}$ selection subsequence and uses block boundaries $\lfloor r^{5/2} \rfloor$. The point and inference bandwidth constants are 0.75 and 0.65. The reporting grids follow the nested activation rule in Section S9. The reporting resolutions have 33 point-estimation grid points and 65 inference grid points. Each method uses the same 41 quadrature values per component. Continuing inference uses 999 studentized resampling draws in each replication.

Selection on this reserved subsequence with k retained observations uses bandwidth $0.85k^{-1/5}$, $\lceil 4k^{1/5} \rceil$ screened candidates and the constant 0.8 in the penalty formula of Section S20 with k in place of n_L. The first estimation boundary is $1{,}000$, blocks with fewer than 80 estimation observations are not fitted, and the support is updated once at least 150 reserved observations are available and after every 75 further reserved observations.

We compare the continuing estimator with the planned-epoch estimator from the final block, the full-data known-support fit and the latest-block fit. Each design has 100 independent replications. Planned-epoch selection uses penalty multiplier 0.8, matching the continuing multiplier. Its screening, selection and estimation blocks, bandwidths and final unpenalized fit follow Section S20. The selection sample sizes still differ. This comparison isolates neither the effect of selection sample size nor that of estimation sample size.

Table S9 compares data use, support and component error at the common penalty multiplier. Both procedures recover all five active effects and exclude every null group in all 100 runs per design.

Table S9: Selection, component error and joint coverage for continuing estimation. Numbers in parentheses represent standard deviations.

			Exact support		Integrated component error			
Design	n	Used fraction	Continuing	Planned	Continuing	Planned	Full oracle	Coverage
A1	12,000	0.833	1.000	1.000	0.0094 (0.0020)	0.0156 (0.0033)	0.0071 (0.0014)	0.970
A2	18,000	0.865	1.000	1.000	0.0073 (0.0014)	0.0121 (0.0030)	0.0055 (0.0010)	0.940
A3	30,000	0.894	1.000	1.000	0.0055 (0.0011)	0.0089 (0.0018)	0.0042 (0.0008)	0.920

Continuing error is smaller in 99, 95 and 99 paired runs in A1–A3. Its mean reductions are 39.8, 39.5 and 38.1 percent. The mean paired decreases are 0.00622, 0.00478 and 0.00340, with Monte Carlo standard errors 0.00032, 0.00029 and 0.00017. These gains concern component estimation using the retained earlier moments, not improved support recovery in these designs. The full-data Oracle remains more accurate with its known support and complete data use.

Joint coverage of the simultaneous intervals at the reported values is 0.970, 0.940 and 0.920, with corresponding Monte Carlo standard errors 0.017, 0.024 and 0.027. The mean width decreases from 0.4904 in A1 to 0.3181 in A3, as reported in Table S15. Compared with the latest-block fit at the same selected support, retaining earlier moments reduces component error at these sample sizes.

S23 Hourly bike rentals

The Capital Bikeshare data record hourly rentals and weather (Fanaee-T, 2013; Fanaee-T and Gama, 2014). We estimate additive associations with log rental demand. The hour of the day enters as one categorical effect, as in Section S3.

S23.1 Data and fitting

To estimate the component associations in a recorded sample, we use the last 8,000 of the source's 17,379 recorded hours, from January 31 to December 31, 2012, with response $\log(1 + \mathrm{cnt})$. Here cnt is the recorded number of rentals in that hour. The predictors are temperature, apparent temperature, humidity, wind speed, hour and the response exactly 24 calendar hours earlier. Lags are matched by exact calendar time because 24 preceding rows can span more than one day; rows with missing lagged responses are dropped. This restriction excludes 35 rows and leaves 7,965 observations. Weather variables retain their released normalization, hour has levels zero to 23, and the lagged response is divided by $\log(1001)$ and restricted to $[0, 1]$. No retained lag exceeds this fixed scale.

We add 50 independent synthetic null covariates, each obtained by applying the normal distribution function to a stationary Gaussian autoregression with coefficient 0.3. They are generated on the original 8,000-row index independently of the rental series and then restricted to rows with a lag. The descriptive epoch has screening, selection and final block sizes 1,973, 2,762 and 3,158, separated by two gaps of 36 observations. The response is standardized using the screening mean and sample standard deviation, and component estimates are returned to the log-response scale. The pointwise bound is applied on the standardized scale.

S23.2 Selected components and nominal uncertainty

The descriptive fit selects temperature, humidity, hour and previous-day demand. These same four variables are selected in all 100 replications of the synthetic null covariates, with no null selected. The observed series is fixed across replications; these frequencies measure stability against the specified null variables and not sampling probabilities across rental systems. Figure S1 gives the fitted effects and nominal simultaneous intervals.

Inference uses 41 values between the empirical 10th and 90th percentiles for each selected continuous covariate and all 24 hour levels. The 3,158 final observations form 19 consecutive deletion groups of size 166 or 167. This gives a joint critical value of 5.793 for the simultaneous intervals. These groups span approximately a week; missing records prevent exact calendar-week blocks. This choice accounts for the daily cycle and replaces the deleted-block length rule used in the simulation designs.

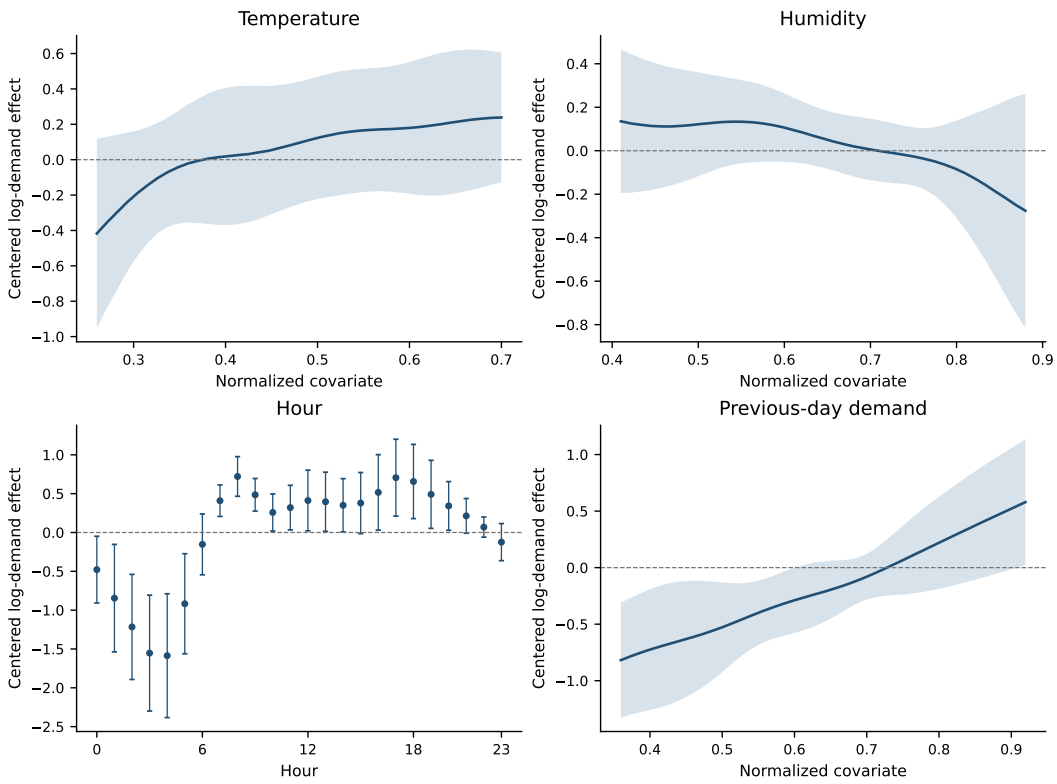


Figure S1: Centered bike-rental effects and nominal simultaneous intervals.

That rule would give 62 blocks of size 50 or 51. Each deletion holds the support, bandwidth and evaluation values fixed and recomputes component means and local-polynomial normalizations.

The hour effect ranges from approximately -1.587 to 0.721, and the previous-day effect from -0.818 to 0.579, both on the log-response scale. These centered conditional associations adjust for the other selected components and have no causal interpretation. The deterministic clock, correlated weather and empirical evaluation grid are outside the proved population conditions. The intervals are nominal model-based summaries; their coverage for this series is not established by the theorems.

S23.3 Hour contrasts

Differences between hour effects remove their common empirical center. We form all 276 pairwise contrasts from the complete fit and each of its 19 exact deleted fits, and we apply the transformation in Section S19 before estimating covariance. The 999 studentized resamples give critical value 7.1344, with nominal intervals excluding zero for 122 pairs. The 08:00–05:00 difference is 1.6390 with interval $[0.6251, 2.6530]$, and the 17:00–05:00 difference is 1.6233 with interval $[0.2495, 2.9972]$. The 17:00–08:00 interval contains zero. All pairs enter the same maximum, including any pair chosen after inspecting the fitted curve.

S23.4 Sequential response scaling

The screening response scale can be computed from raw moments. We define I_0 as the screening block of size n_0, with response mean $\bar{Y}_0$ and sample standard deviation $\widehat{\sigma}_0 > 0$. The standardized summaries are

$$b_{I_0}^{\mathrm{std}} = (b_{I_0} - \bar{Y}_0 a_{I_0})/\widehat{\sigma}_0, \quad y_{I_0}^{\mathrm{std}} = 0, \quad y_{I_0,2}^{\mathrm{std}} = (y_{I_0,2} - n_0 \bar{Y}_0^2)/\widehat{\sigma}_0^2.$$

These identities follow by summing the standardized responses and their squares. Design moments are unchanged, and subsequent responses are standardized when observed. This yields the same scores and penalty scale as batch standardization.

Table S10: SCAD selection frequencies under 100 consecutive-block deletions.

Group	Selection frequency
lights	0.64
T1	1.00
T3	1.00
T4	0.99
T8	1.00
lights_lag1	1.00
T1_lag1	1.00
T4_lag1	1.00
T2_lag6	0.95
T3_lag6	1.00
T8_lag6	0.87
lights_lag144	0.87
Press_mm_hg_lag144	1.00
hour	1.00

S24 Appliance energy use

The Appliances Energy Prediction data record energy use, room conditions, lighting and nearby weather every ten minutes (Candanedo, 2017; Candanedo, Feldheim and Deramaix, 2017). We study selection among observed and lagged covariates without adding synthetic variables.

S24.1 Data and chronological construction

The source contains 19,735 observations from January 11 to May 27, 2016; appliance and lighting energy are measured in watt-hours. We remove its two artificial random columns to leave 25 observed predictors. Each enters at lags 0, 1, 6 and 144 ten-minute intervals. Hour and weekday add categorical groups with 24 and seven levels, which yields 102 candidate groups. After the one-day lag, 19,591 rows remain, from January 12 at 17:00 to May 27 at 18:00. The response is the logarithm of one plus appliance energy; its lags are excluded from the additive candidate set.

The analysis uses the first 15,672 rows, ending at 12:50 on April 30, 2016. Response location and scale, screening, selection and fitting use these observations. Continuous groups are scaled by their empirical 0.005 and 0.995 quantiles and clipped to $[0, 1]$. No future value enters a lagged covariate.

S24.2 Selection stability

SCAD-refit selects 14 groups. To measure stability, each of 100 fits deletes a consecutive block containing ten percent of the selection observations. Starting positions are equally spaced and rounded down; the blocks can overlap. The screened candidate set stays fixed, and each fit recomputes the initial group Lasso, SCAD weights and weighted estimate. Final refitted components do not enter the selection frequency.

Table S10 gives frequencies from 0.64 to 1.00. Nine groups are selected in every deletion and two more in at least 95. Current lighting is the least stable selected group. These frequencies concern one recorded household series, without a sampling interpretation across households or periods.

S24.3 Selected effects and nominal uncertainty

Each selected continuous group is evaluated at 21 quantiles of the final estimation block between 0.10 and 0.90, with duplicates removed. The three lighting groups each have two values, and hour has 24; this yields 240 values in total. Exact refits after deleting 83 consecutive blocks of 75 or 76 observations recompute component centers and within-group normalizations. The 999 studentized resamples give critical value 3.674 and mean width 0.644 on the log-response scale. Figure S2 presents five selected effects.

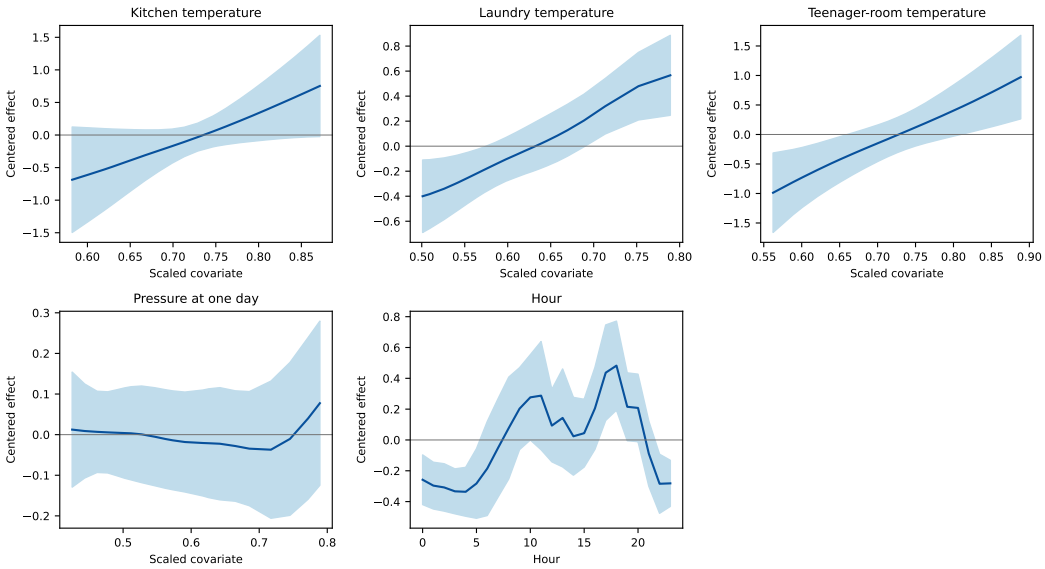


Figure S2: Selected appliance-energy effects and nominal simultaneous intervals.

The indoor-temperature associations differ across rooms and lags, and the hour effect varies after adjustment for the selected measurements. These centered associations do not identify causal effects of temperature, lighting or time. The intervals are nominal because the stationary transition and Gaussian error conditions are not established for this series.

S25 Simulation models

The regression functions and stationary transitions below specify the population targets used in the component-error and coverage comparisons.

S25.1 Regression functions and their population centers

We start with the following five continuous functions:

$$2\sin(2\pi x), \quad 12(x-1/2)^2-1, \quad 4(x-1/2), \quad 1.6\cos(2\pi x), \quad 2\{1-2(x-1/2)^2\}\cos(\pi x).$$

All five functions integrate to zero under the uniform distribution; the fifth is antisymmetric about $1/2$.

For D7, we subtract from each function its integral against $p(x) = 1/2 + x$. These fixed centers are computed by 64-point Gaussian quadrature and are not estimated from a simulated sample. If U is uniform, then

$$X = -\tfrac{1}{2} + \sqrt{\tfrac{1}{4} + 2U}$$

has this density because its distribution function is $(x + x^2)/2$. The population mean of X is $7/12$. A sample-centered linear fit is compared with a function centered at this population value.

For D8, the fifth coordinate is uniform on $\{0, \ldots, 23\}$, with effect

$$f_5(c) = 1.5\sin(2\pi c/24) + 0.7\cos(4\pi c/24).$$

Each trigonometric term sums to zero over the 24 levels. The estimator represents this effect by these 24 level values subject to one empirical centering constraint, and it does not fit two unrestricted functions of sine and cosine.

S25.2 Stationary dependence and fixed error variance

For D1–D6, we generate stationary Gaussian vectors Z_i with covariance $\mathrm{Cov}(Z_{ij}, Z_{ik}) = 0.5^{|j-k|}$ and temporal evolution $Z_i = \rho Z_{i-1} + \sqrt{1-\rho^2}\,\xi_i$. The innovation vectors have the same spatial covariance, and the initial vector has the stationary distribution. Setting $X_{ij} = \Phi(Z_{ij})$ makes every coordinate marginally uniform.

For D7 and D8, we draw the complete initial vector from its product marginal. At each later time, we independently retain the previous vector with probability ρ, the retention probability, or draw a new product vector with probability $1-\rho$. The same retention decision applies to all coordinates. Coordinates are independent at each time and share the redraw times.

We generate the regression errors as a scalar stationary autoregression, independently of every design innovation and redraw decision:

$$\varepsilon_i = \vartheta\varepsilon_{i-1} + 1.5\sqrt{1-\vartheta^2}\,e_i, \qquad e_i \text{ independent } N(0,1), \qquad \varepsilon_1 \sim N(0, 1.5^2).$$

Then $Y_i = \sum_{j=1}^{5} f_j(X_{ij}) + \varepsilon_i$. The error standard deviation is fixed across replications.

S26 Sampling variation of component centers

A component is centered in its population marginal, but the fitted criterion uses the empirical marginal. Section S18 shows that this distinction contributes to component error even for an exactly linear regression function. The following paired experiment separates that contribution from smoothing bias and variable selection.

S26.1 Designs with an exact local representation

We use 100 covariates and either one or two active effects. C1 has one continuous active effect $f_1(x) = 6(x - 1/2)$ with uniform covariates and $n = 1{,}500$. C2 has the same sample size, marginal density $p(x) = 1/2 + x$, and effect $f_1(x) = 6(x - 7/12)$. These two functions are exactly linear and have zero population mean. Their local-polynomial approximation error equals zero, but their empirical centers remain random.

C3 has $n = 3{,}000$ and one categorical active effect,

$$f_1(c) = 1.5\sin(2\pi c/24) + 0.7\cos(4\pi c/24), \qquad c = 0, \ldots, 23.$$

This categorical covariate has 24 equally likely levels. The remaining covariates are independent uniform continuous variables. C4 has $n = 3{,}000$ and two active effects, the continuous effect $6(x - 7/12)$ of a nonuniform covariate and the categorical effect above. The continuous covariates in C4 have density $1/2 + x$. All other effects are zero.

The design vector is stationary and is independently redrawn with probability 0.7 at each time. Errors follow an independent stationary Gaussian autoregression with coefficient 0.3 and standard deviation 0.5. Thus the signal-to-noise ratio is larger than in the primary simulations. This setting makes uncertainty in the covariate center statistically relevant without increasing local-polynomial bias. Selection, bandwidths, the deleted-block length and the number of resampling draws follow the same rules as in the primary simulations.

S26.2 A paired change in the centering calculation

The proposed full calculation and the fixed-center calculation use identical observations, selected support, fitted curves, deleted samples and resampling indices. The full calculation uses the exact deleted-block fit, including its new empirical center, whereas the fixed-center calculation recenters each deleted curve in the complete final sample. More precisely, if $\widehat{f}_{j,-b}$ is centered in the sample with block b removed, the fixed-center calculation replaces it by

$$\widehat{f}^{\,0}_{j,-b}(x) = \widehat{f}_{j,-b}(x) - P_N\widehat{f}_{j,-b}.$$

Here P_N is the empirical distribution of the complete final sample, and $P_N\widehat{f}_{j,-b}$ is the average of the deleted-block curve over that sample. The complete-sample estimate already satisfies $P_N\widehat{f}_j = 0$ and

Table S11: Centering variation and simultaneous component inference.

Design	Full coverage	Fixed-center coverage	Full width	Fixed-center width
C1	0.970	0.515	0.576 (0.057)	0.287 (0.025)
C2	0.975	0.505	0.567 (0.055)	0.298 (0.025)
C3	0.930	0.790	0.597 (0.033)	0.519 (0.028)
C4	0.965	0.550	0.561 (0.030)	0.395 (0.018)

Numbers in parentheses represent standard deviations. Both calculations use the same fitted effects and selected support.

is unchanged. We then form the deleted-block vectors and their studentized resampling distribution from these recentered curves. This change removes the first-order variation of the empirical center while retaining the estimated regression shape. It does not change the point estimate or its population target.

In the fixed-center calculation, the derivative of the fitted component with respect to an observation weight loses the centering term $-\widehat{f}_{ij}$ in (S5). The population counterpart of this term for a categorical level is $-f_j(X_{ij})$ in Proposition S7. The comparison isolates this single term of the proposed uncertainty calculation. It is not a comparison between different smoothing bandwidths or different selected models.

S26.3 Coverage and interval width

We use 200 independent replications per design. Every selected continuous component is evaluated at 29 equally spaced points between 0.15 and 0.85, and every categorical level is included. Simultaneous coverage requires every true active component to be selected and all selected component values to be covered. Mean width is the average over these evaluation points and categorical levels.

Table S11 shows coverage from 0.930 to 0.975 for the full calculation, compared with 0.505 to 0.790 for the fixed-center calculation. In C1 and C2, the mean widths of the full calculation are 0.5761 and 0.5672, compared with 0.2866 and 0.2977 when centering variation is removed. The difference is consistent with the exact centering-risk identity in (S51): an accurately estimated linear shape still has uncertainty about its population center.

Exact support recovery is 1.000, 1.000, 0.990 and 1.000 in C1–C4. Missing or additional variables cannot explain the coverage difference between the paired calculations, because their selected supports are identical within every replication. The categorical comparison also agrees with (S52). In this formula, the centering contribution is common to all levels. This contribution is needed for level inference but cancels from a contrast between two levels.

S27 Increasing grids and bands between evaluation points

A fixed evaluation grid gives intervals only at its own values and does not assess uncertainty over an entire covariate interval. Theorem 3 permits a grid that becomes finer as the sample size increases, and Corollary S3 specifies the additional width needed between the grid values. The experiment below examines this construction without changing the selected estimator.

S27.1 Designs and curvature bound

E1500, E3000 and E6000 use the nonuniform product design of D7 at epoch sizes 1,500, 3,000 and 6,000. M3000 and M6000 use the mixed design of D8 at epoch sizes 3,000 and 6,000. All five designs have dimension 200, five active effects, redraw persistence 0.3, error autoregression coefficient 0.3 and error standard deviation 1.5. We use 100 new independent replications per design.

For each selected continuous component, we use

$$M_N = \max\{29, \lceil 8/h_I \rceil + 1\}$$

equally spaced evaluation points on $[0.15, 0.85]$. Every level of a selected categorical variable is also included in the common resampling maximum. The support, estimation criterion and bandwidths

Table S12: Increasing evaluation grids and interpolation bands.

Design	N	Grid size	Coverage	Mean width	Added half-width
E1500	577	51	0.920	1.272 (0.074)	0.00196
E3000	1175	60	0.940	0.920 (0.039)	0.00141
E6000	2373	71	0.940	0.686 (0.021)	0.00100
M3000	1175	60	0.940	1.121 (0.048)	0.00141
M6000	2373	71	0.950	0.801 (0.022)	0.00100

Numbers in parentheses represent standard deviations. Grid size is per continuous component; every categorical level is included. Added half-width is the maximum curvature correction.

follow the primary procedure. The number of evaluation points per continuous component increases from 51 to 60 and 71 across the three epoch sizes and is of order h_I^{-1}, as required for a grid of comparable local resolution.

We use the fixed curvature bound $L = 80$ for all continuous effects. This value is justified directly by the generating functions. Their second derivatives have absolute bounds $8\pi^2$, 24, zero, $6.4\pi^2$ and $8 + 8\pi + 2\pi^2$, respectively, all of which are less than 80. Subtracting a population center leaves the second derivative unchanged, and a null component has second derivative zero. The curvature bound is thus specified for the complete generating class; it is not estimated from fitted curves.

S27.2 Evaluation between grid values

The joint critical value uses 999 studentized block resamples at all selected grid values. Between adjacent points x_a and x_{a+1}, we linearly interpolate the lower and upper limits and then subtract from the lower limit and add to the upper limit

$$40(x - x_a)(x_{a+1} - x).$$

This is the construction of Corollary S3 with $L = 80$. The maximum added half-width is $10\Delta_N^2$. Here $\Delta_N = 0.7/(M_N - 1)$. Categorical intervals are unchanged.

For the numerical check, we evaluate every interpolated continuous band on an additional grid of 1,001 points. This finite check is not used as a proof of continuum coverage. The deterministic curvature inequality gives coverage at every intervening point whenever all original grid values are covered. The computation also verifies this implication on the additional grid in every replication.

Table S12 reports the joint coverage of the original evaluation grid, which is also a sufficient event for coverage by the interpolated bands at every intervening point. An omitted active component counts as noncoverage. Mean width is first averaged over the 1,001 check points for a continuous component or over all categorical levels, and then over selected components. The added half-width column reports the maximum deterministic curvature correction.

S27.3 Coverage and resolution

Table S12 shows simultaneous coverage from 0.920 to 0.950. With 100 replications, a coverage probability of 0.95 has Monte Carlo standard error about 0.022. The 95% Wilson interval for the coverage probability contains 0.95 in every design. No replication with complete coverage at the original grid failed the additional interpolation check.

For the three nonuniform designs, mean band width decreases from 1.2718 to 0.9195 and 0.6856 across the three epoch sizes. The largest curvature correction decreases from 0.00196 to 0.00141 and 0.00100. The uncertainty from estimating the curve is larger than this interpolation correction at the evaluated resolutions. These results show that the denser grid permits simultaneous coverage between evaluation points with the stated curvature correction.

For the two mixed designs, mean band width decreases from 1.1207 to 0.8008 between epoch sizes 3,000 and 6,000. The categorical values require no interpolation, but they enter the same simultaneous maximum as the continuous effects. The comparison illustrates the common uncertainty calculation for continuous and categorical covariates. It does not establish an adaptive band for an unknown curvature class.

S28 CORRELATED DESIGNS AND PAIRED COMPARISONS

We specify the stationary model and paired comparisons for R1–R3 in Section S22.

S28.1 STATIONARY DISTRIBUTION AND REGRESSION FUNCTIONS

The first three predictors have the joint density (S9). The third predictor is the inactive binary variable and is stored as the two categories 0 and 1, corresponding to $Z = -1$ and $Z = 1$.

All other predictors are independent of this triple and of one another in the stationary distribution. Predictor 6 is uniform on the 24 integers from 0 to 23. Every remaining marginal is uniform on $[0, 1]$. We define $u(x) = \sqrt{3}(2x - 1)$; the five nonzero effects are

$$f_1(x) = f_2(x) = 1.25u(x), \qquad f_4(x) = 1.2\sin(2\pi x), \qquad f_5(x) = 1.5\cos(2\pi x),$$
$$f_6(c) = 1.3\sin(2\pi c/24), \qquad c = 0, \ldots, 23.$$

Every effect has zero mean under its marginal. All other effects equal zero. The error process has marginal standard deviation 1.5 and follows a stationary Gaussian autoregression with coefficient 0.3, independently of the entire design process. Its initial observation has this stationary Gaussian distribution.

We make an independent draw of the first two predictors by rejection sampling from independent uniform proposals, accepting a proposed pair with probability $\{1 + 0.1u(x_1)u(x_2)\}/1.3$. Given this pair, we draw $Z = 1$ with probability $\{1 + 0.99\,\mathrm{sign}(x_1 + x_2 - 1)\}/2$. These probabilities give exactly (S9); no numerical normalization of the density is used.

The process begins with an independent stationary draw. At each later time, it is redrawn independently with probability 0.5 in R1 and 0.25 in R2 and R3. Between such draws, a joint proposal moves the first two covariates by independent uniform increments on $[-0.35, 0.35]$, reflected at 0 and 1, and changes the sign of the binary coordinate with probability one half. We accept this joint proposal with the Metropolis probability $1 \wedge \{q(x_1', x_2', z')/q(x_1, x_2, z)\}$. Here q is the joint density of the first three predictors. The remaining continuous coordinates use independent increments with the same reflection rule. The categorical coordinate moves by -1, 0 or 1, with equal probabilities and addition modulo 24.

Because the reflected and discrete proposals are symmetric, the acceptance probability makes the transition for the correlated triple reversible with respect to its stated density. The other transitions preserve their uniform marginals. The product of these transitions preserves the stationary distribution Q and supplies the transition K in (S8). The independent-draw decisions, proposal randomness and Gaussian innovations are mutually independent. This construction verifies stationarity and the geometric dependence bound for the simulated process.

S28.2 SELECTION AND PAIRED ERRORS

The sample sizes, dimensions and common fitting rules are specified in Section S22. Table S13 shows that SCAD and Lasso-refit have the same true-positive rates and differ mainly in selecting the inactive binary variable. SCAD recovers the exact support while Lasso-refit does not in 64, 76 and 74 paired replications. No replication gives exact recovery for Lasso-refit but not for SCAD.

The mean paired decreases in component error from Lasso-refit to SCAD-refit are 0.02375, 0.01036 and 0.00509, with Monte Carlo standard errors 0.00424, 0.00155 and 0.00088. Replications in which SCAD-refit has the larger error and replications with equal error both remain in these averages. Absolute differences at most 10^{-10} count as ties; the tied fractions are 0.33, 0.21 and 0.25. Oracle errors are 0.217, 0.121 and 0.061, compared with the selected-fit errors in Table S5.

All selected continuous values and categorical levels enter the resampling maximum, with an omitted active group counting as noncoverage. The simultaneous coverage frequencies 0.96, 0.93 and 0.93 have Monte Carlo standard errors 0.020, 0.026 and 0.026, and Wilson intervals $[0.902, 0.984]$, $[0.863, 0.966]$ and $[0.863, 0.966]$. These intervals describe simulation uncertainty and all contain the nominal value 0.95.

Table S13: Selection of active effects and the inactive binary variable.

Design	Method	TPR	FDP	Exact	Z selected
R1	SCAD-refit	0.998	0.015	0.920	0.000
R1	Lasso-refit	0.998	0.130	0.280	0.650
R1	Lasso	0.998	0.137	0.240	0.740
R2	SCAD-refit	1.000	0.005	0.970	0.000
R2	Lasso-refit	1.000	0.139	0.210	0.770
R2	Lasso	1.000	0.124	0.270	0.700
R3	SCAD-refit	1.000	0.002	0.990	0.000
R3	Lasso-refit	1.000	0.127	0.250	0.730
R3	Lasso	1.000	0.133	0.210	0.790

Table S14: Initial-norm ordering and refitting diagnostics in the Gaussian-copula designs. Numbers in parentheses represent standard deviations.

Design	Reversed order	Best threshold exact	Minimum Hamming	Lasso-refit exact	Lasso-refit error	Oracle error
G1	1.000	0.000	1.00	0.000	0.0240 (0.0013)	0.0200 (0.0008)
G2	0.990	0.010	0.99	0.000	0.0178 (0.0011)	0.0148 (0.0006)
G3	1.000	0.000	1.00	0.000	0.0093 (0.0005)	0.0079 (0.0002)

S29 Additional comparisons for SCAD and continuing estimation

We report the initial-norm diagnostic, continuing-state details and online comparisons referenced in Section 4.

S29.1 Initial-norm ordering

For each Gaussian-copula replication, we order the initial group-Lasso norms and consider every distinct support obtained by applying a common cutoff. The minimum Hamming distance is the fewest group additions or removals needed to obtain the two-variable active set from any of these supports. Lasso-refit uses the complete nonzero initial support and the same unpenalized final fit as the other methods.

Table S14 gives a minimum Hamming distance near one. Lasso-refit retains the inactive third variable and never recovers the exact support. Lasso-refit and SCAD-refit use the same final unpenalized criterion.

S29.2 Resolution and latest-block estimation

The latest-block estimator uses the same support and point resolution as the continuing estimate but discards moments accumulated in earlier blocks with that support.

Table S15 gives inference fractions 0.821–0.889 and mean deletion-block counts 25.9–44.7. At the same support and resolution, the latest-block errors exceed the continuing errors in Table S9. This reduction shows the gain from retaining earlier moments.

Table S15: Resolution and interval diagnostics for continuing estimation. Numbers in parentheses represent standard deviations.

Design	Inference fraction	Point grid	Inference grid	Blocks	Last-block error	Mean width
A1	0.821	33	65	25.9	0.1715 (0.0345)	0.4904
A2	0.858	33	65	33.9	0.4020 (0.0911)	0.3884
A3	0.889	33	65	44.7	0.1515 (0.0292)	0.3181

Table S16: Component error with order-two cosine Sobolev tuning.

Design	Continuing	F-SGD	Sieve-SGD
A1	0.0094	0.1259	0.1128
A2	0.0073	0.1357	0.2246
A3	0.0055	0.0979	0.1373

S29.3 Comparison with online additive SGD

We compare the continuing estimator with F-SGD (Chen and Klusowski, forthcoming) and additive Sieve-SGD (Zhang and Simon, 2022). We use i to denote the observation index, j the basis index, J_i the number of basis functions per covariate, and $A, B > 0$ the tuning constants. Both estimators use the bounded orthonormal basis $\sqrt{2}\cos(\pi jx)$ and smoothness parameter 2. F-SGD uses its last iterate, step size A/i and $J_i = \lfloor Bi^{1/5} \rfloor$. Sieve-SGD uses Polyak averaging, j^{-4} component steps and $J_i = 1 \vee \lfloor Bi^{0.21} \rfloor$. Its overall step is $(A/d)i^{-1/5}$, scaled by the covariate dimension d.

The smoothness parameter 2 is not an oracle value for this cosine basis. For the first simulated effect $2\sin(2\pi x)$, direct integration gives, for odd j,

$$\int_0^1 2\sin(2\pi x)\sqrt{2}\cos(\pi jx)\,dx = \frac{8\sqrt{2}}{\pi(4-j^2)}.$$

The squared cosine coefficients of this effect weighted by j^4 have a divergent sum. This effect has bounded second derivatives, as used in Theorem 4, but does not belong to the order-two Sobolev ellipsoid for the stated basis (Zhang and Simon, 2022; Chen and Klusowski, forthcoming). The cited SGD guarantees assume independent observations; in A2 and A3 we compare the same implementations under temporal dependence.

For each design, we select the two tuning constants A and B by minimizing the mean component error over 20 independent pilot streams. They are fixed before the 100 production streams. These streams share observations, seeds and 41-point integration values with the continuing and planned-epoch fits. Every replication is retained.

Table S16 reports the A1–A3 mean errors at smoothness parameter two; the continuing mean errors are 92.5–94.6 percent lower than the F-SGD errors and 91.7–96.7 percent lower than the Sieve-SGD errors. Continuing error is lower in all 100 paired replications of each design. The SGD coefficient states are much smaller than the continuing state. The continuing state also retains the observations and moments needed for support selection and simultaneous uncertainty.

All five simulated effects have square-integrable first derivatives. Integration by parts and Bessel's inequality give, for each effect f,

$$\sum_{k\geq 1} k^2 \left\{\int_0^1 f(x)\sqrt{2}\cos(\pi kx)\,dx\right\}^2 \leq \pi^{-2}\int_0^1 \{f'(x)\}^2\,dx < \infty.$$

This implies that the order-one cosine Sobolev condition holds.

We compare both SGD estimators at this smoothness in the independent design A1, using the same 100 streams and continuing estimates. F-SGD uses $J_i = \lfloor Bi^{1/3} \rfloor$ and step A/i. Sieve-SGD uses $J_i = 1 \vee \lfloor Bi^{1/3+0.01} \rfloor$, step $(A/d)i^{-1/3}$ and coordinate weights j^{-2}. Its reported coefficients average the iterates, including the initial zero vector. No further upper bound is imposed on J_i.

Twenty independent pilot streams select the tuning pair (A, B) by minimum mean total component error. For F-SGD, the candidates are $A \in \{0.025, 0.05, 0.1, 0.2, 0.35, 0.5, 0.7, 1, 1.5, 2\}$ and $B \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.8\}$. For Sieve-SGD, they are $A \in \{0.5, 1, 2, 3, 4, 6\}$ and $B \in \{0.1, 0.25, 0.4, 0.6, 0.8, 1\}$. The selected pairs are $(1, 0.4)$ and $(3, 0.6)$, respectively. Both are interior to their grids. We fix these pairs before evaluating the 100 production streams, whose seeds are disjoint from the pilot seeds. These empirical constants are not asserted to satisfy every finite-constant condition of the cited rate theorems. Re-evaluating all production SGD fits at 161 integration points changes the 41-point errors by less than 4×10^{-15}.

Table S17: Component error in A1 after expanded independent tuning. Numbers in parentheses represent standard deviations.

Method	Total error	Active error	Inactive error
Continuing	0.00939 (0.00204)	0.00939	0.00000
F-SGD	0.03790 (0.00319)	0.01935	0.01854
Sieve-SGD	0.04176 (0.00405)	0.02025	0.02150

Table S17 shows lower continuing error in all 100 paired replications, with mean reductions of 75.2 and 77.5 percent relative to F-SGD and Sieve-SGD. The mean paired decreases are 0.02851 and 0.03236, with Monte Carlo standard errors 0.00030 and 0.00039. The continuing estimator has smaller active-component error, 51.5 and 53.6 percent below F-SGD and Sieve-SGD, and avoids estimating the fifteen inactive effects. Here the simulated effects satisfy the cosine Sobolev condition used for tuning.

S29.4 Comparison with online smooth backfitting

In the independent design A1, we compare with the online smooth backfitting estimator of Yang, Yao and Zhao (2024) with the identity link, using all twenty covariates, including the fifteen inactive ones. Its dynamic bandwidth updates use $K_h(v) = h^{-1}K(v/h)$ and five candidate states of marginal and pairwise product-kernel moments. The first block has 1,000 observations and later blocks have 500. Each component linearly interpolates its local intercepts at 41 equally weighted nodes on $[0, 1]$.

Twenty independent pilot streams choose the component bandwidth constants C_j in $h_{Nj} = C_j N^{-1/5}$ using the true component functions. We evaluate common constants $\{0.35, 0.55, 0.85, 1.3, 3, 8, 32\}$, choose the best value for each component, and compare that vector of constants with the best common setting. The vector has smaller pilot total error and is fixed for the 100 comparison streams. This oracle calibration permits bandwidths above one and differs from the published online plug-in selector. The streams and 41-point integration rule are the same as for continuing estimation.

Table S18: Component error under independent observations. Numbers in parentheses represent standard deviations.

Method	Total	Active	Inactive
Continuing	0.0094 (0.0020)	0.0094 (0.0020)	0.0000 (0.0000)
Online SBF	0.0240 (0.0036)	0.0149 (0.0028)	0.0091 (0.0021)

Table S18 gives total errors 0.0094 and 0.0240, a mean reduction of 60.8 percent. The paired difference is 0.01457 with Monte Carlo standard error 0.00035. Continuing estimation has smaller total and active-component errors in all 100 pairs; the mean active error is 37.0 percent lower, and correct selection removes inactive error. This comparison concerns the complete procedures, which have different smoothing criteria and rules for reporting component curves.

S29.5 Inference bandwidth and block length

The simultaneous intervals use the inference bandwidth $h_I = c_h N^{-1/4}$ and the block length $L_N = \lceil c_L N^{3/5} \rceil$ with the primary constants $(c_h, c_L) = (0.8, 0.4)$. We vary one constant at a time in D3, D7 and D8. Each design uses 100 new independent datasets whose seeds are disjoint from the primary runs. Within a dataset, all choices share the selected support and the point estimate, and only the inference state changes.

Table S19 gives simultaneous coverage from 0.920 to 0.980 across the fifteen design and constant pairs, and no calculation fails. With 100 replications, a coverage probability of 0.95 has Monte Carlo standard error about 0.022. Longer blocks and smaller bandwidth constants give wider intervals. The shortest blocks give the lowest coverage, 0.920–0.930.

Table S19: Simultaneous coverage for alternative inference-bandwidth and block-length constants.

Design	c_h	c_L	Coverage	Mean width	Failures
D3	0.70	0.40	0.940	1.314 (0.068)	0
D3	0.80	0.25	0.920	1.133 (0.055)	0
D3	0.80	0.40	0.930	1.216 (0.062)	0
D3	0.80	0.65	0.950	1.395 (0.105)	0
D3	0.90	0.40	0.930	1.139 (0.059)	0
D7	0.70	0.40	0.940	1.384 (0.093)	0
D7	0.80	0.25	0.920	1.179 (0.079)	0
D7	0.80	0.40	0.930	1.263 (0.086)	0
D7	0.80	0.65	0.960	1.442 (0.109)	0
D7	0.90	0.40	0.940	1.167 (0.079)	0
D8	0.70	0.40	0.980	1.155 (0.051)	0
D8	0.80	0.25	0.930	1.035 (0.039)	0
D8	0.80	0.40	0.970	1.087 (0.047)	0
D8	0.80	0.65	0.980	1.201 (0.058)	0
D8	0.90	0.40	0.970	1.030 (0.046)	0

Numbers in parentheses represent standard deviations. Coverage includes all selected values and counts a missing active component or a computational failure as noncoverage.

S30 Continuing inference with correlated coordinates

We extend the continuing procedure to correlated coordinates. The extension concerns the actual pooled criterion: the local lines share coefficients across blocks, while the node probabilities use each block's original bandwidth. Population additive coercivity replaces independence. For a fixed active distribution, bounded measurable pairwise density ratios suffice for joint inference. A dual approximation controls the categorical bias without derivatives of these ratios. Neither a product distribution nor a finite-rank dependence representation is required.

S30.1 Population conditions and the computed procedure

Write P_j for the marginal distribution of X_j, S for the active set and $s = |S|$. All constants below are fixed, $1 \le s < \infty$, and s does not increase. The dimension may increase under the selection restrictions below. For a continuous coordinate, the support is $[0, 1]$; categorical coordinates have finitely many levels. Norms without a subscript are Euclidean or operator norms, as appropriate.

Condition S2. The following population restrictions hold uniformly over the sample sizes under consideration. The active joint distribution Q_S is fixed. Additional null coordinates may be included as the sample size increases.

(a) The process starts in its stationary distribution Q. We write p_j for the density of a continuous marginal P_j. The centered effects satisfy $\sum_{j\in S} \|f_j\|_\infty^2 \le B^2$. Continuous effects and marginal densities have two bounded derivatives, and $0 < c_p \le p_j \le C_p$. Every categorical coordinate has at least two levels, a bounded number of levels, and all level probabilities are at least c_p. We write δ_x for the point mass at x. The stationary covariate transition and error covariance are

$$P(x, \cdot) = \rho\delta_x + (1-\rho)Q, \qquad \mathrm{Cov}(\varepsilon_i, \varepsilon_l) = \sigma^2\vartheta^{|i-l|}, \qquad 0 \le \rho \le \rho_0 < 1, \quad |\vartheta| \le \vartheta_0 < 1. \tag{S55}$$

The errors are centered and jointly Gaussian, independent of the complete covariate process, and $0 < \sigma_- \le \sigma \le \sigma_+$. The regression function is $F(X) = \mu + \sum_{j\in S} f_j(X_j)$.

(b) For all centered $g_j \in L^2(P_j)$, $j \in S$,

$$\mathbb{E}_Q\left\{\sum_{j\in S} g_j(X_j)\right\}^2 \ge \kappa \sum_{j\in S} \|g_j\|_{P_j}^2, \qquad \kappa > 0. \tag{S56}$$

For distinct active coordinates, let Q_{jk} be their joint marginal distribution and $q_{jk} = dQ_{jk}/d(P_j \otimes P_k)$. These ratios exist and are bounded by a fixed constant. They need not be continuous. No positive lower bound on a pairwise ratio or smooth full joint density is required.

(c) One of the following selection specifications holds. In the first specification, at retained selection size k, use $h_L \asymp k^{-1/5}$, $G_L \asymp h_L^{-1}$ and the population definitions of M_k, Γ_{M_k}, θ^*, θ^L, w and $\bar{\theta}$ preceding Theorem 1, with sample size k. The feature, projection-residual and coefficient bounds in Condition S1 hold, $S \subset M_k$, $\lambda_k = O(1)$, and

$$\lambda_{\min}(\Gamma_{M_k}) \geq \kappa_\Gamma > 0, \qquad m_k r_k = o(\lambda_k),$$
$$\min_{j \in S} \|\bar{\theta}_j\| \geq c\lambda_k, \qquad \min_{j \in M_k \setminus S} \left[w_j - \|\{\Gamma_{M_k}(\bar{\theta} - \theta^*)\}_j\| \right] \geq c\lambda_k. \tag{S57}$$

Here $m_k = |M_k| \leq d \wedge Ck^{1/5}$, as in the prescribed screening rule, and

$$r_k = \sqrt{G_L \log(dk)/k} + G_L \log k \log(dk)/k + h_L^2 \sqrt{G_L \log(dk)/k}.$$

If $m_k < d$, the population screening scores separate M_k from its complement by a fixed positive margin. This restriction is omitted when $m_k = d$. The inactive-margin inequality is omitted when $M_k = S$. These restrictions are uniform for $(n/\log n)^{3/4} \leq k \leq n^{3/4}$, including the retained sizes at temporary reports. A penalty multiplied by the empirical response standard deviation uses the corresponding population standard deviation in these population definitions.

In the second specification, $d \geq 2$ is fixed and all candidate marginals satisfy part (a). For centered $g \in L^2(P_l)$ define $T_{jl} g = \mathbb{E}\{g(X_l) \mid X_j\}$. Suppose

$$\max_{1 \leq j \leq d} \sum_{l \neq j} \|T_{jl}\|_{\mathrm{op}} \leq \gamma < \tfrac{1}{4}, \qquad \min_{j \in S} \|f_j\|_{P_j} \geq b > 0. \tag{S58}$$

Use $h_L \asymp k^{-1/5}$, $G_L \asymp h_L^{-1}$ and $\lambda_k = c_\lambda \widehat{\sigma}_Y \sqrt{\log d/(kh_L)}$, where $\widehat{\sigma}_Y$ is the retained sample standard deviation and c_λ is a sufficiently large fixed constant. Use the original screening cap $m_k = \min(d, \lceil 4k^{1/5} \rceil)$, SCAD shape $a > 2$, support tolerances and $100 + \lceil 30 \log k \rceil$ sweeps.

Part (b) identifies the population-centered effects. Part (c) concerns population minimizers and residuals, not successful sample selection. It is an explicit restriction: additive coercivity on S alone does not imply the selection margins on M_k. The examples below verify these premises. The second selection specification covers sufficiently weak correlation with the original penalty form and sweep cap. For continuous-effect risk, usable fractions and storage alone, bounded marginal densities and second derivatives of the effects suffice in (a), and the pairwise density restriction can be omitted from (b).

We retain the selection indices $\lfloor k^{4/3} \rfloor$, the time blocks $t_r = \lfloor r^{5/2} \rfloor$, support resets, early grid activation, point and inference bandwidths, temporary reports, and failure rules of Section S9. One computational addition is needed for the full class. If either convex selection fit has not attained its k^{-3} group optimality tolerance at the stated sweep limit, continue to a certificate by the following iteration on that same normalized criterion. For its matrix H, score b and fixed group weights ω_j, set $L_k = \lambda_{\max}(H)$ and restart at zero. Use

$$z^{(a)} = \theta^{(a)} - \{H\theta^{(a)} - b\}/L_k, \qquad \theta_j^{(a+1)} = \left(1 - \frac{\omega_j}{L_k \|z_j^{(a)}\|}\right)_+ z_j^{(a)}, \tag{S59}$$

with the value zero when $z_j^{(a)} = 0$. Stop when the maximum group optimality error defined in Section S20 is at most k^{-3}. A finite bound can be computed before the continuation. If $c_H = \lambda_{\min}(H) > 0$, $q_H = 1 - c_H/L_k$ and $D = \|\theta^{(1)} - \theta^{(0)}\|$, then at iterate $a+1$ the residual is at most $L_k q_H^a D$. Choose the least nonnegative integer a for which this bound is at most k^{-3}. If $q_H = 0$, one update is sufficient. Apply this rule separately to the initial and weighted fits. Reject an unstable matrix as before. This addition does not change the statistical criterion, support threshold, moments or access to observations. The second selection specification already gives convergence within the original sweep cap. The first specification can require additional iterations.

A report requires only its pooled fit and the deleted fits entering its covariance. An optional most-recent-block-only comparison is not allowed to invalidate these fits. This distinction matters in the reference implementation, whose general report also computes that comparison: a partial block can have only a fixed number of records even when the pooled sample tends to infinity. Both reporting procedures return an unavailable flag for that optional comparison if it fails, without suppressing an

otherwise valid pooled report. The reference implementation also omits partial blocks with fewer than 80 retained records. Keeping or omitting such a block has the same conclusions below; the proof permits any positive partial-block size and explicitly includes the empty case by omission. Neither convention needs a lower bound increasing with n on the size of the last block.

Lemma S1. *On a positive-definite normalized criterion, the iteration in (S59) reaches its tolerance after finitely many steps. If $H \succeq cI$, it needs at most $Cm_k \log k$ additional steps when the starting coefficient and score norms are polynomial in k. Under Condition S2(c), the final certified fit has the same selected support as the exact two-stage procedure with probability $1 - o(k^{-A})$ for every fixed $A > 0$.*

Proof. Because $H_{jj} = I$ and H is positive semidefinite, Cauchy–Schwarz in its quadratic form gives

$$\theta^{\mathsf{T}} H\theta \le \left(\sum_j \|\theta_j\| \right)^2 \le m_k \|\theta\|^2.$$

If $H \succeq cI$, the affine map in (S59) contracts distances by at most $1 - c/L_k$. Group soft thresholding is nonexpansive: this follows by adding the two subgradient inequalities for the convex group norm. The full iteration has the same contraction factor and its unique fixed point is the minimizer. Its update supplies a stationarity residual at $\theta^{(a+1)}$ equal to

$$v^{(a+1)} = (L_k I - H)(\theta^{(a)} - \theta^{(a+1)}).$$

Thus the computed optimality error is at most $L_k \|\theta^{(a)} - \theta^{(a+1)}\|$, which decreases geometrically. The iteration count follows from $L_k \le m_k$. Starting at zero, the minimizer norm is at most $\|b\|/c_H$: multiply its optimality equation by the minimizer and drop the nonnegative penalty. Thus $\lceil (L_k/c_H) \max\{0, \log(2L_k\|b\|/(c_H k^{-3}))\} \rceil + 2$ is another sufficient finite bound when $b \neq 0$. If the initial residual meets tolerance, return zero. For any positive minimum eigenvalue, the same argument proves finite termination, with a constant depending on that eigenvalue. A normalized minimum eigenvalue below k^{-1} is rejected when continuation is needed. Such a rejection has probability smaller than every fixed power of k under the population lower bound.

Under the first selection specification, strong convexity bounds the error of a fit with maximum group residual k^{-3} by $\sqrt{m_k} k^{-3}/c$. The SCAD derivative is Lipschitz in its norm argument with constant $1/(a-1)$. The error in the initial fit changes its weight vector by at most this constant times the initial coefficient error. Both errors are $o(\lambda_k)$ by (S57). The population selection argument gives active norms and inactive score gaps at least a fixed multiple of λ_k on the sample with probability $1 - o(k^{-A})$. Proposition S8, applied after the weight perturbation, then preserves its support. Lemma S2 below proves this probability assertion on the retained schedule. For the second specification, Lemma S2 proves the strict margins and convergence within the original sweep cap directly. This proof uses no fixed-sweep guarantee for an arbitrary coercive matrix. It concerns the exact-real-arithmetic procedure with a checked residual, not a finite-precision coverage guarantee. This completes the proof. □

The inference result below requires no continuity of a pairwise density ratio. A quantitative condition gives categorical risk rates and also permits changing active distributions with a common regularity bound. For $0 < \zeta \le 1$, suppose that every active pair satisfies

$$|q_{jk}(x, y) - q_{jk}(x', y)| \le C|x - x'|^{\zeta} \tag{S60}$$

in each continuous argument, uniformly in the other argument. No inequality is imposed in a categorical argument.

The next theorem includes selection, discarded prefixes, unsuccessful states and all required deleted fits. It does not condition on a correct selected set.

Theorem S7. *Under Condition S2, use the continuing procedure with the certificate addition just specified. At each deterministic reporting time n, $N_{E,n}/n \to_p 1$ and $N_{I,n}/n \to_p 1$. For each active continuous component,*

$$\mathbb{E}\|\widehat{f}_j^{\mathrm{C,T}} - f_j\|_{P_j}^2 \le C\{h_{E,n}^4 + (nh_{E,n})^{-1}\} \le Cn^{-4/5}, \qquad h_{E,n} = c_E n^{-1/5}. \tag{S61}$$

For $h_{I,n} = c_I n^{-1/4}$, fix a nonempty finite collection of active component values, with every continuous value in a fixed interior interval. The actual recomputed-center delete-block covariance and conditional resampling quantile satisfy

$$\Pr\left\{\max_{j,x}\frac{|\widehat{f}_j^{\mathrm{C}}(x) - f_j(x)|}{\widehat{\mathrm{se}}_j^{\mathrm{C}}(x)} \le c_{1-\alpha}^{\mathrm{C}}\right\} = 1 - \alpha + o(1). \tag{S62}$$

If the bounded pairwise ratios also satisfy the separate Hölder condition (S60) with exponent $0 < \zeta \le 1$, each categorical group has expected error at most $C\{n^{-1} + n^{-(4+2\zeta)/5}\}$. In particular, $\zeta \ge 1/2$ gives the rate n^{-1}. An unavailable report makes this event false. Continuous standard errors have order $(nh_{I,n})^{-1/2}$ and categorical standard errors have order $n^{-1/2}$, in probability. The storage bound of Proposition S6 is unchanged:

$$O\Big\{dn^{3/4} + s^2\ell^2_{n\log\log n}n^{2/5}(\log\log n)^{2/5} + s^2\ell^2_{n\log\log n}n^{9/10}(\log\log n)^{1/2}\Big\} \tag{S63}$$

with probability tending to one. All completed estimation observations remain discarded. The theorem concerns a fixed finite value set at a fixed time, not simultaneous coverage over reporting times or a growing value set.

We prove the theorem through the following lemmas. The population coordinate changes in these proofs are used only for analysis. The procedure does not estimate Q, a density ratio, an additive inverse, or any population coordinate matrix.

S30.2 Selection times and the historical bandwidths

Lemma S2. *Under either selection specification in part (c) of Condition S2, the probability of any incorrect or unavailable selection update after $n/\log n$, apart from the delay to the first eligible boundary, is $o(n^{-D})$ for every fixed $D > 0$. Both current states then equal deterministic known-support states with index sets $\mathcal{I}_n$ satisfying*

$$\begin{aligned} N = |\mathcal{I}_n| = n - o(n), \qquad & \min\mathcal{I}_n \ge cn/\log\log n, \\ J \asymp n^{2/5}, \qquad & L_{\max} = O(n^{3/5}), \qquad JL_{\max}/N = O(1). \end{aligned} \tag{S64}$$

Here J counts nonempty estimation blocks, including an incomplete reporting block when present. That block may have size one.

Proof. Let $a_l = \lfloor l^{4/3}\rfloor$ be the selection indices. The transition from X_{a_l} to $X_{a_{l+1}}$ is

$$\rho^{a_{l+1}-a_l}\delta_x + (1 - \rho^{a_{l+1}-a_l})Q.$$

Its failure to redraw has probability at most ρ_0. The Gaussian covariance on any retained subset is a principal submatrix of the physical-time covariance and has eigenvalues between $\sigma_-^2(1-\vartheta_0)/(1+\vartheta_0)$ and $\sigma_+^2(1+\vartheta_0)/(1-\vartheta_0)$. The latter bounds follow by extending a finite coefficient sequence by zero and using the Fourier series $\sum_{r\in\mathbb{Z}}\vartheta^{|r|}e^{irv} = (1-\vartheta^2)/(1-2\vartheta\cos v + \vartheta^2)$.

For the second selection specification, screening eventually retains all d candidates. At a common selection bandwidth the population normalized cross block is a product of two conditional-average contractions and T_{jl}. Its operator norm is at most $\|T_{jl}\|$. Nodal normalization matrices and cross blocks concentrate, with a probability complement smaller than any power of k. Here is a direct bound. A normalized matrix summand has norm CG_L, variance bound CG_L, and dimension $O(dG_L)$. Split the ordered retained indices into $\lceil C_A\log k\rceil$ residue classes and couple each class to independent stationary vectors. The coupling error is at most $k\rho_0^{C_A\log k}$. Matrix Bernstein, with both eigenvalue tails and a union over the fixed number of group pairs, gives error $O_A\{\sqrt{G_L\log^2 k/k} + G_L\log^2 k/k\} = o(1)$. The same argument bounds the empirical centers. Thus the empirical normalized row sum is at most $1/4$, with positive slack, and its minimum eigenvalue is bounded away from zero. Raw nodal eigenvalues are at least of order $k^{-3/5} \gg k^{-3}$, and all numerical rejection thresholds are inactive.

Represent each active component by its Taylor lines and recenter them empirically. Their coefficient norms tend to $\|f_j\|_{P_j} \ge b$. Write θ^0 for these coefficients and e for the normalized score error

relative to their empirical quadratic. For a grouped vector v, write $\|v\|_{2,\infty} = \max_j \|v_j\|_2$. The response residual after this representation is the Gaussian error plus a uniformly bounded term of size Ch_L^2. Cauchy–Schwarz in the empirical within-group norm bounds the residual contribution to every normalized score by Ch_L^2. Conditional on all retained designs, each Gaussian group score has covariance at most vI/k in a dimension at most $2G_L$. The Gaussian exponential moment of its squared norm implies

$$\Pr\{\|\text{group score}\| > C\sqrt{G_L/k} \mid X\} \le e^{-cG_L},$$

with constants made as large as required by increasing C. For example, let λ_r be its covariance eigenvalues and let Z_r be independent standard normal variables. Use $\mathbb{E}\exp(t\sum_r \lambda_r Z_r^2) = \prod_r (1 - 2t\lambda_r)^{-1/2}$ at $t = k/(4v)$. The sample standard deviation of Y is bounded above and away from zero with a probability complement smaller than any power of k. To see this, apply the preceding residue-class scalar bound to the bounded signal and use the same Gaussian quadratic exponential moment for the centered error vector; the Gaussian linear cross term has variance $O(k)$. The lower covariance eigenvalue supplies the lower variance bound. A sufficiently large fixed c_λ hence gives $\|e\|_{2,\infty} \le \lambda_k/4$.

We give the deterministic selection argument. If $\delta = \widehat{\theta}^L - \theta^0$, the group optimality equations give $\|\delta\|_{2,\infty} \le (\lambda_k + \|e\|_{2,\infty})/(1 - \gamma_H) \le 5\lambda_k/3$, where $\gamma_H = \max_j \sum_{l \ne j} \|H_{jl}\|_{\mathrm{op}} \le 1/4$. For an inactive group its soft-threshold argument has norm at most $\lambda_k/4 + (1/4)(5\lambda_k/3) < \lambda_k$, and its coefficient is exactly zero. The fixed active norm lower bound puts every active initial norm above $a\lambda_k$ for large k. The SCAD weights are thus zero on S and λ_k elsewhere. The restricted unpenalized error is at most $\lambda_k/3$ in maximum group norm, and its inactive score is at most $\lambda_k/4 + \lambda_k/12 < \lambda_k$. This proves the weighted support assertion. A group-coordinate sweep contracts maximum group error by at most $1/4$: apply nonexpansiveness successively to already updated and not yet updated groups. Thus the original sweep cap yields error at most $C4^{-30\log k} = O(k^{-30\log 4})$ from a bounded initial iterate. Its stationarity residual is below k^{-3}, and the strict inactive gaps give exactly zero inactive groups in the last sweep. This proves the second specification without additional optimization. For the first specification we use the following concentration calculation.

The population comparison in Theorem 1 applies after the following direct concentration calculation for the retained schedule. Conditional on all retained redraw indicators, the retained observations are constant over segments of lengths $L_1, L_2, \dots$ and their distinct segment values are independent Q draws. The lengths need not be independent. Couple their indicators to independent forced redraws with probability $1 - \rho_0$. Additional redraws only split these forced segments. Thus their sum of squared lengths is bounded by the corresponding sum for forced segments, and their maximum length is bounded by the maximum forced length. The latter segments have independent geometric lengths. Truncating the first $k+1$ geometric lengths at $C_A \log k$ and applying scalar Bernstein to their squares proves

$$\sum_a L_a^2 \le Ck, \qquad \max_a L_a \le C_A \log k$$

except on a set of probability $o(k^{-A})$ for any fixed A. Indeed, the truncation error is at most $(k+1)\rho_0^{C_A \log k}$, and the probability that the truncated square sum exceeds twice a fixed multiple of k is at most $\exp\{-ck/\log^2 k\}$. This argument includes the initial and terminal pieces.

For a centered matrix or vector g, the sample sum is conditionally $\sum_a L_a g(W_a)$, with independent $W_a \sim Q$. A normalized moment block has norm bound CG_L and matrix variance bound CG_L per observation; the cross-block variance uses the pair-moment bound in part (c). Conditional matrix Bernstein now has variance parameter CkG_L and range $C_A G_L \log k$. It gives block error

$$C_A\left\{\sqrt{G_L \log(dk)/k} + G_L \log k \log(dk)/k\right\}.$$

For a feature mean or bounded-response score use the column-vector version, whose scalar variance parameter is the expected squared vector norm, at most CG_L. For the projection residual it is at most $CG_L h_L^4$. The signal is bounded by $\sqrt{s}B$, and the independent Gaussian error score is controlled conditionally by the retained error covariance bound. Its covariance trace is at most CkG_L, and diagonalizing that covariance gives a norm bound at most $C_A\sqrt{kG_L \log(dk)}$. These estimates yield $C_A r_k$ uniformly over all groups before screening, with exceptional probability $o(k^{-A})$. A

union bound over groups changes $\log k$ in the moment bound to $\log(dk)$, but does not multiply the redraw-length exceptional probability by d.

The centering and normalization comparison can be made explicit. In the population-normalized coordinates, let $\widehat{D}_j$ be the empirical within-group augmented matrix and $\widehat{c}_j$ its empirical feature mean. Their centered matrix is

$$W_j = \widehat{D}_j - \widehat{c}_j\widehat{c}_j^{\mathsf{T}} = I + O_{\mathrm{op}}(r_k).$$

Its aligned empirical feature is $W_j^{-1/2}(z_j - \widehat{c}_j)$. For $\|W_j - I\| \le 1/2$, scalar spectral calculus gives $\|W_j^{-1/2} - I\| \le C\|W_j - I\|$. The empirical cross block is

$$W_j^{-1/2}\left\{k^{-1}\sum_l z_j(X_{a_l,j})z_v(X_{a_l,v})^{\mathsf{T}} - \widehat{c}_j\widehat{c}_v^{\mathsf{T}}\right\}W_v^{-1/2}, \qquad j \ne v.$$

Substitution of the proved moment bounds gives an $O(r_k)$ error per block and the same error for each centered score. Any other normalized basis differs by a within-group orthogonal matrix, which leaves coefficient norms, the penalty and selection unchanged. The response scale is controlled by the preceding scalar calculation for the bounded signal, the Gaussian square exponential moment, and the conditional Gaussian signal–error cross term. Because $\sigma_Y \ge \sigma_-$, its relative error is at most $C_A\{\sqrt{\log k/k} + \log^2 k/k\}$. Multiplication by $\sqrt{m_k}\lambda_k$ is bounded by $C_A m_k r_k$. A strict population screening gap is preserved by these bounds on squared scores; when $m_k = d$, screening is exact by definition. This establishes the required rate without an extra factor from a residue-class union bound. The remaining perturbation argument is deterministic. The normalized Gram perturbation is at most $C_A m_k r_k$. Its positive minimum eigenvalue is preserved. Subtract the sample and population initial optimality equations and use monotonicity of the group-norm subgradient. The coefficient error is at most $C_A m_k r_k$, because the population coefficient norm is bounded. The SCAD weight error has the same bound. On S, the restricted weighted coefficient error is at most $C_A m_k r_k$. Each row from S to an inactive group has norm at most $\sqrt{s}$, by positive semidefiniteness and identity diagonal blocks. Its inactive score error is again at most $C_A m_k r_k$. Because $m_k r_k = o(\lambda_k)$, the active norms remain positive and every inactive inequality is strict. The restricted solution is the full minimizer. The coefficient and residual bounds in part (c) are exactly the population bounds needed for the preceding moment and perturbation estimates; no empirical expansion is assumed. The certificate proof above transfers the result to the computed fits. Choosing the constants in the concentration inequalities for any fixed A gives a support-error probability $o(k^{-A})$, uniformly over the stated retained sizes.

Put $k_0 = \lceil (n/\log n)^{3/4} \rceil$. There are at most $n^{3/4} + 1$ relevant retained sizes, including a temporary query. A union bound gives

$$\Pr(\text{some later selection fails}) \le (n^{3/4} + 1) \sup_{k \ge k_0} o(k^{-A}) = o(n^{-D})$$

by taking A sufficiently large. The rule requiring 75 new observations delays the first such update by $o(n/\log n)$ in physical time, and the next block boundary adds $O((n/\log n)^{3/5})$. Both are $o(n/\log\log n)$.

For the current grid, its first reporting time u_l is between cn and n, by nesting and slow variation of ℓ_n. Its activation is of order $n/\log\log n$. Thus all incorrect supports and resets precede activation on the event just proved. Remove the deterministic activation prefix, the selection indices, and the block meeting the activation boundary. Their counts are $O(n/\log\log n)$, $O(n^{3/4})$ and $O(n^{3/5})$. All sufficiently late complete blocks exceed the minimum count 80. The incomplete block follows its prescribed temporary rule. These facts give $N = n - o(n)$ and the deterministic oracle index set. Finally, $t_r \asymp r^{5/2}$ gives $J \asymp n^{2/5}$ and $L_{\max} = O(n^{3/5})$. Stochastic arguments below are proved on this deterministic state and then transferred by equality of procedures; they do not condition the data distribution on selection success. This completes the proof. □

Historical averaging has a critical fourth moment at the inference bandwidth. We keep this factor explicitly.

Lemma S3. *For either current state put $h = cn^{-\gamma}$, with $\gamma = 1/5$ or $1/4$, and let h_i be the original bandwidth of observation $i \in \mathcal{I}_n$. Write $H_r = N^{-1}\sum_i h_i^r$, $h_{\max} = \max_i h_i$ and $T_n = \max_i h_i/h$.*

Then

$$h \le h_i, \quad T_n \le C(\log\log n)^\gamma, \quad H_r \le Ch^r \ (r=1,2,3), \quad H_2 \asymp h^2, \quad H_4 \le Ch^4 B_n, \tag{S65}$$

where $B_n = 1$ for $\gamma = 1/5$ and $B_n = 1 + \log\log\log(n + e^{e^e})$ for $\gamma = 1/4$. At least a fixed fraction of the retained indices have h_i/h bounded by a fixed constant. The current grid spacing satisfies $\Delta \asymp h/\ell_n$.

Proof. For a completed block ending at t_r, $h_i = ct_r^{-\gamma}$ and $t_r/i = 1 + O(i^{-2/5})$ uniformly after activation. Thus $h_i = ci^{-\gamma}\{1 + o(1)\}$. A partial block uses h and satisfies all upper bounds directly. Integral comparison on $[cn/\log\log n, n]$ gives, for $r\gamma < 1$,

$$\frac{1}{N}\sum_i h_i^r \le \frac{Cn^{1-r\gamma}}{N} \le Ch^r.$$

For $r\gamma = 1$, the integral is at most $C\log\{n/(cn/\log\log n)\} \le CB_n$, giving $H_4 \le Ch^4 B_n$ in the inference state. The last half of physical time contains a fixed fraction of the retained indices and has bounded h_i/h, proving the remaining moment statements. The definition of the nested reporting grid gives $\ell_n \le h/\Delta < 2\ell_n$. This completes the proof. □

S30.3 The actual pooled operator and evaluation rows

For this subsection fix the deterministic oracle state. Use the common powers $(1, (x-u_g)/h)$ at each continuous node; this is a nonsingular change of coordinates from the unscaled powers stored by the procedure. Let $v_j(g,x)$ be the raw vector with these two entries in node g and zeros elsewhere. For a categorical group, let $v_j(x)$ be its level indicator vector. Define

$$\begin{aligned} a_{ij}(x) &= \sum_g \pi_{i,jg}(x) v_j(g,x), \qquad & D_{ij}(x) &= \sum_g \pi_{i,jg}(x) v_j(g,x) v_j(g,x)^\mathsf{T}, \\ \pi_{i,jg}(x) &= \frac{q_g K((x-u_g)/h_i)}{\sum_a q_a K((x-u_a)/h_i)}, \qquad & \bar D_j &= \frac{1}{N}\sum_i \mathbb{E} D_{ij}(X_j), \quad \bar c_j = \frac{1}{N}\sum_i \mathbb{E} a_{ij}(X_j). \end{aligned} \tag{S66}$$

Here $K(t) = 3(1-t^2)_+/4$ and q_g are the trapezoidal grid weights. The categorical definitions have no bandwidth. Let d_j represent the constant polynomial one, or the all-one level vector. Then $\bar D_j d_j = \bar c_j$ and $d_j^\mathsf{T} \bar D_j d_j = 1$. Choose a deterministic matrix R_j with

$$\bar c_j^\mathsf{T} R_j = 0, \quad R_j^\mathsf{T} \bar D_j R_j = I, \qquad R_j R_j^\mathsf{T} = \bar D_j^{-1} - d_j d_j^\mathsf{T}. \tag{S67}$$

Such a matrix is obtained by taking an orthonormal basis perpendicular to $\bar D_j^{1/2} d_j$ and multiplying by $\bar D_j^{-1/2}$. This construction is not performed by the estimator.

Conditionally on (i, X), let auxiliary nodes U_j be independent with probabilities $\pi_{i,jg}(X_j)$. They only express the sums in the criterion. Put

$$\begin{aligned} \phi_i &= (1, (R_j^\mathsf{T} v_j(U_j, X_j))_{j\in S})^\mathsf{T}, \quad & M_i &= \mathbb{E}_U(\phi_i \phi_i^\mathsf{T} \mid X_i), \\ z_i &= \mathbb{E}_U(\phi_i \mid X_i), & \Gamma &= \frac{1}{N}\sum_i \mathbb{E} M_i. \end{aligned} \tag{S68}$$

We write z_{ij} for the group-j block of z_i. The population diagonal blocks of Γ are identity matrices, and its intercept cross blocks are zero. For a reported pair $k = (j,x)$, let e_k be a row vector with $a_{hj}(x)^\mathsf{T} R_j$ in group j and zeros elsewhere; a_{hj} uses the reporting bandwidth h. We also write $e_j(x) = e_k$ for this row.

Lemma S4. *Under parts (a)–(b) of Condition S2, the pooled matrix and rows satisfy*

$$cI \preceq \Gamma \preceq CI, \quad \|M_i\| \le CG\{1 + (h_i/h)^2\}, \quad \|\mathbb{E} M_i\| \le C\{1 + (h_i/h)^2\}. \tag{S69}$$

For continuous j,

$$\sup_x \|e_j(x)\|^2 \le C/h, \qquad \left\| \int e_j(x)^\mathsf{T} e_j(x)\, dP_j(x) \right\| \le C. \tag{S70}$$

For categorical j these bounds hold with h replaced by one. All constants are independent of the number of grid nodes and the historical bandwidth ratio.

Proof. The kernel denominator in (S66) is between ch_i and Ch_i, including boundary points, because $\Delta/h_i \le 1/3$. Each 2 by 2 scaled nodal block of $\bar{D}_j$ has upper eigenvalue at most $C\Delta\{1+H_2/h^2\} \le C\Delta$. On the last half of the sample, h_i/h is bounded. Integrating over an interval of length ch within the kernel support gives a lower eigenvalue $c\Delta$: the moment matrix of $(1,t)$ on an interval of fixed positive length, possibly one-sided, is positive definite with a uniform lower bound. Thus every scaled nodal block has eigenvalues in $[c\Delta, C\Delta]$. The same bounds hold for the one-bandwidth matrix $D_{j,h} = \mathbb{E}D_{hj}(X_j)$, and $\bar{D}_j \asymp D_{j,h}$. The categorical assertion follows from the level probabilities.

At bandwidth h, write a component polynomial as $H_j(U_j, X_j)$ and let $m_j(X_j) = \mathbb{E}_U H_j$. After centering each component at its reference population mean, conditional independence of the auxiliary nodes and (S56) give

$$\mathbb{E}\left(\alpha + \sum_j H_j\right)^2 = \alpha^2 + \sum_j \mathbb{E}\,\mathrm{Var}_U(H_j \mid X_j) + \mathbb{E}\left(\sum_j m_j(X_j)\right)^2$$
$$\ge \min(1,\kappa)\left\{\alpha^2 + \sum_j \mathbb{E}H_j^2\right\}.$$

This independence is only the notational independence of the auxiliary nodes; X_j need not be independent. For the last half of the observations, $h \le h_i \le Ch$ and $\pi_{i,jg}(x) \ge c\pi_{h,jg}(x)$ on the latter support. Multiplying this inequality over the fixed s groups bounds the pooled augmented criterion below by a constant times the reference criterion.

To compare reference centering with pooled centering, decompose a pooled-centered raw coefficient as $\beta = \beta^0 + \mu_h d_j$, where β^0 has reference mean zero. Because $\bar{c}_j^{\mathsf{T}}\beta = 0$, $|\mu_h|^2 = |\bar{c}_j^{\mathsf{T}}\beta^0|^2 \le (\beta^0)^{\mathsf{T}}\bar{D}_j\beta^0 \le C(\beta^0)^{\mathsf{T}}D_{j,h}\beta^0$. Let $\alpha_h = \alpha + \sum_j \mu_{h,j}$ be the reference intercept. The inequalities just proved imply

$$\alpha^2 + \sum_j \beta_j^{\mathsf{T}}\bar{D}_j\beta_j \le C\left\{\alpha_h^2 + \sum_j (\beta_j^0)^{\mathsf{T}}D_{j,h}\beta_j^0\right\}.$$

This proves the lower bound for Γ in the pooled coordinates. The upper bound follows from $(\alpha + \sum_j H_j)^2 \le (s+1)(\alpha^2 + \sum_j H_j^2)$ and (S67).

At observation i, the squared norm of the scaled raw features is at most $C\{1 + (h_i/h)^2\}$ before normalization. The inverse nodal bound gives the individual bound for M_i. Integrating each nodal block over its support of length $O(h_i)$ gives the same bound without G for $\mathbb{E}M_i$. The cross blocks satisfy the corresponding bound by Cauchy–Schwarz and fixed s.

At the reporting bandwidth, at most Ch/Δ raw nodal blocks of $a_{hj}(x)$ are nonzero, each with norm at most $C\Delta/h$. The inverse nodal bound gives $\|e_j(x)\|^2 \le C/h$. Conditional Jensen gives $\int a_{hj}a_{hj}^{\mathsf{T}}\,dP_j \preceq D_{j,h} \preceq C\bar{D}_j$. Multiplication by R_j proves the integrated bound. This also proves the categorical version directly. This completes the proof. □

The next bound prevents a factor $h^{-1/2}$ from multiplying the full correlated projection error. A bound on $\|\Gamma^{-1}\|$ alone would not suffice at the stated inference bandwidth.

Lemma S5. *Put $C_n = \Gamma - I$. For any reported component row,*

$$\|e_k C_n\| \le C, \qquad e_k\Gamma^{-1} = e_k - e_k C_n \Gamma^{-1}. \tag{S71}$$

For a continuous group define

$$K_{ij}(x,z) = a_{hj}(x)^{\mathsf{T}}\bar{D}_j^{-1}a_{ij}(z). \tag{S72}$$

Then $e_j(x)z_{ij}(z) = K_{ij}(x,z) - 1$, and

$$K_{ij}(x,z) = 0 \text{ if } |x-z| > h + h_i, \qquad \int |K_{ij}(x,z)|\,dP_j(z) \le C(1 + h_i/h). \tag{S73}$$

In particular, the second term in (S71), applied to the mean feature z_i, has squared L^2 norm averaged over i bounded by a constant.

Proof. The identity $e_j z_{ij} = K_{ij} - 1$ follows from (S67). The raw inverse is block diagonal over nodes. A common nonzero node requires both $|x - u_g| \le h$ and $|z - u_g| \le h_i$, proving the support assertion. Also,

$$\int \pi_{i,jg}(z) \left(1 + \frac{|z - u_g|}{h}\right) dP_j(z) \le C\Delta(1 + h_i/h).$$

The inverse block has norm at most C/Δ, while the norms of the reporting raw blocks sum to at most a constant. Summing gives (S73).

Only off-diagonal group blocks occur in C_n. For $k \neq j$ and a unit coefficient vector b_k, put $m_{ik}(X_k) = z_{ik}^{\mathsf{T}} b_k$. The pairwise upper density bound gives

$$\begin{aligned} |e_j(C_n)_{jk} b_k| &\le \frac{C}{N} \sum_i (1 + h_i/h) \mathbb{E}|m_{ik}(X_k)| \\ &\le C \left\{ \frac{1}{N} \sum_i (1 + h_i/h)^2 \right\}^{1/2} \left\{ \frac{1}{N} \sum_i \mathbb{E} m_{ik}^2 \right\}^{1/2} \le C. \end{aligned}$$

The last average is at most one by conditional Jensen and normalization. There are a fixed number of active groups. For a categorical row, bounded row norm and bounded Γ give the same result. The resolvent identity is exact. Finally, $N^{-1} \sum_i \mathbb{E} z_i z_i^{\mathsf{T}} \preceq \Gamma$, which proves the averaged L^2 bound for its correction. No factorization of a mixture of bandwidth-specific pairwise operators has been used. This completes the proof. □

S30.4 Population approximation and categorical bias

Let Π_n denote orthogonal projection in the Hilbert space with distribution

$$i \sim \mathrm{Unif}(\mathcal{I}_n), \quad X \sim Q, \quad U_j \mid (i, X) \sim \pi_{i,j\cdot}(X_j)$$

onto the intercept and the common local additive coefficient space. Define $\eta^* = \Gamma^{-1} N^{-1} \sum_i \mathbb{E}(\phi_i F)$ and $g_k^* = e_k \eta^*$. These are population projections of the actual heterogeneous-bandwidth criterion.

Lemma S6. *For any active continuous value, including the boundary for the upper bound,*

$$|g_j^*(x) - f_j(x)| \le Ch^2. \tag{S74}$$

For a fixed active distribution and a categorical coordinate and level x,

$$|g_j^*(x) - f_j(x)| = o(h^2). \tag{S75}$$

The full augmented residual and each component residual satisfy

$$\max_i \operatorname*{ess\,sup}_{X,U} |F(X) - \phi_i^{\mathsf{T}} \eta^*| \le C\{h_{\max}^2 + \sqrt{G}(1 + T_n)h^2\} =: E_n. \tag{S76}$$

For a component residual, replace F and $\phi_i^{\mathsf{T}} \eta^*$ by f_j and its fitted local component. At $h = c_I n^{-1/4}$, $E_n \to 0$. The biases in (S74) and (S75) are $o((nh)^{-1/2})$ and $o(n^{-1/2})$, respectively.

Proof. First project each f_j separately with its pooled marginal criterion, and call its coefficients η_j^0. At a continuous node use the Taylor line $f_j(u_g) + f_j'(u_g)(x - u_g)$. Its residual at observation i is bounded by Ch_i^2. Its pooled nodal score in scaled coordinates is at most $C\Delta(H_2 + H_3/h) \le C\Delta h^2$. The inverse nodal bound in Lemma S4 gives coefficient error $O(h^2)$ in both scaled entries. The historical residual r_{ij}^0 thus satisfies $|r_{ij}^0| \le Ch_i^2$, because $h^2(1 + h_i/h) \le 2h_i^2$. The reported error is $O(h^2)$. Constants are reproduced, and hence the pooled mean of each separate projection is zero. Categorical components are represented exactly.

Stack these projections and the intercept into η^0, and put $r_i^0 = \sum_j r_{ij}^0$. The individual moment bound in (S69) and Cauchy–Schwarz give

$$\begin{aligned} &d_i = \mathbb{E}(\phi_i r_i^0), && \|d_i\| \le \|\mathbb{E} M_i\|^{1/2} \|r_i^0\|_2 \le C(1 + h_i/h)h_i^2, \\ &d = N^{-1} \sum_i d_i, \quad \|d\| + \|\eta^* - \eta^0\| \le C(H_2 + H_3/h) \le Ch^2, && \eta^* - \eta^0 = \Gamma^{-1} d. \end{aligned} \tag{S77}$$

The actual joint projection residual $R_i = F - \phi_i^{\mathsf{T}}\eta^*$ obeys

$$\|R_i\|_2 \le Ch_i^2 + C(1 + h_i/h)h^2 \le Ch_i^2. \tag{S78}$$

These two bounds use additive coercivity and marginal moment bounds only.

The own-group and intercept scores of each separate residual are zero after pooling. For a continuous reported row, its remaining scores use other coordinates. The bounded pairwise ratios and (S73) imply $|e_j d| \le CN^{-1}\sum_i (1 + h_i/h)h_i^2 \le Ch^2$. Apply the exact row identity (S71) to d. Its correction is $O(h^2)$ by (S77), which proves (S74). The feature bound gives (S76) for the full residual and for each component. The intercept correction is zero because its pooled residual score is zero and its cross blocks in Γ are zero.

For a categorical value, consider $\mathcal{H} = \bigoplus_{j\in S} L_0^2(P_j)$, $Ag = \sum_j g_j(X_j)$ and $T = A^*A$. Additive coercivity gives $\kappa I \preceq T \preceq sI$. The bounded self-adjoint operator T is onto: its range is closed by the lower bound, and the orthogonal complement of that range is its zero kernel. For the requested level x of coordinate c, define $r \in \mathcal{H}$ by $r_c(z) = \mathbf{1}\{z = x\}/P_c\{x\} - 1$ and $r_j = 0$ for $j \ne c$. There is a unique $a = T^{-1}r$, with $\|a\|_{\mathcal{H}} \le \kappa^{-1}\|r\|_{\mathcal{H}}$. Put $G_c = Aa$. The normal equations imply

$$\mathbb{E}(G_c \mid X_j) = r_j(X_j), \qquad \mathbb{E}G_c = 0. \tag{S79}$$

For any common local additive coefficient vector β, conditioning on each coordinate and its auxiliary node gives $\langle G_c, \phi^{\mathsf{T}}\beta\rangle = \beta_c(x) - \mathbb{E}\beta_c(X_c)$. Here $\beta_c(x)$ denotes the categorical function represented by the group-c coefficients of β. This is exactly the categorical report in the pooled-centered coordinates. Also $\langle G_c, F\rangle = f_c(x)$. Orthogonal projection gives

$$g_c^*(x) - f_c(x) = -\langle G_c, (I - \Pi_n)F\rangle = -\langle (I - \Pi_n)G_c, (I - \Pi_n)F\rangle. \tag{S80}$$

The functions a_j are fixed when the active distribution is fixed. For any $\epsilon > 0$, choose fixed twice continuously differentiable functions $a_{j,\epsilon}$ with $\|a_j - a_{j,\epsilon}\|_{P_j} \le \epsilon$ on continuous coordinates. Such functions are dense: finite interval step functions approximate an L^2 function, and rounding their finitely many jumps on intervals of arbitrarily small total length gives smooth approximants. The bounded marginal density transfers this approximation to $L^2(P_j)$. Center the approximants and keep categorical components exact. Their sum $G_{c,\epsilon}$ satisfies $\|G_c - G_{c,\epsilon}\|_2 \le C\epsilon$. Let $G_{c,\epsilon}^T$ be their common nodal Taylor representation. Its error at historical index i is bounded by $C_\epsilon h_i^2$. Orthogonality in (S80) and (S78) now give

$$|g_c^*(x) - f_c(x)| \le C\epsilon H_2 + C_\epsilon H_4. \tag{S81}$$

For each fixed ϵ, $H_4/h^2 \le Ch^2 B_n \to 0$. First let n tend to infinity and then let ϵ tend to zero. This proves (S75). The conclusion is pointwise in the fixed active distribution, not uniform over arbitrary changing distributions whose dual functions need increasingly fine approximations.

At inference bandwidth, $\sqrt{nh}\,h^2 = O(n^{-1/8}) \to 0$ and $\sqrt{n}\,o(h^2) = o(1)$. Because $G = O(\ell_n/h)$ and T_n grows only as a power of $\log\log n$, $E_n \to 0$ and $\sqrt{G}(1 + T_n)E_n \to 0$. This completes the proof. □

The continuous components of the dual satisfy $a_j(x) = -\sum_{l\ne j}\int q_{jl}(x, y)a_l(y)dP_l(y)$. Cauchy–Schwarz and (S60) give a bounded Hölder norm of order ζ for each such a_j. Using its nodal values with zero slopes gives squared approximation error at most $CH_{2\zeta}$. Equation (S80) yields

$$|g_c^*(x) - f_c(x)| \le C\sqrt{H_{2\zeta}H_4}. \tag{S82}$$

At point bandwidth this is $O(h^{2+\zeta})$. At inference bandwidth its product with $\sqrt{n}$ is $O(n^{-\zeta/4}\sqrt{B_n}) = o(1)$. Thus the same inference proof holds for sequences of active distributions satisfying these common bounds, without the fixed-distribution premise.

S30.5 Concentration and expected risk

We give the concentration estimates used for the whole fit and for every block deletion. Throughout this subsection $p_n = (\log n)^C$ denotes a deterministic bound with some fixed sufficiently large exponent. It bounds only stochastic remainders and bandwidth ratios; the sharper factor B_n in the population bias is kept separately.

Lemma S7. *For the deterministic oracle state, $\widehat{M} = N^{-1}\sum_i M_i$ has eigenvalues in a fixed positive interval except on a set of probability $o(n^{-D})$, for any fixed $D > 0$. In either the point or inference state, define*

$$u_i = \mathbb{E}_U\{\phi_i(Y_i - \phi_i^{\mathsf{T}}\eta^*) \mid X_i, Y_i\}.$$

Uniformly over all estimation blocks b with $a = |b| \geq 1$,

$$\begin{aligned}
\|\widehat{M} - \Gamma\| &= O_p(p_n\sqrt{G/N}),\\
\left\|\frac{1}{N}\sum_i u_i\right\| + \left\|\frac{1}{N}\sum_i z_{ij}\right\| &= O_p(p_n\sqrt{G/N}),\\
\left\|\sum_{i\in b}(M_i - \mathbb{E}M_i)\right\| &= O_p\{p_n(\sqrt{aG} + G)\},\\
\left\|\sum_{i\in b}(u_i - \mathbb{E}u_i)\right\| + \left\|\sum_{i\in b}(z_{ij} - \mathbb{E}z_{ij})\right\| &= O_p(p_n\sqrt{aG}).
\end{aligned} \tag{S83}$$

The bounds hold jointly for the finitely many reported coordinates and the $O(n^{2/5})$ blocks. For every fixed $D > 0$, each bound in (S83) holds with a constant depending on D outside an event of probability $o(n^{-D})$. The complete and every deleted fit retain all positive raw directions and pass the stated normalized eigenvalue threshold with probability $1 - o(n^{-D})$.

Proof. Separate physical times into $L = \lceil C_D \log n\rceil$ residue classes. Within a class, couple successive covariates to independent stationary draws by replacing the segment if there has been no redraw during the intervening L times. The total failure probability is at most $n\rho_0^L$. In each class the deterministic bandwidths need not agree. Put $R_n = CG(1 + T_n^2)$. For $W_i = M_i - \mathbb{E}M_i$, positivity gives

$$\|W_i\| \leq 2R_n, \qquad \mathbb{E}W_i^2 \preceq \mathbb{E}M_i^2 \preceq R_n\mathbb{E}M_i.$$

Thus a class with a_l indices has matrix variance norm at most $Ca_lG(1 + T_n^2)^2$. Theorem 6.1(ii) of Tropp (2012), applied to W_i and $-W_i$, gives a bound $C_D\{\sqrt{a_lG\log n}\,(1 + T_n^2) + G(1 + T_n^2)\log n\}$. Sum over the L classes and use $\sum_l \sqrt{a_l} \leq \sqrt{La}$. The dimension is $O(G)$ and the number of blocks is polynomial; a union bound gives the matrix bounds in (S83), with p_n enlarged. This explicitly reduces the dependent triangular array to the independent-matrix theorem.

For the centered vector $z_{ij} - \mathbb{E}z_{ij}$, the norm is at most $C\sqrt{G}(1 + T_n)$ and its squared norm has expectation at most $CG(1 + T_n^2)$. Use the rectangular version of the same Bernstein inequality, Theorem 1.6 or Remark 6.3 of Tropp (2012), on a column vector. Its row-space variance is the sum of squared Euclidean norms and is at most $Ca_lG(1 + T_n^2)$. This yields the last bound with $\sqrt{aG}$, also when $a = 1$, after increasing p_n.

Write $u_i = z_i\varepsilon_i + r_i(X_i)$, where $r_i = \mathbb{E}_U\{\phi_i(F - \phi_i^{\mathsf{T}}\eta^*) \mid X_i\}$. Conditional Cauchy–Schwarz gives $\|r_i\| \leq CE_n\sqrt{G}(1 + T_n)$ and $\mathbb{E}r_ir_i^{\mathsf{T}} \preceq E_n^2\mathbb{E}M_i$. The same vector argument applies, with $E_n \leq 1$ eventually. Conditional on the complete design, $\sum_{i\in b} z_i\varepsilon_i$ is Gaussian and its covariance trace is bounded by $C\sum_{i\in b}\|z_i\|^2 \leq CaG(1 + T_n^2)$. Its norm has the required bound by diagonalizing that covariance and using the exponential moment of independent Gaussian squares. The pooled normal equation gives $\sum_i \mathbb{E}u_i = 0$, and pooled centering gives $\sum_i \mathbb{E}z_{ij} = 0$. These identities establish the complete-sample vector rates.

The matrix rate tends to zero, yielding a fixed positive lower eigenvalue with probability $1 - o(n^{-D})$ by choosing the constants in the tail bound. Deleting $a \leq L_{\max}$ terms changes the average by a stochastic quantity of order $p_n(\sqrt{aG} + G)/N$ and a deterministic quantity of order $a(1 + T_n^2)/N = o(1)$. Thus all deleted matrices remain positive definite. In the unscaled raw powers, the minimum continuous nodal eigenvalue has order Δh^2; empirical comparison and centering preserve a fixed multiple of this order. Both $\Delta h^2 \gg N^{-3}$ and the normalized lower bound exceeds N^{-1}. The same holds with $N - a \asymp N$. The finite categorical eigenvalues are bounded below. This proves the assertion about the numerical thresholds. This completes the proof. □

Proof of (S61). Use the point state and $h = c_E n^{-1/5}$. Center local Taylor coefficients under the pooled population means and transfer their constants to the intercept, giving η^T. Its augmented

residual satisfies $|r_i^T| \le Ch_i^2$. On the stable design event of Lemma S7, the exact fit satisfies

$$\widehat{\eta} - \eta^T = \widehat{M}^{-1}(b_r + b_\varepsilon), \quad b_r = \frac{1}{N}\sum_i \mathbb{E}_U(\phi_i r_i^T \mid X_i), \quad b_\varepsilon = \frac{1}{N}\sum_i z_i\varepsilon_i.$$

For every vector v, conditional Cauchy–Schwarz bounds $|v^\mathsf{T} b_r|^2 \le (v^\mathsf{T}\widehat{M}v)N^{-1}\sum_i \mathbb{E}_U(r_i^T)^2$. Taking $v = \widehat{M}^{-1}b_r$ gives

$$b_r^\mathsf{T}\widehat{M}^{-1}b_r \le CH_4 \le Ch^4. \tag{S84}$$

The Gaussian error covariance bound and $z_i z_i^\mathsf{T} \preceq M_i$ give

$$\mathrm{Cov}(\widehat{M}^{-1}b_\varepsilon \mid X) \preceq \frac{C}{N}\widehat{M}^{-1}. \tag{S85}$$

The empirical center of group j is $\bar{z}_j = N^{-1}\sum_i z_{ij}$. In the fixed population coordinates, the reported effect is exactly $(e_j(x) - \bar{z}_j^\mathsf{T})\widehat{\eta}$, with the center row embedded in group j. The intercept row of $\widehat{M}$ and its bounded norm imply $\|\bar{z}_j\| \le C$ on this event. Equation (S70) then gives an integrated row operator norm at most C and integrated squared row norm at most C/h. Equations (S84)–(S85) imply integrated squared fit error at most $Ch^4 + C/(Nh)$ in expectation on this design event.

The empirically centered Taylor curve differs from f_j by a uniform deterministic approximation $Ch^2 + CH_2$ and the sample mean $N^{-1}\sum_i f_j(X_{ij})$. The latter has second moment at most C/N: the hold-or-redraw model gives covariance $\rho^{|i-l|}\|f_j\|_{P_j}^2$ on distinct retained indices. Pooled population centering cancels when the Taylor curve is empirically centered. Combining these bounds proves the stated oracle risk.

For N large, $\log N > B$, and clipping cannot increase squared error relative to f_j. On the remaining design or selection event, the prescribed output, clipped or zero, has loss at most $C(1 + \log^2 n)$. Lemmas S2 and S7 make its expected contribution $o(n^{-4/5})$. This comparison is unconditional and includes the dependence between selection and estimation. It establishes (S61). The observation-count conclusion was proved in Lemma S2. This completes the proof. □

The following proposition gives the categorical risk under quantitative pairwise regularity. An exponent of at least $1/2$ gives the parametric rate.

Proposition S10. *Under Condition S2 and (S60), each active categorical component satisfies*

$$\mathbb{E}\|\widehat{f}_j^{\mathrm{C,T}} - f_j\|_{P_j}^2 \le C\{n^{-1} + n^{-(4+2\zeta)/5}\}.$$

Proof. Use the point bandwidth $h = c_E n^{-1/5}$. For one categorical value, the exact report in pooled coordinates is $e_k\widehat{\eta} - \bar{z}_j^\mathsf{T}\widehat{\eta}_j$. Let u_i be as in Lemma S7, and define $\psi_{ik} = e_k\Gamma^{-1}u_i - z_{ij}^\mathsf{T}\eta_j^*$. The pooled normal equation and centering give $\sum_i \mathbb{E}\psi_{ik} = 0$. The inverse identity and the concentration bounds give, with probability $1 - o(n^{-D})$ for any fixed D,

$$\widehat{f}_k - g_k^* = \frac{1}{N}\sum_i \psi_{ik} + R_n, \qquad |R_n| \le C_D p_n G/N.$$

Here $\|e_k\| \le C$, and hence there is no evaluation factor involving $h^{-1/2}$.

Write the influence as a design term plus $(e_k\Gamma^{-1}z_i)\varepsilon_i$. Its noise sum has variance at most C/N by the Gaussian covariance bound and $N^{-1}\sum_i \mathbb{E}z_i z_i^\mathsf{T} \preceq \Gamma$. The design term is uniformly bounded: its residual part is bounded by $C\sqrt{G}(1 + T_n)E_n = o(1)$, and its centering part differs uniformly from the bounded f_j by $o(1)$. Geometric redraw covariance summation bounds the variance of its average by C/N. Its covariance with the noise term is zero. Thus $\mathbb{E}|N^{-1}\sum_i \psi_{ik}|^2 \le C/N$.

Equation (S82) gives squared population bias at most $Ch^{4+2\zeta}$. The squared remainder is at most $C_D p_n^2 G^2/N^2 = o(N^{-1})$, because $G = O(\ell_n n^{1/5})$. On the complementary concentration or selection event, clipping bounds the loss by $C(1 + \log^2 n)$, whose expected contribution is $o(N^{-1})$. Summing over the fixed number of categorical levels proves the result. This completes the proof. □

S30.6 INFLUENCE EXPANSION AND THE CORRELATED LONG-RUN COVARIANCE

Use the inference state from this point forward. Let D_n be diagonal on the fixed reported vector, with entries $\sqrt{h}$ for continuous values and one for categorical values. For a report $k = (j, x)$ define

$$\psi_{ik} = e_k \Gamma^{-1} u_i - z_{ij}^{\mathsf{T}} \eta_j^*, \qquad \psi_i^0 = \psi_i - \mathbb{E}\psi_i, \qquad \Omega_n = \operatorname{Cov}\left\{ N^{-1/2} \sum_i D_n \psi_i^0 \right\}. \tag{S86}$$

Here ψ_i collects the ψ_{ik} over the fixed reported pairs, and $\bar{\psi} = N^{-1} \sum_i \psi_i$ is its sample mean. The second term is the random empirical center. In general $\mathbb{E}\psi_i \neq 0$, because historical bandwidths differ, but $\sum_i \mathbb{E}\psi_i = 0$.

Lemma S8. *The exact selected report, except on an event of probability tending to zero, has the expansion*

$$\sqrt{N} D_n (\widehat{f}^{\mathrm{C}} - g^*) = N^{-1/2} \sum_i D_n \psi_i^0 + o_p(1). \tag{S87}$$

The centered sum has a Gaussian approximation with covariance Ω_n, and

$$0 < c \leq \min_k (\Omega_n)_{kk} \leq \max_k (\Omega_n)_{kk} \leq C. \tag{S88}$$

Neither a common limiting covariance matrix nor nonsingularity of the entire reported covariance is required.

Proof. First work with the deterministic oracle state. Its coefficient error is exactly $\delta = \widehat{\eta} - \eta^* = \widehat{M}^{-1} \bar{u}$, where $\bar{u} = N^{-1} \sum_i u_i$. Its group report is

$$\widehat{f}_k - g_k^* = e_k \delta - \bar{z}_j^{\mathsf{T}} \eta_j^* - \bar{z}_j^{\mathsf{T}} \delta_j.$$

Substitute $\widehat{M}^{-1} = \Gamma^{-1} - \Gamma^{-1}(\widehat{M} - \Gamma)\widehat{M}^{-1}$. By Lemma S7 and $\|D_{n,kk} e_k\| \leq C$, the scaled remainder is at most $O_p(p_n G/\sqrt{N}) = o_p(1)$. This proves (S87). It also proves directly why the empirical center cannot be held fixed.

Decompose $Z_i = D_n \psi_i^0$ as

$$Z_i = b_i(X_i) + v_i(X_i)\varepsilon_i, \quad \mathbb{E}b_i = 0, \quad v_{ik}(X) = D_{n,kk} e_k \Gamma^{-1} z_i(X). \tag{S89}$$

The part of b_i from the augmented projection residual is uniformly bounded by $C\sqrt{G}(1 + T_n)E_n = o(1)$. The component approximation bound (S76) shows, uniformly in i, that its center term differs by $o(1)$ from zero for continuous coordinates and from $-f_j(X_j)$ for categorical coordinates. Thus b_i is uniformly close to a bounded vector with continuous entries zero and categorical entries $-f_j(X_j)$. For any such bounded differences, covariance of functions of X_i and X_l decreases by $\rho^{|i-l|}$. Their normalized sum is $o_p(1)$.

A central limit argument for the remaining bounded categorical vector can be given directly. Partition physical time into blocks of length $\lfloor n^{1/4} \rfloor$, separated by $\lceil C \log n \rceil$ discarded times. The variance of the discarded sum divided by N tends to zero by geometric covariance summation. The remaining blocks can be coupled independently with error at most $n\rho_0^{C \log n}$. Their norms divided by $\sqrt{N}$ are bounded by $Cn^{-1/4}$, even after the deterministic selection holes. Their covariance sum is the original covariance up to $o(1)$. Expanding their characteristic functions to second order, with the remainder bounded by the maximum normalized block norm times their summed second moments, proves the multivariate Gaussian approximation. This argument also covers zero limiting variances for some linear combinations.

Conditional on the complete covariate process, the noise sum is Gaussian with covariance

$$Q_n = \frac{\sigma^2}{N} \sum_{i,l \in \mathcal{I}_n} \vartheta^{|i-l|} v_i(X_i) v_l(X_l)^{\mathsf{T}}. \tag{S90}$$

Jensen's inequality and Lemma S4 give

$$\frac{1}{N} \sum_i \mathbb{E}\|v_i\|^2 \leq C, \quad \max_i \|v_i\|_\infty^2 \leq CG(1 + T_n^2), \quad \max_i \mathbb{E}\|v_i\|^2 \leq C(1 + T_n^2).$$

For a physical lag r, products of two v values have second moment at most $CG(1+T_n^2)^2$. Two lag segments separated by a times have covariance bounded by this second-moment bound times ρ_0^a, by conditioning on a redraw between them. Overlapping segments give at most $2r+1$ terms per index. Thus the variance of their average is at most $CG(1+T_n^2)^2(r+1)/N$. Truncate (S90) at $r = C\log n$; its tails have expectation at most $C(1+T_n^2)\vartheta_0^{C\log n}$. Summing the lag bounds gives

$$Q_n - \mathbb{E}Q_n = O_p(p_n\sqrt{G/N}) = o_p(1).$$

The conditional Gaussian characteristic function and the bounded design central limit calculation now give a Gaussian approximation for $N^{-1/2}\sum_i Z_i$. Indeed, replacing the conditional covariance by its deterministic expectation in that characteristic function has error tending to zero in L^1. This proves asymptotic independence of the two limiting contributions without asserting their finite-sample independence.

The upper diagonal bound in (S88) follows from the averaged L^2 row bound and the upper Gaussian covariance bound, plus the bounded design term. For a continuous interior value, the direct uncentered row is $K_{ij}(x,\cdot)$ from (S72). For $i \geq n/2$,

$$\int K_{ij}(x,z)\,dP_j(z) = 1 + O(h). \tag{S91}$$

Here is the required nodal calculation. For an interior node on an infinite equally spaced grid, the kernel denominator is periodic with period Δ. Partitioning the real line into grid cells shows $\int \pi_{i,jg} = \Delta$, and reflection around u_g gives $\int (z-u_g)\pi_{i,jg}(z)\,dz = 0$. These identities are exact on the finite grid for nodes at distance greater than $2h_{\max}$ from the boundary. A second-order expansion of p_j shows that, for $h_i = O(h)$, the nodal entries of $c_{ij} = \mathbb{E}a_{ij}(X_j)$ differ from those of $\bar{c}_j$ by $O(\Delta h^2)$ in the intercept and $O(\Delta h)$ in the scaled slope. Because $\bar{c}_j = \bar{D}_j d_j$, the inverse nodal bound and the bounded sum of reporting raw block norms prove (S91).

The support of this row has probability at most Ch, and hence

$$\int K_{ij}(x,z)^2\,dP_j(z) \geq \frac{\{\int K_{ij}(x,z)\,dP_j(z)\}^2}{P_j\{|X_j - x| \leq Ch\}} \geq c/h.$$

A fixed fraction of the retained indices satisfy $i \geq n/2$. Subtracting one to center the row changes its averaged squared norm by at most order $h^{-1/2}$ plus a constant. The resolvent correction in Lemma S5 has bounded averaged squared norm. After multiplying by h, these corrections cannot cancel the positive lower bound. Thus $N^{-1}\sum_i \mathbb{E}v_{ik}^2 \geq c$ for a continuous report.

For a categorical row, $\|e_k\|^2 = P_j\{x\}^{-1} - 1 \geq c$. Its basis does not contain auxiliary randomness. With $a = \Gamma^{-1}e_k^{\mathsf{T}}$, the group-$j$ block equation is

$$\frac{1}{N}\sum_i \mathbb{E}\{z_{ij}(z_i^{\mathsf{T}}a)\} = e_{k,j}^{\mathsf{T}}, \qquad \frac{1}{N}\sum_i \mathbb{E}z_{ij}z_{ij}^{\mathsf{T}} = I.$$

Cauchy–Schwarz gives $N^{-1}\sum_i \mathbb{E}(z_i^{\mathsf{T}}a)^2 \geq \|e_k\|^2$. The lower Gaussian autoregressive covariance bound then proves a positive diagonal noise variance for every report. The design variance is nonnegative and has zero covariance with the noise part. This proves (S88).

Along any subsequence there is a further subsequence on which the bounded covariance matrices converge. The preceding characteristic-function argument applies on each further subsequence, proving the stated Gaussian approximation without requiring a unique limiting matrix. A finite maximum of standardized Gaussian coordinates has no atoms: an interval of width δ for the maximum is contained in the union of at most twice the number of reported values of intervals of width δ for unit-variance Gaussian coordinates. This also gives uniform anti-concentration when the joint covariance is singular. Selection transfer follows from Lemma S2. This completes the proof. □

For clarity, the covariance in this theorem can be written entirely in terms of deterministic population functions. For $r = |i-l|$, (S55) and (S89) give

$$\Omega_n = \frac{1}{N}\sum_{i,l\in\mathcal{I}_n}\Big[\rho^r\mathbb{E}_Q\{b_i(X)b_l(X)^{\mathsf{T}}\} + \sigma^2\vartheta^r\big\{\rho^r\mathbb{E}_Q[v_i(X)v_l(X)^{\mathsf{T}}] + (1-\rho^r)(\mathbb{E}_Q v_i)(\mathbb{E}_Q v_l)^{\mathsf{T}}\big\}\Big]. \tag{S92}$$

The convention is $\rho^0 = \vartheta^0 = 1$, including at zero. The sum includes both orientations of each pair. Coordinate dependence is retained in the Q expectations, and physical-time gaps include the omitted selection indices. In particular, this is not a stationary single-bandwidth covariance formula substituted into a pooled fit.

S30.7 Exact deletion, unequal blocks and conditional resampling

The following calculation accounts for recomputation of the empirical centers and the normalization after deleting a block. It holds in a fixed population coordinate system, which is a common coordinate representation of every fit; it does not keep any sample center or normalization fixed in the procedure.

Lemma S9. *For $V_b = (N - n_b)(\widehat{f}^{\mathrm{C}} - \widehat{f}^{\mathrm{C}}_{-b})$ and $\widetilde{V}_b = V_b - (n_b/N)\sum_a V_a$, put $S_b = \sum_{i\in b} D_n \psi_i^0$ and $S = \sum_b S_b$. Then*

$$\max_b \frac{\|D_n \widetilde{V}_b - (S_b - n_b S/N)\|}{\sqrt{n_b}} = o_p(1). \tag{S93}$$

The result includes an incomplete block of any positive size.

Proof. Set $A = \widehat{M}$, $A_{-b} = (NA - M_b)/(N - a)$, $a = n_b$, $M_b = \sum_{i\in b} M_i$, $U_b = \sum_{i\in b} u_i$, and $Z_{b,j} = \sum_{i\in b} z_{ij}$. The coefficient error is $\delta = A^{-1}\bar{u}$. Subtracting the two exact normal equations gives

$$(N - a)(\widehat{\eta} - \widehat{\eta}_{-b}) = A_{-b}^{-1}(U_b - M_b\delta). \tag{S94}$$

We write $\bar{z}_{j,-b} = (N - a)^{-1}\sum_{i\notin b} z_{ij}$ for the deleted empirical mean. For a report in group j, use $(N - a)(\bar{z}_j - \bar{z}_{j,-b}) = Z_{b,j} - a\bar{z}_j$. Its exact deleted report is

$$V_{b,k} = (e_k - \bar{z}_{j,-b}^{\mathsf{T}})A_{-b}^{-1}(U_b - M_b\delta) - (Z_{b,j} - a\bar{z}_j)^{\mathsf{T}}\widehat{\eta}_j. \tag{S95}$$

Center rows in this equation are embedded in group j. This formula includes the changes that would be lost if the center were held fixed.

Put $Q_b = U_b - a\bar{u}$ and $H_b = M_b - aA$. Then $U_b - M_b\delta = Q_b - H_b\delta$. The first-order term in (S95) is

$$e_k\Gamma^{-1}Q_b - (Z_{b,j} - a\bar{z}_j)^{\mathsf{T}}\eta_j^* = \sum_{i\in b}(\psi_{ik} - \bar{\psi}_k).$$

Its error is exactly the sum of

$$[e_k(A_{-b}^{-1} - \Gamma^{-1}) - \bar{z}_{j,-b}^{\mathsf{T}}A_{-b}^{-1}]Q_b,$$
$$-(e_k - \bar{z}_{j,-b}^{\mathsf{T}})A_{-b}^{-1}H_b\delta, \qquad -(Z_{b,j} - a\bar{z}_j)^{\mathsf{T}}\delta_j.$$

We bound all three terms, including their deterministic block means. Lemmas S4 and S6 give

$$\|\mathbb{E}u_i\| \le C(1 + T_n)E_n, \qquad \|\mathbb{E}M_b - a\Gamma\| \le Ca(1 + T_n^2), \qquad \|\mathbb{E}z_{ij}\| \le C(1 + T_n).$$

Using Lemma S7, $G = O(\ell_n n^{1/4})$ and $L_{\max} = O(n^{3/5})$, uniformly in b we obtain

$$\begin{aligned}
\|\delta\| &= O_p(p_n\sqrt{G/N}),\\
\|A_{-b} - \Gamma\| + \|\bar{z}_{j,-b}\| &= O_p(p_n\sqrt{G/N}),\\
\|Q_b\| &= O_p(p_n\sqrt{aG}),\\
\|H_b\| &= O_p\{p_n(\sqrt{aG} + G + a)\},\\
\|Z_{b,j} - a\bar{z}_j\| &= O_p\{p_n(\sqrt{aG} + a)\}.
\end{aligned}$$

For the second line, the deterministic deletion term $L_{\max}(1 + T_n^2)/N$ is bounded by $p_n\sqrt{G/N}$. For the third, the deterministic contribution $a(1 + T_n)E_n$ is bounded by $p_n\sqrt{aG}$, because $\sqrt{L_{\max}}h^2$ times an iterated-logarithm factor tends to zero. The same estimates cover $a = 1$.

Because $\|D_{n,kk}e_k\| \le C$, division of the three exact error terms by $\sqrt{a}$ gives the joint bound

$$O_p\left[p_n\left\{\frac{G}{\sqrt{N}} + \frac{G^{3/2}}{\sqrt{Na}} + \sqrt{\frac{aG}{N}}\right\}\right]. \tag{S96}$$

To replace ψ_i by ψ_i^0, note that

$$\max_i \|D_n \mathbb{E}\psi_i\| \le C(1+T_n)E_n \le \sqrt{G}h^2 p_n.$$

For the center term in this inequality, each projected component differs from f_j, whose population mean is zero, by at most the bound in (S76). The deterministic block means divided by $\sqrt{a}$ are at most $\sqrt{L_{\max} G}\, h^2 p_n = o(1)$. In (S96), the worst powers of n are $-1/4$, $-1/8$ and $-3/40$ for $a \ge 1$; the mean bound has power $-3/40$. Thus all errors tend to zero uniformly.

Finally, the first-order terms $S_b - aS/N$ sum to zero. If the remainder in block b is at most $\epsilon_n\sqrt{n_b}$, where $\epsilon_n = o_p(1)$, then its size-weighted recentering adds at most

$$\epsilon_n \frac{\sqrt{n_b}}{N} \sum_a \sqrt{n_a} \le \epsilon_n \sqrt{J n_b / N} = O(\epsilon_n).$$

The last inequality uses $JL_{\max}/N = O(1)$. This proves (S93) for the actual $\widetilde{V}_b$. This completes the proof. □

Lemma S10. *For the actual covariance*

$$\widehat{\Sigma} = \frac{J}{(J-1)N^2} \sum_b \widetilde{V}_b \widetilde{V}_b^{\mathsf{T}},$$

we have

$$N D_n \widehat{\Sigma} D_n - \Omega_n \to_p 0. \tag{S97}$$

Conditional independent uniform resampling of J vectors from $\{\widetilde{V}_1, \dots, \widetilde{V}_J\}$, with the resampled standard deviation in each coordinate, gives the same limiting studentized maximum as the original report.

Proof. For a consecutive retained segment of size a, write its standardized influence sum as $S_I = \sum_{i\in I} Z_i$, with the decomposition in (S89). Its moments obey

$$\mathbb{E}\|S_I\|^2 \le C p_n a, \qquad \mathbb{E}\|S_I\|^4 \le C p_n (a^2 + aG). \tag{S98}$$

Here and below the reported vector has fixed dimension. To see these bounds explicitly, the centered b_i are uniformly bounded. For ordered times $i_1 \le i_2 \le i_3 \le i_4$, conditioning on redraws in the first and last gaps bounds a fourth product by $C\rho_0^{\max(i_2-i_1, i_4-i_3)}$. Summing the two exterior gaps and the possible middle gap yields Ca^2, also on a thinned sequence. For the noise sum, conditional Gaussian fourth moments and the upper error covariance bound give

$$\mathbb{E}\left|\sum_{i\in I} v_{ik}\varepsilon_i\right|^4 \le C\mathbb{E}\left(\sum_{i\in I} v_{ik}^2\right)^2 \le C p_n (a^2 + aG).$$

For the second inequality, use $\max_i \mathbb{E}v_{ik}^2 \le C p_n$, $\max_i \|v_{ik}\|_\infty^2 \le C p_n G$, and the exact redraw contraction for the covariance of v_{ik}^2 and v_{lk}^2. Second moments follow from the same calculation. The fourth moment of their sum is bounded by eight times the sum of fourth moments. This proves (S98).

Equation (S92) shows that the norm of a covariance at physical lag r is at most $C p_n (\rho_0 \vee \vartheta_0)^r$. Across one block boundary, at most r pairs have lag r. The sum of all omitted cross-block covariances divided by N is thus at most $C p_n J/N = o(1)$.

We justify the variance calculation for block outer products without assuming independent blocks. Remove $g = \lceil C \log n \rceil$ physical times from the beginning of each complete block. The total discarded count is at most Jg. By (S98), the expected sum of squared discarded block sums is at most $C p_n J g$. Cauchy–Schwarz then shows that their removal changes $N^{-1}\sum_b S_b S_b^{\mathsf{T}}$ by $o_p(1)$, because $p_n Jg/N \to 0$. All late complete blocks are much longer than g. A shorter incomplete block can be omitted entirely: its expected squared sum divided by N is at most $C p_n L_{\max}/N = o(1)$.

For each remaining complete block, an independent redraw in its removed prefix makes its retained design segment independent of earlier blocks, except with probability at most ρ_0^g. For the errors use their innovation representation $\varepsilon_i = \sum_{r\ge 0} \vartheta^r \zeta_{i-r}$, where ζ_i are independent $N(0, \sigma^2(1-\vartheta^2))$.

Start an independent stationary error at the beginning of the removed prefix of each block and use the same innovations within the block. The L^2 difference at the first retained time is at most $C\vartheta_0^g$, and decreases geometrically thereafter. Multiplication by $\|v_i\|_\infty \le Cp_n\sqrt{G}$ keeps the total perturbation negligible for C large. Thus the retained block sums can be replaced by independent, nonidentically distributed sums S_b°, with negligible effect on the covariance. This is an approximation only in the proof; no observations are removed by the algorithm.

For any covariance entry, independence and (S98) give

$$\mathbb{E}\left|\frac{1}{N}\sum_b\{S_{b,k}^\circ S_{b,l}^\circ - \mathbb{E}S_{b,k}^\circ S_{b,l}^\circ\}\right|^2 \le \frac{Cp_n}{N^2}\sum_b(n_b^2+n_bG) \le Cp_n\left(\frac{L_{\max}}{N}+\frac{G}{N}\right)\to 0.$$

Restore the removed times and the partial block. The cross-block covariance bound proves $N^{-1}\sum_b S_bS_b^{\mathsf{T}} - \Omega_n \to_p 0$. Also $S/\sqrt{N} = O_p(1)$ by Lemma S8, and $\sum_b(n_b/N)^2 \le L_{\max}/N = o(1)$. Cauchy–Schwarz shows that replacing S_b by $S_b - n_bS/N$ does not affect the limit. Lemma S9 then proves (S97); its remainders have summed squares $o_p(N)$, and $J/(J-1)\to 1$.

For the resampling calculation let $A_b = \sqrt{J/N}\,D_n\widetilde{V}_b$. Their empirical mean is exactly zero and their empirical covariance converges to Ω_n. Also,

$$\frac{1}{J^2}\sum_b\|A_b\|^4 = \frac{1}{N^2}\sum_b\|D_n\widetilde{V}_b\|^4 = o_p(1). \tag{S99}$$

This follows from (S98), the bound $p_n(L_{\max}/N+G/N)\to 0$, and the uniform deletion remainder: if its size is $\epsilon_n\sqrt{n_b}$, its fourth-power sum divided by N^2 is at most $\epsilon_n^4 L_{\max}/N$. The size-weighted term n_bS/N has fourth-power sum divided by N^2 bounded by $\|S/\sqrt{N}\|^4(L_{\max}/N)^3 = o_p(1)$, which also justifies the recentering in this fourth-moment calculation. Equation (S99) is the conditional Lyapunov bound for J independent draws from the empirical distribution of the A_b. It proves the conditional central limit approximation. It also bounds the conditional variance of each resampled empirical second moment, which makes each resampled standard deviation consistent. The positive coordinate variances in (S88) permit studentization.

Apply these arguments along covariance-convergent further subsequences as in Lemma S8. Gaussian anti-concentration for the fixed maximum then transfers conditional quantiles. The exact conditional resampling quantile has the claimed limit; no fixed finite number of Monte Carlo resamples is asserted to give an exact quantile. This proves the lemma. This completes the proof. □

Completion of Theorem S7. The count and continuous-risk claims were proved above. Proposition S10 gives the categorical rate under its additional Hölder condition. The count, continuous-risk and storage proofs require no pairwise density condition or fixed active distribution: their population matrix, Taylor residual, concentration and truncation bounds use only the marginal bounds, additive coercivity and selection premises. Lemma S6 makes the population bias negligible in (S87). Lemmas S8 and S10 prove (S62) and the stated standard-error orders for the oracle state. The probability of a selected or unavailable state differing from that state tends to zero by Lemma S2 and the eigenvalue bounds; it bounds the difference between the two coverage probabilities.

The certificate addition stores a constant number of coefficient and residual vectors in the current selection quadratic and no additional observations. The selection records use $O(dn^{3/4})$ scalars and the open block uses $O(dn^{3/5})$. A grid activated by time n has first reporting time at most $Cn\log\log n$. Thus the largest active point and inference grids have orders $\ell_{n\log\log n}(n\log\log n)^{1/5}$ and $\ell_{n\log\log n}(n\log\log n)^{1/4}$. The sums of squared dimensions over nested grids are bounded by a constant times the largest squared dimension, even if older grids are retained. There is one point total and at most $O(n^{2/5})$ inference block states per grid, each of order s^2G^2 after support stabilization. The resulting counts are exactly (S63). The temporary selection quadratic has order $m_k^2G_L^2 \le Cdk$ for the prescribed candidate rule $m_k \le d\wedge Ck^{1/5}$ and $G_L \asymp k^{1/5}$: indeed $m_k^2G_L^2 \le Cdk^{3/5} \le Cdk$. The extra vectors are smaller. Thus it is covered by the selection-record term. At most $O(\log n)$ distinct nested grid levels have been activated by time n if no obsolete level is deleted; the geometric sum, not an assumption of a constant number of levels, proves the scalar bound. Completed estimation records are never revisited. This completes the proof. □

S30.8 NONPRODUCT EXAMPLES WITH VERIFIED SELECTION PREMISES

The first example has two curved continuous effects and an active correlated categorical effect. It also contains an inactive variable, making support selection part of the example.

Proposition S11. *Let X_1, X_2 have uniform marginals on $[0,1]$, let Z have probabilities $1/2$ on $\{-1,1\}$, and let X_4 be an independent uniform null variable. Relative to the product of these marginals, give (X_1, X_2, Z) density*

$$q(x_1,x_2,z) = 1 + au(x_1)u(x_2) + bu(x_1)z + cu(x_2)z, \qquad u(x) = \sqrt{3}(2x-1), \tag{S100}$$

where $|a| \le 1/20$, $|b|, |c| \le 1/10$, and at least one of a, b, c is nonzero. Put

$$\begin{aligned} f_1(x) &= b_0 u(x) + \eta \frac{\sqrt{5}}{2}\{3(2x-1)^2 - 1\}, \\ f_2(x) &= b_0 u(x) + \eta \frac{\sqrt{7}}{2}\{5(2x-1)^3 - 3(2x-1)\}, \\ f_3(z) &= b_0 z, \qquad f_4 = 0, \qquad b_0 = B/10, \quad \eta = b_0/10. \end{aligned} \tag{S101}$$

For $h_L \asymp k^{-1/5}$, $G_L \asymp k^{1/5}$, $m_k = d = 4$ and $\lambda_k = c_\lambda k^{-1/4}$, with $c_\lambda > 0$ fixed, every premise of Theorem S7 holds for sufficiently large k. Any fixed SCAD parameter $a_{\mathrm{SCAD}} > 2$ is allowed. The original fixed-sweep solver also attains its tolerance on this class with probability $1 - o(k^{-A})$ for every fixed $A > 0$.

Proof. The perturbation in (S100) is bounded in absolute value by $3/20 + \sqrt{3}/5 < 1/2$. Thus $q > 1/2$ and $q < 3/2$. Integrating any centered factor gives the stated marginals. Its pairwise ratios are $1 + au_1u_2$, $1 + bu_1z$ and $1 + cu_2z$, all with bounded second derivatives. The normalized dependence matrix on (u_1, u_2, z) has unit diagonal and off-diagonal entries a, b, c; its minimum eigenvalue is at least $4/5$ by the row-sum bound. Decompose each centered effect into its projection on u_j or z and its orthogonal residual. Only the three projections have nonzero cross-covariances. This proves (S56) with $\kappa = 4/5$.

The stated polynomials are centered. The crude bounds $\sqrt{3} < 2$, $\sqrt{5} < 3$ and $\sqrt{7} < 3$ give $\sum_j \|f_j\|_\infty^2 \le (2 \cdot 2.3^2 + 1)b_0^2 < B^2$. Their derivatives are bounded by fixed multiples of B. Choose any permitted ρ, ϑ, σ in (S55).

At selection bandwidth, an affine function is represented exactly by its nodal lines. The separate local projection residuals of f_1, f_2 are orthogonal to 1 and u in the augmented marginal space. All pairwise interactions are through u and z, and hence every cross score of these residuals is zero. The full population projection is the stack of the separate projections. Its residual is uniformly $O(h_L^2)$. The normalized coefficient of each active group has norm at least b_0, by taking its inner product with the unit coefficient representing u or z. The null group is independent, and its coefficients are zero.

The normalized population Gram has minimum eigenvalue at least $4/5$ and off-diagonal block row sums at most $1/5$. Its initial Lasso solution has zero null block and satisfies

$$\|\theta^L - \theta^*\| \le \frac{\sqrt{3}\lambda_k}{4/5}.$$

Thus all three active norms exceed $a_{\mathrm{SCAD}}\lambda_k$ eventually, giving zero active SCAD weights, while the null weight is λ_k. The restricted weighted solution is $\bar\theta = \theta^*$; its inactive score is zero. The margins in (S57) are explicit. All coefficient norms are bounded, and the normalized feature bound $C\sqrt{G_L}$ and pair-moment bound CG_L follow from the nodal moment bounds and conditional Jensen. There is no screening boundary because $m_k = d$. Finally, $4r_k = O(k^{-2/5}\sqrt{\log k} + k^{-4/5}\log^2 k) = o(k^{-1/4})$. This verifies all selection restrictions uniformly over the retained range.

On the same concentration event, sample off-diagonal block row sums are below $1/4$ for all sufficiently large k. Group soft thresholding is nonexpansive, and each cyclic sweep contracts the maximum group error by at most this row sum: already updated errors have the contracted bound and nonupdated errors have the preceding bound. Thus $100 + \lceil 30 \log k \rceil$ sweeps give an error smaller than k^{-3}, and the certificate margin argument gives the stated support. The certificate iteration is not needed on these samples. The variance lower bounds were proved in Lemma S8, including for the correlated categorical component. This completes the proof. □

The next example has infinitely many nonzero pairwise dependence coefficients.

Proposition S12. *We define* $\varphi_r(x) = \sqrt{2}\sin(2\pi r x)$. *We replace* (S100) *by*

$$q(x_1, x_2, z) = 1 + \epsilon \sum_{r=1}^{\infty} r^{-6} \varphi_r(x_1)\varphi_r(x_2) + bu(x_1)z, \quad 0 < \epsilon \le 1/16, \quad 0 < |b| \le 1/20. \tag{S102}$$

We keep the independent null X_4 *and choose* $f_1 = b_0 u$, $f_2 = b_0 u$, $f_3 = b_0 z$, $f_4 = 0$, *with* $b_0 = B/10$. *This distribution has infinite-rank pairwise dependence and satisfies every premise and conclusion of Theorem S7 with the selection choices in Proposition S11. The original fixed-sweep solver is again covered.*

Proof. Because $\sum_{r\ge1} r^{-6} \le 1 + \int_1^\infty t^{-6} dt = 6/5$ and $\sqrt{3} < 2$,

$$|q - 1| \le 2\epsilon \sum_{r\ge1} r^{-6} + \sqrt{3}|b| < 3/20 + 1/10 = 1/4.$$

Thus $3/4 < q < 5/4$. All marginals are the specified marginals. The series and its first two derivatives in either argument converge uniformly, because their absolute bounds are constant multiples of $\sum r^{-6}$, $\sum r^{-5}$ and $\sum r^{-4}$. The density satisfies the pairwise bounds, and its product-density lower bound implies (S56) with $\kappa = 3/4$. The pair (X_1, X_2) has correlation operator eigenvalues ϵr^{-6} on the infinitely many orthonormal functions φ_r, proving infinite rank.

Every nonzero effect is exactly represented by the augmented local lines at every bandwidth, and hence the full population residual is zero. All active normalized coefficient norms equal b_0. The independent null group is zero in the population initial Lasso solution. Strong convexity gives $\|\theta^L - \theta^*\| \le \sqrt{3}\lambda_k/(3/4)$. The same zero-active-weight, positive-null-weight calculation as above verifies every selection margin and bounded coefficient condition. The normalized feature and pair-moment bounds again follow from the marginals alone. The total signal bound is immediate from $7b_0^2 < B^2$.

At a single selection bandwidth the normalized additive averaging maps are contractions in marginal L^2. The cross-operator norm between groups 1 and 2 is at most ϵ, and that between group 1 and Z is at most $|b|$; all other cross blocks are zero. Thus their maximum off-diagonal block row sum is at most $1/16 + 1/20 < 1/4$. Sample concentration preserves a bound strictly below $1/4$ eventually. The preceding sweep argument proves the final assertion. This completes the proof. □

Proposition S13. *We assume* X_1, X_2, X_4 *have uniform marginals on* $[0, 1]$, *and we assume* Z *has probability* $1/2$ *at each of* $-1, 1$. *Relative to their product distribution, we use the density*

$$q(x_1, x_2, z, x_4) = 1 + \tau \operatorname{sign}(x_1 - 1/2)z, \qquad 0 < \tau \le 1/10. \tag{S103}$$

We use fixed nonzero centered C^2 *effects for* X_1, X_2, *a fixed nonzero effect* bz *for* Z, *and a zero effect for* X_4, *with total signal within the stated bound. Under the second selection specification in Condition S2(c), all conclusions of Theorem S7 hold. The categorical point risk is* $O(n^{-1})$, *although its pairwise density ratio is discontinuous.*

Proof. The density is between $9/10$ and $11/10$. Integrating the centered step or z leaves the stated marginals. The product lower bound gives additive coercivity with $\kappa = 9/10$. The only nonzero conditional-expectation operators connect X_1 and Z, and their norms are τ. Thus (S58) holds. Fixed nonzero effects give the signal lower bound. The selection lemma verifies the original penalty form with a sufficiently large fixed multiplier and its sweep cap. The remaining inference conditions follow directly from the bounded fixed density.

For the categorical contrast z, we define $s(x) = \operatorname{sign}(x - 1/2)$. Solving the dual equations on the span of $s(X_1)$ and Z gives $G_Z = (Z - \tau s(X_1))/(1 - \tau^2)$. The two level reports are fixed multiples of this contrast. Common nodal constants represent s exactly except when a node and its observed value are on opposite sides of $1/2$. Such a value lies within h_i of the jump, an event of probability at most $2h_i$. The squared augmented approximation error of the dual is at most CH_1. Equation (S80) bounds squared categorical bias by CH_1H_4. At point bandwidth, $H_1H_4 \le Ch^5 = O(n^{-1})$. The variance and remainder bounds in Proposition S10 do not use Hölder continuity and remain $O(n^{-1})$. At inference bandwidth, $\sqrt{nH_1H_4} \le Cn^{-1/8}\sqrt{B_n} \to 0$. This proves the categorical risk and confirms the inference conclusion directly. This completes the proof. □

S30.9 Scope and relation to the independent proof

Coordinate independence was used in four distinct places in the earlier continuing argument: its selection corollaries, the identity reference additive Gram matrix, the small off-diagonal pooled correction, and exact zero categorical population bias. Lemmas S2, S4, S5 and S6, respectively, replace those steps. Temporal redraw and independence of the entire error process from the design remain unchanged. The independent-coordinate theorems retain their original selection alternatives. The first specification in part (c) can be stronger at a fixed penalty multiplier; the second specification supplies a direct fixed-multiplier result under weak correlation.

The finite-rank pairwise class of (S10) satisfies the new active-design assumptions whenever its dependence functions have two bounded derivatives and its additive correlation matrix is bounded below. Decomposing each centered function into its finite dependence span and orthogonal complement proves coercivity, and bounded smooth functions give the required pairwise ratios. Historical pooling, however, averages products of bandwidth-specific projection operators. It is generally false that this average equals the product of the two averaged operators. The proof above uses the actual pooled augmented criterion and does not make this substitution.

The inference extension still requires bounded pairwise density ratios, smooth marginal densities and the specified temporal model. Fixed-distribution categorical inference uses an L^2 approximation of its dual. Uniform inference over changing distributions follows under the common Hölder bound in (S60). Continuous-effect risk does not require a pairwise density. Neither argument covers general temporal mixing, non-Gaussian errors, growing active sets, data-dependent reporting times or growing collections of reported values. The unmodified sweep cap is covered under the second selection specification; the first specification uses the certificate when needed.

References for the Appendix

Candanedo, L. (2017). *Appliances Energy Prediction*. UCI Machine Learning Repository. https://doi.org/10.24432/C5VC8G.

Candanedo, L. M., Feldheim, V. and Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings* **140** 81–97.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields* **162** 47–70.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability* **45** 2309–2352.

Cohen, A., Davenport, M. A. and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics* **13** 819–834.

Dueck, J., Edelmann, D. and Richards, D. (2017). Distance correlation coefficients for Lancaster distributions. *Journal of Multivariate Analysis* **154** 19–39.

Fanaee-T, H. (2013). *Bike Sharing*. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.

Fanaee-T, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2** 113–127.

Guo, Z. and Zhang, C.-H. (2022). Extreme eigenvalues of nonlinear correlation matrices with applications to additive models. *Stochastic Processes and their Applications* **150** 1037–1058.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer.

Mammen, E. and Park, B. U. (2006). A simple smooth backfitting method for additive models. *Annals of Statistics* **34** 2252–2271.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* **12** 389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.