arXiv:2304.07855v2 [econ.EM] 15 Nov 2026

Selective Inference for Survey-Weighted Generalized Linear Models: Growing Dimension and Average Marginal Effects*

Joann Jasiak† Purevdorj Tuvaandorj‡

November 17, 2026

Abstract

This paper extends the ℓ_1-penalized likelihood Selective Inference method to survey settings. We establish its asymptotic validity for generalized linear models with an increasing number of covariates and survey weights, allowing inference on pseudo-true parameters of possibly misspecified selected models. Moreover, we generalize the method to facilitate inference on nonlinear functionals, such as the average marginal effect in survey-weighted logit models. The effectiveness of the proposed approach is illustrated using simulated data and the 2020 Canadian Internet Use Survey.

Keywords: Average Marginal Effect, Growing dimension, ℓ_1-regularization, Logit, Selective Inference, Survey data, Yurinskii coupling

1 Introduction

Survey data are widely used across many disciplines in the social sciences. The statistical methodology for analyzing survey samples has been extensively developed, resulting in a

*The authors thank Denis Chetverikov, the co-editor, and two anonymous referees for their constructive comments and suggestions, which substantially improved the paper; Lucas Van Den Driesschen for research assistance; and participants at the 2024 Canadian Economics Association (CEA) Conference for helpful discussions. The authors acknowledge financial support of the Catalyzing Interdisciplinary Research Clusters Initiative (CIRC), York University. This research is part of the projects "Data Economy" and "Digital Currencies," approved and funded by the office of VPRI and participating faculties of York University. The authors also acknowledge access to the data provided by RDC and Statistics Canada (Project 21-MAPA YRK-721).

†York University, jasiakj@yorku.ca

‡York University, tpujee@yorku.ca

large body of literature (see, e.g., Cameron and Trivedi, 2009; Wooldridge, 2010; Fuller, 2011; Thompson, 2012).

Despite the rapid development of machine learning and high-dimensional econometric methods, together with the increasing availability of large datasets, relatively little work has addressed how these methods can be adapted to survey data. This paper contributes to filling this gap by extending the ℓ_1-penalized likelihood Selective Inference (SI) method to a survey setting. Our goal is to expand the toolbox available to applied researchers who wish to employ high-dimensional regression techniques with survey data.

Penalized prediction-oriented methods, including the Lasso, involve a bias-variance trade-off and produce estimates that are not directly suitable for conventional inference on model coefficients. Several methods addressing the inferential difficulties associated with regularization and model selection have been proposed in the literature. Among these, Zhang and Zhang (2014) and Javanmard and Montanari (2014) propose debiased Lasso procedures based on a one-step correction of the initial Lasso estimator. Belloni et al. (2014) and Belloni et al. (2016) develop double-selection and $C(\alpha)$-type procedures for generalized linear models and partially linear models under sparsity assumptions. A related recent contribution is Chetverikov and Sørensen (2025), who propose a tuning parameter selection method termed bootstrapping after cross-validation (BCV). Another post-selection inference approach is the PoSI method proposed by Berk et al. (2013). This approach formulates post-selection inference as a simultaneous inference problem over a prespecified family of models.

An inferential procedure conceptually distinct from these approaches is SI, which is the focus of this paper. The SI framework for linear models selected by the Lasso was developed by Lee et al. (2016) and subsequently extended by Taylor and Tibshirani (2018) to generalized linear models (GLMs) estimated by ℓ_1-penalized maximum likelihood (PML). In SI, inference on the parameters of the selected model is conducted conditional on the selection event. This distinguishes SI from PoSI, which constructs simultaneous confidence intervals across a prespecified family of models without conditioning on the particular model selected.

In applications of this kind, the ℓ_1-PML serves as a model-selection device rather than as an end in itself. Once the data have been used to choose a specification, the questions of interest are which selected associations remain distinguishable from zero after accounting for selection, how large they are, and how precisely they are estimated. Conventional Wald intervals computed from the selected model ignore the selection step and may not attain their nominal coverage. SI answers these questions for the population parameters of the selected model, which we define as pseudo-true values so that they remain well defined when

the selected model omits relevant regressors. These targets differ from those of the debiased Lasso, double-selection, and other Neyman-orthogonal procedures such as the $C(\alpha)$ test, which deliver inference on a low-dimensional parameter of the full model specified before selection. As with any regression-based summary, the SI targets are associational; a causal interpretation requires additional identifying assumptions.

This paper provides two main results. First, we extend the SI framework to survey-weighted GLMs estimated via ℓ_1-PML and establish its asymptotic validity when the number of covariates grows with the sample size.[1] Our survey framework follows Wooldridge (2001). With survey weights, the usual information matrix equality fails to hold, and the statistics associated with the active and inactive parts of the Karush–Kuhn–Tucker (KKT) conditions need not be asymptotically independent. Consequently, the simplification used by Taylor and Tibshirani (2018), which relies on this independence, is not generally applicable. We therefore conduct inference conditional on the full selection event, characterized jointly by the active and inactive KKT conditions. Second, we establish the asymptotic validity of the proposed procedure for inference on nonlinear functionals, including average marginal effects (AMEs) in survey-weighted logit models.

Several recent papers study penalized estimation and bias-corrected inference for related quantities. Wooldridge and Zhu (2020) develop inference on average partial effects (APEs) in approximately sparse correlated random effects probit models with panel data; Hirshberg and Wager (2020) propose debiased inference on APEs in single-index models; and Chiang (2022) studies inference on many APEs with an application to text regression. Unlike these approaches, we develop SI for AMEs in survey-weighted GLMs.

There are relatively few studies applying Lasso-type methods in survey settings. For example, McConville et al. (2017) study model-assisted survey regression estimation using the Lasso and establish finite-population asymptotic results with a fixed number of regressors. In contrast, we consider survey-weighted GLMs and establish the asymptotic validity of the SI procedure for pseudo-true parameters of selected models, without requiring sure screening, under the standard infinite-population framework commonly used in econometrics (see, e.g., Wooldridge (2001, 2010); Cameron and Trivedi (2009)).

The remainder of the paper is organized as follows. Section 2 introduces the model framework. Section 3 develops the selective inference procedure for survey-weighted GLMs. Section 4 presents simulation evidence on the properties of the proposed method, and Section

[1] We thank the co-editor for suggesting this extension using Yurinskii's coupling.

5 provides an empirical application using the 2020 Canadian Internet Use Survey. Section 6 concludes.

Notations and terminology Let $1(\cdot)$ denote the indicator function, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of a symmetric matrix A, respectively. For a $k \times 1$ vector $a = (a_1, \ldots, a_k)'$, define $\text{supp}(a) \equiv \{j \in \{1, \ldots, k\} : a_j \neq 0\}$, $\|a\|_0 \equiv |\text{supp}(a)|$, $\|a\|_1 \equiv \sum_{j=1}^k |a_j|$. For a real matrix $A = (a_{ij})$, let $\|A\|_\infty \equiv \max_{i,j} |a_{ij}|$ denote its entrywise maximum norm, and let $\|A\| = \sqrt{\text{tr}(A'A)}$ and $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$ denote its Frobenius and spectral norms, respectively. The sub-Gaussian norm of a random variable X is defined as $\|X\|_{\psi_2} \equiv \sup_{m \geq 1} m^{-1/2} (\text{E}[|X|^m])^{1/m}$. A random variable X is called sub-Gaussian if $\|X\|_{\psi_2} \leq C < \infty$ for a constant $C > 0$. A random vector $X \in \mathbb{R}^p$ is called sub-Gaussian if the one-dimensional marginals $X'b$ are sub-Gaussian random variables for all $b \in \mathbb{R}^p$. The sub-Gaussian norm for the random vector is defined as $\|X\|_{\psi_2} \equiv \sup_{\|b\|=1} \|X'b\|_{\psi_2}$. The sub-exponential norm of a random variable $X \in \mathbb{R}$ is $\|X\|_{\psi_1} \equiv \inf\{t > 0 : \text{E}[\exp(|X|/t)] \leq 2\}$. Moreover, define $\mathbf{1}_m \equiv (1, \ldots, 1)' \in \mathbb{R}^m$, $\mathbf{0}_m \equiv (0, \ldots, 0)' \in \mathbb{R}^m$, and let $\mathbf{0}_{l \times m}$ denote the $l \times m$ matrix of zeros, and $\boldsymbol{e}_k$ denote the unit vector whose k-th element is 1 and all remaining elements are 0 (with the dimension understood from the context). We use I_r to denote the $r \times r$ identity matrix. Let $F(x; \mu, \sigma^2, a, b)$ denote the CDF of an $N(\mu, \sigma^2)$ random variable truncated on the interval $[a, b]$: for $x \in [a, b]$, $F(x; \mu, \sigma^2, a, b) \equiv \{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)\}/\{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)\}$, where $\Phi(\cdot)$ denotes the standard normal CDF. The CDF is 0 for $x < a$ and 1 for $x > b$. Also, let $\Lambda(z) \equiv \exp(z)/(1+\exp(z))$ denote the CDF of the standard logistic distribution. For scalar functions of a vector argument, ∇ denotes the column-vector gradient. We abbreviate the central limit theorem, the continuous mapping theorem, and "with probability approaching one" as CLT, CMT, and wpa 1, respectively.

2 Model

We consider a GLM that specifies the conditional density of a scalar outcome variable y_i given a $(d+p) \times 1$ vector of covariates x_i as

$$f(y_i|x_i, \theta_0) = \exp\left(y_i x_i'\theta_0 - a(x_i'\theta_0)\right) c(y_i), \quad i = 1, \ldots, n,$$

where θ_0 is the true value of the parameter vector $\theta \in \mathbb{R}^{d+p}$, and $a(\cdot)$ and $c(\cdot)$ are known functions. To each observation $(y_i, x_i')'$ there corresponds a positive, bounded, and possibly random survey weight w_i, which may depend on the unit's own characteristics, for example through its stratum.[2] Let $g(y, x'\theta) \equiv -\log f(y, x'\theta)$ and define the weighted log-likelihood function as

$$L(\theta) \equiv -n^{-1} \sum_{i=1}^{n} w_i \, g(y_i, x_i'\theta). \tag{2.1}$$

The weighted likelihood framework is widely used in survey data analysis (Manski and Lerman, 1977; Cameron and Trivedi, 2009; Wooldridge, 2010) and accommodates, among others, the following stratification schemes.

Example 1 (Standard stratified sampling)**.** Let $\mathcal{Z}$ be the population for $z = (y, x')'$, assumed infinite (or containing a large number of units). $\mathcal{Z}$ is stratified into J nonempty, mutually exclusive and exhaustive strata $\mathcal{Z} = \bigcup_{j=1}^{J} \mathcal{Z}_j$. A sample of n_j observations $\{(y_{ij}, x_{ij}')'\}_{i=1}^{n_j}$ is drawn independently from stratum $\mathcal{Z}_j$, $j = 1, \ldots, J$. The strata sample sizes n_j are non-random, and the population frequencies $q_j = P[z \in \mathcal{Z}_j] > 0$ are known. Weights for stratum j are $w_i = q_j/(n_j/n)$ for $i \in$ stratum j, with $n = \sum_{j=1}^{J} n_j$. Pooling and reindexing the observations as $i = 1, \ldots, n$, the corresponding weighted log-likelihood function is

$$L(\theta) = -\sum_{j=1}^{J} q_j \left(n_j^{-1} \sum_{i=1}^{n_j} g(y_{ij}, x_{ij}'\theta) \right) = -n^{-1} \sum_{i=1}^{n} w_i \, g(y_i, x_i'\theta). \tag{2.2}$$

Example 2 (Exogenous stratification)**.** Let $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$. The population is stratified into J strata according to a deterministic function of x_i: $\mathcal{X} = \cup_{j=1}^{J} \mathcal{X}_j$, where $\mathcal{X}_j$, $j = 1, \ldots, J$, are mutually exclusive. Population frequencies $q_j = P[x \in \mathcal{X}_j] > 0$ are known. Given $n = \sum_{j=1}^{J} n_j$ observations drawn independently from each stratum, the likelihood function takes the form (2.2).

Wooldridge (2001) established the asymptotic properties of M-estimators under these sampling schemes. We establish the asymptotic validity of the proposed SI procedure more generally for independent, non-identically distributed triangular arrays, without requiring either particular sampling scheme.

[2] In our framework, $\{(y_i, x_i', w_i)'\}_{i=1}^{n}$ forms a triangular array $\{(y_{ni}, x_{ni}', w_{ni})'\}_{i=1}^{n}$. We drop the index n for notational simplicity.

The score function and negative Hessian corresponding to (2.1) are

$$S(\theta) \equiv \frac{\partial L(\theta)}{\partial \theta} = -n^{-1} \sum_{i=1}^{n} w_i x_i \dot{g}(y_i, x_i'\theta), \quad \dot{g}(y,t) \equiv \frac{\partial g(y,t)}{\partial t}, \tag{2.3}$$

$$\hat{H}(\theta) \equiv -\frac{\partial^2 L(\theta)}{\partial \theta \, \partial \theta'} = n^{-1} \sum_{i=1}^{n} w_i x_i x_i' \ddot{g}(y_i, x_i'\theta), \quad \ddot{g}(y,t) \equiv \frac{\partial^2 g(y,t)}{\partial t^2}. \tag{2.4}$$

Survey weights and information equality. Since $g(y,t) = -yt + a(t) - \log c(y)$, we have $\dot{g}(y,t) = a'(t) - y$ and $\ddot{g}(y,t) = a''(t)$. Suppose, for illustration, that the model is correctly specified and that w_i is a function of x_i. Then the score contribution $w_i x_i \dot{g}(y_i, x_i'\theta_0)$ has conditional mean zero and conditional variance $w_i^2 a''(x_i'\theta_0) x_i x_i'$, whereas the corresponding contribution to the negative Hessian is $w_i a''(x_i'\theta_0) x_i x_i'$. Each conditional variance thus equals w_i times the corresponding Hessian contribution, and this factor varies across observations. Hence $n^{-1} \sum_i w_i^2 a''(x_i'\theta_0) x_i x_i'$ is in general not equal, or even proportional, to $n^{-1} \sum_i w_i a''(x_i'\theta_0) x_i x_i'$. We refer to this feature as heteroskedasticity of the weighted score. As a consequence, the information matrix equality $\mathrm{Var}(n^{1/2} S(\theta_0)) = \mathrm{E}[\hat{H}(\theta_0)]$ fails in general, even under correct specification. Section 3.1 shows that this failure makes the random quantities that define the active and inactive KKT constraints statistically dependent.

Parameter partition. Partition the covariate and parameter vectors as $x_i = (x_{1i}', x_{2i}')' \in \mathbb{R}^{d+p}$ and $\theta = (\alpha', \beta')' \in \mathbb{R}^{d+p}$, where $x_{1i} \in \mathbb{R}^d$ collects the constant and any pre-specified regressors of interest, $\alpha \in \mathbb{R}^d$ are the corresponding unpenalized parameters, $x_{2i} \in \mathbb{R}^p$ are candidate regressors subject to selection, and $\beta \in \mathbb{R}^p$ are their coefficients. The linear index is $x_i'\theta = x_{1i}'\alpha + x_{2i}'\beta$. The special case $d = 1$, $x_{1i} = 1$ recovers the intercept-only formulation of the previous literature.

ℓ_1-PML. Variable selection and initial estimation are performed using the survey-weighted ℓ_1-PML estimator, obtained by minimizing the penalized negative log-likelihood

$$\min_{\theta = (\alpha', \beta')' \in \mathbb{R}^{d+p}} \Big(-L(\theta) + \lambda \|\beta\|_1 \Big), \tag{2.5}$$

where $\lambda \geq 0$ is a tuning parameter. Only the coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ are penalized; α is always estimated without regularization. Henceforth, the j-th elements of θ and θ_0 are denoted as $\theta_{(j)}$ and $\theta_{0(j)}$, respectively.

Selected model. For a given λ, the solution to (2.5) produces a random index set

$$\hat{M} \equiv \{1, \ldots, d\} \cup \{d + j : \hat{\beta}_j \neq 0,\ j = 1, \ldots, p\} \subseteq \{1, \ldots, d + p\},$$

which we call the *selected model*. The first d indices, corresponding to the unpenalized α, are always included in $\hat{M}$; the remaining indices in $\hat{M}$ are the components of β with non-zero ℓ_1-PML estimates.

For a generic (non-random) index set $M \subseteq \{1, \ldots, d + p\}$ with $\{1, \ldots, d\} \subseteq M$, let $|M|$ denote its cardinality and $x_{iM} \in \mathbb{R}^{|M|}$ denote the subvector of x_i corresponding to the indices in M. Let $X = [x_1, \ldots, x_n]'$ denote the $n \times (d + p)$ design matrix and $X_M = [x_{1M}, \ldots, x_{nM}]'$ its submatrix corresponding to M. Write $\theta_M = (\alpha', \beta_M')' \in \mathbb{R}^{|M|}$ and $\hat{\theta}_M = (\hat{\alpha}', \hat{\beta}_M')' \in \mathbb{R}^{|M|}$, where $\beta_M \in \mathbb{R}^{|M|-d}$ and $\hat{\beta}_M \in \mathbb{R}^{|M|-d}$ are the subvectors of β and $\hat{\beta}$ corresponding to the selected indices in $M \setminus \{1, \ldots, d\}$. Define $\hat{\mathbf{s}}_M \equiv \text{sign}(\hat{\beta}_M)$. On the event $\{\hat{M} = M\}$, $\hat{\mathbf{s}}_M \in \{-1, 1\}^{|M|-d}$.

Target parameter. For a generic model M, let $\theta_{0,M}$ denote the coordinate subvector of the full-model parameter θ_0 corresponding to M. The target of selective inference (see Section 3.1 below) is instead the population parameter of the restricted model,

$$\theta_M^* = \arg \min_{\theta_M \in \mathbb{R}^{|M|}} \mathcal{E}_M(\theta_M), \qquad \mathcal{E}_M(\theta_M) \equiv \mathrm{E}\left[n^{-1} \sum_{i=1}^{n} w_i\, g(y_i, x_{iM}'\theta_M) \right]. \tag{2.6}$$

Here and below, expectations are taken with respect to the joint distribution of the observations $(y_i, x_i', w_i)'$, $i = 1, \ldots, n$, including the survey weights when they are random.

Write $\theta_M^* = (\alpha_M^{*\prime}, \beta_M^{*\prime})'$. In general, θ_M^* need not equal $\theta_{0,M}$, since omitting an active regressor can change the population coefficients on the included regressors. Let $\mathcal{S}_0 \equiv \text{supp}(\theta_0)$. If $\mathcal{S}_0 \subseteq M$ and $\mathrm{E}[S(\theta_0)] = 0$, then $\theta_{0,M}$ minimizes $\mathcal{E}_M$. When this minimizer is unique, it follows that

$$\theta_M^* = \theta_{0,M}. \tag{2.7}$$

The proposed SI procedure targets θ_M^* and does not require $\mathcal{S}_0 \subseteq M$. In particular, always including the first d coordinates does not by itself imply $\alpha_M^* = \alpha_0$; that equality follows under the conditions stated above. Let $-M \equiv \{1, \ldots, d + p\} \setminus M$. Because $\{1, \ldots, d\} \subseteq M$, the indices in $-M$ correspond only to penalized regressors. For a restricted parameter $\theta_M \in \mathbb{R}^{|M|}$,

define $L_M(\theta_M) \equiv -n^{-1}\sum_{i=1}^n w_i g(y_i, x_{iM}'\theta_M)$,

$$S(\theta_M) \equiv -n^{-1}\sum_{i=1}^n w_i x_i \dot{g}(y_i, x_{iM}'\theta_M), \quad \hat{H}(\theta_M) \equiv n^{-1}\sum_{i=1}^n w_i x_i x_i' \ddot{g}(y_i, x_{iM}'\theta_M).$$

Thus, $S(\theta_M)$ and $\hat{H}(\theta_M)$ denote the full-dimensional score and negative Hessian evaluated at the full parameter vector with coordinates outside M set to zero. For $A, B \in \{M, -M\}$, define $S_A(\theta_M) \equiv [S(\theta_M)]_A$ and $\hat{H}_{A,B}(\theta_M) \equiv [\hat{H}(\theta_M)]_{A,B}$. Accordingly, after ordering the coordinates as $(M, -M)$,

$$S(\theta_M) = \begin{bmatrix} S_M(\theta_M) \\ S_{-M}(\theta_M) \end{bmatrix}, \qquad \hat{H}(\theta_M) = \begin{bmatrix} \hat{H}_{M,M}(\theta_M) & \hat{H}_{M,-M}(\theta_M) \\ \hat{H}_{-M,M}(\theta_M) & \hat{H}_{-M,-M}(\theta_M) \end{bmatrix}.$$

We abbreviate $\hat{H}_M(\theta_M) \equiv \hat{H}_{M,M}(\theta_M)$. Here $S_M(\theta_M) \in \mathbb{R}^{|M|}$ is the score for the restricted model, while $S_{-M}(\theta_M)$ records the score in the excluded directions. By the definition of θ_M^*, its population first-order condition is

$$\mathrm{E}[S_M(\theta_M^*)] = 0. \tag{2.8}$$

Throughout, we assume that the optimization problem (2.5) admits an almost surely unique solution. Sufficient conditions for existence and uniqueness of ℓ_1-penalized likelihood estimators are provided by Ali and Tibshirani (2019). Consequently, $\hat{\theta}$, $\hat{M}$, and the corresponding coefficient-sign vector are well-defined almost surely.

For a fixed model M containing $\{1, \ldots, d\}$ and a fixed sign vector $\mathtt{s}_M \in \{-1, 1\}^{|M|-d}$, the selection event $\{\hat{M} = M, \hat{\mathtt{s}}_M = \mathtt{s}_M\}$ occurs if and only if there exist $\hat{\theta}_M \in \mathbb{R}^{|M|}$ and $\mathtt{u} \in \mathbb{R}^{d+p-|M|}$ satisfying the Karush–Kuhn–Tucker (KKT) conditions for (2.5) (see Lee et al. (2016) and Taylor and Tibshirani (2018)):

$$S_M(\hat{\theta}_M) - (\mathbf{0}_d', \lambda\, \mathtt{s}_M')' = 0, \quad \mathtt{s}_M = \mathrm{sign}(\hat{\beta}_M), \tag{2.9}$$

$$S_{-M}(\hat{\theta}_M) - \lambda\, \mathtt{u} = 0, \quad \|\mathtt{u}\|_\infty \le 1. \tag{2.10}$$

Average marginal effects and scalar functionals. The conditional mean implied by the exponential family specification is $\mathrm{E}[y_i|x_i] = a'(x_i'\theta_0)$. For a continuous regressor x_{ij} with coefficient $\theta_{0(j)}$, the marginal effect on the conditional mean is $\frac{\partial}{\partial x_{ij}}\mathrm{E}[y_i|x_i] = a''(x_i'\theta_0)\,\theta_{0(j)}$. More generally, for a parameter vector θ, define $\mathrm{ME}_{ij}(\theta) = a''(x_i'\theta)\theta_{(j)}$ for a continuous regressor. For a binary regressor, the marginal effect is defined as the discrete change $\mathrm{ME}_{ij}(\theta) \equiv$

$a'(x_i'\theta)\big|_{x_{ij}=1} - a'(x_i'\theta)\big|_{x_{ij}=0}$. The corresponding AME is

$$\mathrm{AME}_j(\theta_0) \equiv \mathrm{E}\left[\frac{\sum_{i=1}^n w_i \,\mathrm{ME}_{ij}(\theta_0)}{\sum_{i=1}^n w_i}\right],$$

where the expectation is taken with respect to the joint distribution of the regressors and survey weights. A natural plug-in estimator replaces θ_0 by an estimator $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')'$, for example the survey-weighted ℓ_1-PML estimator:

$$\widehat{\mathrm{AME}}_j(\hat{\theta}) = \frac{\sum_{i=1}^n w_i \,\mathrm{ME}_{ij}(\hat{\theta})}{\sum_{i=1}^n w_i}.$$

For a selected model M, let $\mathrm{ME}_{ij}(\theta_M)$ denote the marginal effect of regressor $j \in M \setminus \{1, \ldots, d\}$, computed as above under the restricted model with linear index $x_{iM}'\theta_M$. Let $\beta_{M,j}$ denote the corresponding coefficient in θ_M. For a continuous regressor in the logit model, $\mathrm{ME}_{ij}(\theta_M) = \beta_{M,j}\Lambda'(x_{iM}'\theta_M)$. The corresponding AME is

$$\mathrm{AME}_{Mj}(\theta_M) \equiv \mathrm{E}\left[\frac{\sum_{i=1}^n w_i \,\mathrm{ME}_{ij}(\theta_M)}{\sum_{i=1}^n w_i}\right].$$

We collect the AMEs into the vector $\mathrm{AME}_M(\theta_M) = (\mathrm{AME}_{Mj}(\theta_M))_{j\in M\setminus\{1,\ldots,d\}} \in \mathbb{R}^{|M|-d}$.

More generally, for a fixed model M, let $\rho_M : \mathbb{R}^{|M|} \to \mathbb{R}$ denote a differentiable scalar population functional of θ_M, and let $\hat{\rho}_M$ denote its sample counterpart. For the k-th selected slope coefficient, take $\rho_M(\theta_M) = (\beta_M)_k$, $k \in \{1, \ldots, |M| - d\}$; for an unpenalized component, take $\rho_M(\theta_M) = \boldsymbol{e}_k'\alpha$, $k \in \{1, \ldots, d\}$. In both cases, $\hat{\rho}_M = \rho_M$. For an AME, take $\rho_M(\theta_M) = \mathrm{AME}_{Mj}(\theta_M)$, with sample counterpart

$$\hat{\rho}_M(\theta_M) = \frac{\sum_{i=1}^n w_i \,\mathrm{ME}_{ij}(\theta_M)}{\sum_{i=1}^n w_i}.$$

The population target is $\rho_M(\theta_M^*)$, and $\hat{\rho}_M(\hat{\theta}_M)$ is its plug-in estimator.

3 Selective Inference

The objective of SI is to construct confidence intervals that account for model selection. Since the selected model $\hat{M}$ is random, so is the population target $\rho_{\hat{M}}(\theta_{\hat{M}}^*)$. Moreover, $\theta_{\hat{M}}^*$ need not coincide with the corresponding subvector $\theta_{0,\hat{M}}$ of the full-model parameter. Conditional on

$\{\hat{M} = M\}$, however, the target is fixed at $\rho_M(\theta^*_M)$.

Conventional confidence intervals that ignore model selection need not attain the desired conditional coverage, since selection affects the sampling distribution of the estimator (Lee et al., 2016). For a fixed model M and scalar functional ρ_M, we build a confidence interval $\mathrm{CI}_{\hat{M}}$ satisfying

$$P\left[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \middle| \hat{M} = M\right] \to 1 - \zeta. \tag{3.1}$$

Following Lee et al. (2016), we first construct confidence intervals by additionally conditioning on the signs of the selected coefficients. This facilitates a polyhedral representation of the selection constraints and allows us to establish the sign-conditional coverage property

$$P\left[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \middle| \hat{M} = M,\ \hat{\mathbf{s}}_M = \mathbf{s}_M\right] \to 1 - \zeta. \tag{3.2}$$

Under suitable conditions, sign-conditional validity in (3.2) then implies the model-conditional coverage property in (3.1).

We establish the asymptotic validity of the proposed SI procedure for survey-weighted GLMs with a growing number of covariates, using the KKT characterization of the selection event. For a deterministic sequence m_n bounding the selected-model size, let

$$\mathcal{M}_n(m_n) \equiv \{M : \{1, \ldots, d\} \subseteq M,\ |M| \leq m_n\}.$$

3.1 Construction of selective confidence intervals

One-step estimator. For a fixed model M, define the restricted penalized estimator by

$$\hat{\theta}_M \in \arg \min_{\theta_M = (\alpha', \beta_M')'} \left(-L_M(\theta_M) + \lambda \|\beta_M\|_1\right). \tag{3.3}$$

On $\{\hat{M} = M, \hat{\mathbf{s}}_M = \mathbf{s}_M\}$, this estimator coincides with the active block of the full ℓ_1-PML estimator. For inference on the scalar population functional $\rho_M(\theta^*_M)$ introduced in Section 2, following Taylor and Tibshirani (2018), we define the one-step estimator of $\rho_M(\theta^*_M)$ by

$$\tilde{\rho}_M \equiv \hat{\rho}_M(\hat{\theta}_M) + \nabla \hat{\rho}_M(\hat{\theta}_M)' \hat{H}_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M). \tag{3.4}$$

For a parameter functional, this reduces to the usual one-step correction. For an AME, it also retains the sampling variation in the average of marginal effects. On the selection event, the KKT condition (2.9) gives $S_M(\hat{\theta}_M) = (\mathbf{0}_d', \lambda \mathbf{s}_M')'$. As in the classical one-step estimator (see,

e.g., Newey and McFadden, 1994), the correction removes the first-order penalization bias.

Let $H_M \equiv \mathrm{E}[\hat{H}_M(\theta_M^*)]$. Taylor expansions of the score and the sample counterpart $\hat{\rho}_M$ around θ_M^* yield, to first order,

$$n^{1/2}\left(\tilde{\rho}_M - \rho_M(\theta_M^*)\right) \approx \nabla\rho_M(\theta_M^*)' H_M^{-1} n^{1/2} S_M(\theta_M^*) + n^{1/2}\left(\hat{\rho}_M(\theta_M^*) - \rho_M(\theta_M^*)\right).$$

The last term vanishes for selected coefficients and unpenalized parameters, but is generally nonzero for population AMEs. To express the active and inactive selection constraints jointly, consider the one-step estimator and the orthogonalized inactive score

$$\tilde{\theta}_M \equiv \hat{\theta}_M + \hat{H}_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M), \tag{3.5}$$

$$\tilde{S}_{-M}(\hat{\theta}_M) \equiv S_{-M}(\hat{\theta}_M) - \hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M). \tag{3.6}$$

We collect these quantities together with the functional estimator in

$$Z \equiv n^{1/2} \begin{bmatrix} \tilde{\rho}_M \\ \tilde{\theta}_M \\ \tilde{S}_{-M}(\hat{\theta}_M) \end{bmatrix} \in \mathbb{R}^{d+p+1}. \tag{3.7}$$

The construction thus treats a general scalar functional as the first coordinate of the joint statistic, without requiring a separate polyhedral representation.

KKT conditions as polyhedral constraints. Next we express the KKT conditions in (2.9)-(2.10) as polyhedral constraints. Let $P \equiv [\mathbf{0}_{(|M|-d)\times d}, I_{|M|-d}]$ extract the β_M coordinates from θ_M. Write $q \equiv |M| - d$ and $r \equiv d + p - |M|$. Using $\|\mathtt{u}\|_\infty \leq 1$, the closure of the selection event $E_{M,\mathtt{s}_M} \equiv \{\hat{M} = M, \hat{\mathtt{s}}_M = \mathtt{s}_M\}$ has the polyhedral representation

$$\{AZ \leq \hat{b}\}, \tag{3.8}$$

where

$$A \equiv \begin{bmatrix} \mathbf{0}_q & -\mathrm{diag}(\mathtt{s}_M)P & \mathbf{0}_{q\times r} \\ \mathbf{0}_r & \mathbf{0}_{r\times|M|} & I_r \\ \mathbf{0}_r & \mathbf{0}_{r\times|M|} & -I_r \end{bmatrix}, \; \hat{b} \equiv n^{1/2} \begin{bmatrix} -\mathrm{diag}(\mathtt{s}_M)P\hat{H}_M(\hat{\theta}_M)^{-1} v_M \\ \lambda\mathbf{1}_r - \hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} v_M \\ \lambda\mathbf{1}_r + \hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} v_M \end{bmatrix}, \; v_M \equiv \begin{bmatrix} \mathbf{0}_d \\ \lambda\mathtt{s}_M \end{bmatrix}. \tag{3.9}$$

The zero first column of A shows that including $\tilde{\rho}_M$ in Z does not alter the selection event.

Following Lee et al. (2016), Proposition 4.2, we show that the active and inactive inequalities in (3.8) are equivalent to the corresponding KKT conditions in (2.9)–(2.10). First, the active constraints satisfy

$$\begin{aligned}\left\{\operatorname{sign}(\hat{\beta}_M) = \mathtt{s}_M\right\} &= \left\{\operatorname{diag}(\mathtt{s}_M) P \left[\tilde{\theta}_M - \hat{H}_M(\hat{\theta}_M)^{-1} v_M\right] > 0\right\} \\ &= \left\{-\operatorname{diag}(\mathtt{s}_M) P \tilde{\theta}_M < -\operatorname{diag}(\mathtt{s}_M) P \hat{H}_M(\hat{\theta}_M)^{-1} v_M\right\}.\end{aligned}$$

To verify the first equality, note that, if $\operatorname{sign}(\hat{\beta}_M) = \mathtt{s}_M$, by the restricted KKT conditions, $S_M(\hat{\theta}_M) = v_M$, and hence (3.5) gives $\tilde{\theta}_M - \hat{H}_M(\hat{\theta}_M)^{-1} v_M = \hat{\theta}_M$. Thus, the sign restrictions imply the active inequalities. Conversely, suppose that the active inequalities hold. By the restricted KKT conditions and (3.5),

$$\begin{aligned}0 &\leq \left[\tilde{\theta}_M - \hat{H}_M(\hat{\theta}_M)^{-1} v_M - \hat{\theta}_M\right]' \left[v_M - S_M(\hat{\theta}_M)\right] \\ &= -\left[S_M(\hat{\theta}_M) - v_M\right]' \hat{H}_M(\hat{\theta}_M)^{-1} \left[S_M(\hat{\theta}_M) - v_M\right] \leq 0.\end{aligned}$$

The first inequality follows from the active sign constraints and the coordinatewise restricted KKT conditions. Since $\hat{H}_M(\hat{\theta}_M)$ is positive definite, we obtain $S_M(\hat{\theta}_M) = v_M$. Hence, by (3.5), the active inequalities reduce to $\operatorname{diag}(\mathtt{s}_M) P \hat{\theta}_M > 0$, or equivalently, $\operatorname{sign}(\hat{\beta}_M) = \mathtt{s}_M$. Next, given $S_M(\hat{\theta}_M) = v_M$, the inactive KKT condition (2.10) is equivalent to

$$\begin{aligned}\left\{\|S_{-M}(\hat{\theta}_M)\|_\infty \leq \lambda\right\} &= \left\{-\lambda \mathbf{1}_r \leq \tilde{S}_{-M}(\hat{\theta}_M) + \hat{H}_{-M,M}(\hat{\theta}_M) \hat{H}_M(\hat{\theta}_M)^{-1} v_M \leq \lambda \mathbf{1}_r\right\} \\ &= \left\{\begin{array}{l} \tilde{S}_{-M}(\hat{\theta}_M) \leq \lambda \mathbf{1}_r - \hat{H}_{-M,M}(\hat{\theta}_M) \hat{H}_M(\hat{\theta}_M)^{-1} v_M, \\ -\tilde{S}_{-M}(\hat{\theta}_M) \leq \lambda \mathbf{1}_r + \hat{H}_{-M,M}(\hat{\theta}_M) \hat{H}_M(\hat{\theta}_M)^{-1} v_M \end{array}\right\}.\end{aligned}$$

These are exactly the remaining inequalities in (3.8).

We have therefore established both directions of the equivalence: the active inequalities recover the coefficient signs and active KKT conditions, while the inactive inequalities recover the inactive KKT conditions. Since the KKT conditions are necessary and sufficient for the penalized optimization problem and its solution is unique, the combined inequalities characterize the selection event $E_{M,\mathtt{s}_M}$. The exact characterization uses strict inequalities for the active signs and weak inequalities for inactive feasibility; (3.8) uses the corresponding boundary-inclusive representation.

Gaussian approximation by Yurinskii coupling. For a generic model $M \in \mathcal{M}_n(m_n)$, define the population full Hessian $H(\theta_M^*) \equiv \mathrm{E}[\hat{H}(\theta_M^*)]$, with blocks $H_M \equiv [H(\theta_M^*)]_{M,M}$ and $H_{-M,M} \equiv [H(\theta_M^*)]_{-M,M}$. Define the full-dimensional score contribution $s_{i,M} \equiv -w_i x_i \dot{g}(y_i, x_{iM}'\theta_M^*)$ and $\xi_{i,M} \equiv s_{i,M} - \mathrm{E}[s_{i,M}]$. The covariance of the centered score is

$$I_M \equiv n^{-1} \sum_{i=1}^{n} \mathrm{E}[\xi_{i,M}\xi_{i,M}'] = n^{-1} \sum_{i=1}^{n} \mathrm{Var}(s_{i,M}). \tag{3.10}$$

For a functional with a sample-dependent counterpart, suppose that

$$n^{1/2}\left(\hat{\rho}_M(\theta_M^*) - \rho_M(\theta_M^*)\right) = n^{-1/2} \sum_{i=1}^{n} \tau_{i,M} + o_p\left((\log 2p)^{-1/2}\right), \tag{3.11}$$

where $\tau_{i,M}$ are independent, mean-zero influence-function contributions. For a parameter functional, set $\tau_{i,M} = 0$. Define the augmented score covariance by

$$\omega_{i,M} \equiv \begin{bmatrix} \xi_{i,M} \\ \tau_{i,M} \end{bmatrix}, \qquad \Omega_M \equiv n^{-1} \sum_{i=1}^{n} \mathrm{E}[\omega_{i,M}\omega_{i,M}']. \tag{3.12}$$

The upper-left block of Ω_M is I_M. Let $\hat{\Omega}_M$ be a symmetric positive-semidefinite estimator of Ω_M; its construction may use grouping variables or replicate weights, as discussed in Section 3.3.

To estimate the covariance matrix of Z, define

$$\hat{\Sigma}_M \equiv \hat{Q}_M \hat{\Omega}_M \hat{Q}_M', \quad \hat{Q}_M \equiv \begin{bmatrix} \nabla\hat{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1} & \mathbf{0}_r' & 1 \\ \hat{H}_M(\hat{\theta}_M)^{-1} & \mathbf{0}_{|M|\times r} & \mathbf{0}_{|M|} \\ -\hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} & I_r & \mathbf{0}_r \end{bmatrix}. \tag{3.13}$$

Its population counterpart is

$$\Sigma_M \equiv Q_M \Omega_M Q_M', \quad Q_M \equiv \begin{bmatrix} \nabla\rho_M(\theta_M^*)' H_M^{-1} & \mathbf{0}_r' & 1 \\ H_M^{-1} & \mathbf{0}_{|M|\times r} & \mathbf{0}_{|M|} \\ -H_{-M,M}H_M^{-1} & I_r & \mathbf{0}_r \end{bmatrix}. \tag{3.14}$$

The restricted population first-order condition gives $\mathrm{E}[S_M(\theta_M^*)] = 0$, but it need not imply $\nu_{-M} \equiv \mathrm{E}[S_{-M}(\theta_M^*)] = 0$. An omitted active regressor can leave a nonzero population score in the excluded directions. If $\mathcal{S}_0 \subseteq M$, $\mathrm{E}[S(\theta_0)] = 0$, and $\theta_M^* = \theta_{0,M}$, then $\nu_{-M} = 0$. Define the

Gaussian mean and target variance by

$$\mu_M \equiv n^{1/2} \begin{bmatrix} \rho_M(\theta_M^*) \\ \theta_M^* \\ \nu_{-M} \end{bmatrix}, \qquad \sigma_M^2 \equiv \boldsymbol{e}_1' \Sigma_M \boldsymbol{e}_1. \tag{3.15}$$

Partition I_M conformably with $(M, -M)$ and let $H_{M,-M} \equiv H'_{-M,M}$. By (3.14), the covariance between the active component $H_M^{-1} n^{1/2} S_M(\theta_M^*)$ and the orthogonalized inactive score $n^{1/2}\left(S_{-M}(\theta_M^*) - H_{-M,M} H_M^{-1} S_M(\theta_M^*)\right)$ is

$$H_M^{-1}\left([I_M]_{M,-M} - [I_M]_{M,M} H_M^{-1} H_{M,-M}\right).$$

Under the information matrix equality $I_M = H(\theta_M^*)$, which holds in the unweighted GLM considered by Taylor and Tibshirani (2018) when the selected model contains the true support, this covariance is zero. The active and inactive components are then asymptotically independent, and, for a parameter functional, the inactive constraints can be dropped. With survey weights, $I_M \neq H(\theta_M^*)$ in general (Section 2), even when the selected model is correctly specified; an underfitted selected model has the same effect. The covariance then need not vanish, and the two components are statistically dependent. We therefore account for their joint covariance in the Gaussian approximation and condition on both sets of constraints. For a population AME, Ω_M additionally captures the variance of the empirical average and its covariance with the selection scores.

Under the rate conditions in Proposition 3.2, Yurinskii coupling provides a Gaussian approximation to Z; see Appendix A.4. Specifically, there exists a Gaussian vector $T_M \sim N(0, \Sigma_M)$ such that

$$\|Z - \mu_M - T_M\|_\infty \xrightarrow{p} 0. \tag{3.16}$$

We summarize this coupling as

$$Z \overset{a}{\sim} N(\mu_M, \Sigma_M). \tag{3.17}$$

Truncated Gaussian pivot. Combined with the polyhedral constraint (3.8), the Gaussian approximation (3.17) leads to a truncated Gaussian pivot. For $z \in \mathbb{R}^k$, a covariance matrix Σ, and a contrast η with $\eta' \Sigma \eta > 0$, define

$$c(\Sigma, \eta) \equiv \frac{\Sigma \eta}{\eta' \Sigma \eta}, \qquad r(z, \Sigma, \eta) \equiv \left(I_k - c(\Sigma, \eta) \eta'\right) z. \tag{3.18}$$

For fixed A, b and $c = c(\Sigma, \eta)$, define

$$\mathcal{V}^-(r) \equiv \max_{j:(Ac)_j<0} \frac{b_j - (Ar)_j}{(Ac)_j}, \tag{3.19}$$

$$\mathcal{V}^+(r) \equiv \min_{j:(Ac)_j>0} \frac{b_j - (Ar)_j}{(Ac)_j}, \tag{3.20}$$

$$\mathcal{V}^0(r) \equiv \min_{j:(Ac)_j=0} (b_j - (Ar)_j). \tag{3.21}$$

Here $(Ac)_j$ denotes the jth element of Ac; we use $\max \varnothing = -\infty$ and $\min \varnothing = +\infty$. The following result collects Lemma 5.1 and Theorem 5.2 of Lee et al. (2016).

Lemma 3.1 (Polyhedral lemma and truncated Gaussian pivot)**.** *Let $G \sim N(\mu, \Sigma)$ and let A, b, and η be fixed, with $\eta'\Sigma\eta > 0$. Write $c = c(\Sigma, \eta)$ and $r = r(G, \Sigma, \eta)$. Then $\eta'G$ is independent of r, and*

$$\{AG \le b\} = \{\mathcal{V}^-(r) \le \eta'G \le \mathcal{V}^+(r),\ \mathcal{V}^0(r) \ge 0\}. \tag{3.22}$$

If $P(AG \le b) > 0$, then

$$F\left(\eta'G; \eta'\mu, \eta'\Sigma\eta, \mathcal{V}^-(r), \mathcal{V}^+(r)\right) \mid \{AG \le b\} \sim U(0,1). \tag{3.23}$$

Selective confidence interval. Set $c_M \equiv c(\Sigma_M, \boldsymbol{e}_1)$. In our application, take $\eta = \boldsymbol{e}_1$ and let $\hat{c}_M \equiv c(\hat{\Sigma}_M, \boldsymbol{e}_1)$ and $\hat{r} \equiv r(Z, \hat{\Sigma}_M, \boldsymbol{e}_1)$. The truncation limits $\mathcal{V}^-(\hat{r})$ and $\mathcal{V}^+(\hat{r})$ are computed using A, $\hat{b}$, and $\hat{c}_M$ in (3.19)–(3.20). For $\zeta \in (0,1)$, define the level $1-\zeta$ SI confidence interval $\mathrm{CI}_{\hat{M}} \equiv [\tilde{q}_l, \tilde{q}_u]$ for $\rho_M(\theta^*_M)$, where the lower and upper endpoints solve

$$F\left(n^{1/2}\tilde{\rho}_M; n^{1/2}\tilde{q}_l, \boldsymbol{e}_1'\hat{\Sigma}_M\boldsymbol{e}_1, \mathcal{V}^-(\hat{r}), \mathcal{V}^+(\hat{r})\right) = 1 - \frac{\zeta}{2}, \tag{3.24}$$

$$F\left(n^{1/2}\tilde{\rho}_M; n^{1/2}\tilde{q}_u, \boldsymbol{e}_1'\hat{\Sigma}_M\boldsymbol{e}_1, \mathcal{V}^-(\hat{r}), \mathcal{V}^+(\hat{r})\right) = \frac{\zeta}{2}. \tag{3.25}$$

Although A is fixed for a given model and sign vector, $\hat{b}$ and $\hat{\Sigma}_M$ are random and Z is only approximately Gaussian. Lemma 3.1 therefore does not by itself establish validity of the feasible sample pivot. The following result supplies the needed rate, stability, and selection probability conditions.

3.2 Asymptotic validity

We now state the main results establishing the asymptotic validity of selective inference for a general parameter functional. The regularity conditions required for these results are collected in Assumptions 1, 2 and 3 below.

Proposition 3.2 (Asymptotic validity of SI)**.** *Fix a deterministic sequence $M = M_n \in \mathcal{M}_n(m_n)$, a sign vector $\mathbf{s}_M$, and a functional ρ_M. Let (3.11) and Assumptions 1, 2, and 3 hold, and suppose that $\lambda = C\sqrt{(\log 2p)/n}$ for some fixed $0 < C < \infty$. If*

$$\frac{p^2(\log 2p)^6}{n} \to 0, \quad m_n \log 2p \sqrt{\frac{p}{n}} \to 0, \quad \text{and} \quad \frac{m_n^3(\log 2p)^3}{n} \to 0,$$

then $\mathrm{CI}_{\hat{M}}$, defined by the endpoints solving (3.24)–(3.25), satisfies (3.2) for every $\zeta \in (0,1)$.

Corollary 3.3 (Model-conditional validity)**.** *Fix a deterministic sequence $M = M_n \in \mathcal{M}_n(m_n)$ and a scalar functional ρ_M. Suppose (3.11) and Assumptions 1, 2, and 3(a) hold, together with the tuning parameter and rate conditions of Proposition 3.2. If*

$$\sup_n(|M| - d) < \infty, \qquad \liminf_{n\to\infty} P(\hat{M} = M) > 0,$$

then the confidence interval $\mathrm{CI}_{\hat{M}}$ defined by (3.24)–(3.25) satisfies (3.1).

The proofs of Proposition 3.2 and Corollary 3.3 are given in Appendix A.4 and Appendix A.5, respectively.

The assumption of fixed p is commonly used in the SI literature (Lee et al., 2016; Tian and Taylor, 2017; Taylor and Tibshirani, 2018; Kuchibhotla. et al., 2022), while Taylor and Tibshirani (2018) provide a heuristic argument for standard GLMs. Our main result establishes asymptotic validity for survey-weighted GLMs with i.n.i.d. observations and a growing number of covariates. Survey weighting and heteroskedasticity may induce dependence between the random quantities defining the active and inactive KKT constraints, which are therefore treated jointly. The result also accommodates inference on pseudo-true parameters of possibly misspecified selected models and nonlinear functionals, including population AMEs.

Corollary 3.3 establishes that the same SI confidence intervals attain asymptotic coverage conditional on the selected model alone. The result follows by averaging over the possible sign configurations, provided the number of selected penalized regressors remains bounded and the model-selection probability does not vanish. Thus, model-conditional validity is achieved without modifying the inference procedure, while allowing the number of candidate regressors

p to grow with the sample size. When the number of selected penalized regressors grows, the averaging argument would require uniform control over an increasing number of sign configurations, and we do not claim model-conditional validity in that case.

The rate conditions are sufficient and may be conservative. In particular, the first condition arises from the Gaussian coupling of the $(d+p+1)$-dimensional augmented score vector. For bounded m_n, it is the most restrictive condition and requires $p = o(\sqrt{n}/(\log 2p)^3)$. Under i.i.d. sampling, sharper Gaussian approximations tailored to the polyhedral selection event could potentially relax this restriction. The present theory does not cover $p \geq n$. Extending SI to that regime would require a different treatment of the high-dimensional Hessian and covariance matrices and of the growing polyhedral selection event, which we leave for future research.

We next provide the regularity conditions underlying Proposition 3.2 and Corollary 3.3. The first set of assumptions imposes regularity conditions at the selected-model population target θ^*_M rather than at the full-model parameter.

Assumption 1 (Sampling and model regularity)**.** *For every deterministic sequence $M = M_n \in \mathcal{M}_n(m_n)$, the following conditions hold:*

(a) $\{(y_i, x_i', w_i)'\}_{i=1}^n$ *are independent across i. For some constants $0 < C_l < C_u < \infty$, $\max_{1\leq i\leq n} \|x_i\|_{\psi_2} \leq C_u$, $\max_{1\leq i\leq n} \|x_i\|_\infty \leq C_u$ a.s., and $C_l \leq w_i \leq C_u$ a.s. for all i.*

(b) *$g(y,t)$ is convex and twice continuously differentiable in $t \in \mathbb{R}$ for every $y \in \mathcal{Y}$. There are constants $\bar{\eta}, L_0, C_g > 0$, independent of M and n, such that the following conditions hold uniformly over i, $y \in \mathcal{Y}$, and $t, \tilde{t}$ satisfying $\max\{|t - x_{iM}'\theta^*_M|, |\tilde{t} - x_{iM}'\theta^*_M|\} \leq \bar{\eta}$:*

$$|\dot{g}(y, x_{iM}'\theta^*_M)| \leq C_g, \quad |\ddot{g}(y,t)| \leq C_g, \quad |\ddot{g}(y,t) - \ddot{g}(y,\tilde{t})| \leq L_0|t - \tilde{t}|.$$

(c) *$\mathcal{E}_M$ in (2.6) has a unique interior minimizer θ^*_M, and $\max_i |x_{iM}'\theta^*_M| \leq C_u$ a.s.*

(d) *There are constants $0 < c_H < C_H < \infty$ and $0 < c_I < C_I < \infty$, independent of M and n, such that $\lambda_{\min}(H_M) \geq c_H$, $\|H(\theta^*_M)\|_2 \leq C_H$, and $c_I \leq \lambda_{\min}(I_M) \leq \lambda_{\max}(I_M) \leq C_I$.*

(e) *The functional ρ_M is continuously differentiable in a neighborhood of θ^*_M, with a locally Lipschitz Jacobian satisfying $0 < c_\rho \leq \|\nabla\rho_M(\theta^*_M)\|_2^2 \leq C_\rho < \infty$.*

The sample functional $\hat{\rho}_M$ has a Lipschitz Jacobian on a fixed neighborhood of θ^*_M, with Lipschitz constant $O_p(1)$, and (3.11) holds with independent, mean-zero $\tau_{i,M}$ having uni-

formly bounded sub-Gaussian norms. Moreover, $\|\nabla\hat{\rho}_M(\theta_M^*) - \nabla\rho_M(\theta_M^*)\|_2 = O_p\left(\sqrt{\frac{m_n}{n}}\right)$, *and* $\inf_n \sigma_M^2 > 0$.

Assumption 1(a) allows independent, non-identically distributed observations, including the stratified sampling designs in Section 2. Survey weights may be random but must be uniformly bounded above and away from zero. The weights need not sum to one. The `R` package `glmnet` rescales observation weights to sum to n, and we adopt this mean-one convention in the numerical implementation.

The remaining conditions impose at θ_M^* analogues of the curvature, smoothness, and variance conditions commonly imposed at θ_0 in analyses of regularized M estimators and GLMs (Negahban et al., 2012; van de Geer et al., 2014; Xia et al., 2023).

In particular, Assumption 1(e) requires a smooth, nondegenerate functional and controls the empirical approximation of its value and derivative. The influence-function expansion (3.11) is immediate for parameter functionals, for which the empirical averaging term vanishes. In Section 3.3, we verify this expansion for the population AME and describe the corresponding covariance estimation.

Assumption 2 (Joint covariance estimation)**.** *For every deterministic sequence* $M = M_n \in \mathcal{M}_n(m_n)$, *there exists a deterministic sequence* $r_{I,n} \to 0$ *such that*

$$\|\hat{\Omega}_M - \Omega_M\|_2 = O_p(r_{I,n}), \qquad (\log 2p)^{1/2} r_{I,n} \to 0. \tag{3.26}$$

Assumption 2 accommodates survey-specific covariance estimators. For a parameter functional, $\tau_{i,M} = 0$, so only the score covariance I_M needs to be estimated. For a population AME, the joint covariance also accounts for the variability of empirical averaging and its covariance with the pseudo-score.

Assumption 3 (Selection-event regularity)**.**

(a) *For some* $\kappa > 0$, *the nonzero polyhedral slopes satisfy* $\min_{j:(Ac_M)_j \neq 0} |(Ac_M)_j| \geq \kappa$ *for all sufficiently large* n.

(b) *The selection event satisfies* $\liminf_{n\to\infty} P[E_{M,\mathbf{s}_M}] > 0$.

Assumption 3 imposes the stability conditions needed for conditional inference. Part (a) keeps the nonzero polyhedral slopes away from zero, while Part (b) ensures that approximation errors remain negligible after conditioning on selection. The growth rates in Proposition 3.2

control the Gaussian coupling, linearization, covariance, and threshold approximation errors. The nonvanishing selection probability condition excludes asymptotically rare selection events, including those associated with underfitted models whose inactive population scores remain bounded away from zero as $\lambda \to 0$.

3.3 Covariance estimation and auxiliary results

We first verify the influence-function and Jacobian conditions in (3.11) and Assumption 1(e) for the population AME in the survey-weighted logit model. We then describe estimators of the score and augmented covariance matrices and establish the sample matrix rates used in the proof of Proposition 3.2.

AME linearization and Jacobian convergence. For a population AME, the additional influence-function contribution in (3.12) is

$$\tau_{i,M} \equiv \frac{w_i \left(\mathrm{ME}_{ij}(\theta_M^*) - \rho_M(\theta_M^*)\right) - \mathrm{E}\left[w_i \left(\mathrm{ME}_{ij}(\theta_M^*) - \rho_M(\theta_M^*)\right)\right]}{n^{-1}\sum_{k=1}^n \mathrm{E}[w_k]}. \tag{3.27}$$

To verify (3.11), recall that $\hat{\rho}_M(\theta_M^*) = (\sum_{i=1}^n w_i \mathrm{ME}_{ij}(\theta_M^*))/(\sum_{i=1}^n w_i)$. Under Assumption 1(a), the observations are independent and the weights are bounded above and away from zero.

The individual marginal effects are uniformly bounded in the weighted logit model. For binary regressors, this follows from the boundedness of Λ. For continuous regressors, $\mathrm{ME}_{ij}(\theta_M) = \beta_{M,j}\Lambda'(x_{iM}'\theta_M)$. The uniform bound $\lambda_{\min}(H_M) \geq c_H > 0$ implies $c_H\|\theta_M^*\|_2^2 \leq \theta_M^{*\prime} H_M \theta_M^* \leq C_u \sup_{t\in\mathbb{R}} t^2\Lambda'(t) < \infty$. Hence $\|\theta_M^*\|_2 = O(1)$ uniformly over $M \in \mathcal{M}_n(m_n)$, and the marginal effects are bounded in this case as well.

A second-order ratio expansion therefore shows that the expectation of the sample ratio differs from the ratio of expectations by $O(n^{-1})$. Since $\rho_M(\theta_M^*) = \mathrm{E}[\hat{\rho}_M(\theta_M^*)]$, a first-order ratio expansion yields

$$n^{1/2}\left(\hat{\rho}_M(\theta_M^*) - \rho_M(\theta_M^*)\right) = n^{-1/2}\sum_{i=1}^n \tau_{i,M} + O_p(n^{-1/2}). \tag{3.28}$$

Since $(\log 2p)/n \to 0$ follows from $p^2/n \to 0$ in Proposition 3.2, the remainder is $o_p((\log 2p)^{-1/2})$, verifying (3.11).

We next verify the Jacobian convergence rate in Assumption 1(e). Differentiation of the sample AME gives $\nabla\hat{\rho}_M(\theta_M^*) = (\sum_{i=1}^n w_i \nabla \mathrm{ME}_{ij}(\theta_M^*))/(\sum_{i=1}^n w_i)$. By differentiation under the

expectation, $\nabla\rho_M(\theta_M^*) = \mathrm{E}[\nabla\hat{\rho}_M(\theta_M^*)]$. The elements of $\nabla\mathrm{ME}_{ij}(\theta_M^*)$ are uniformly bounded. For binary regressors, this follows from bounded regressors and $|\Lambda'(t)| \leq 1/4$. For continuous regressors, differentiation of $\beta_{M,j}\Lambda'(x_{iM}'\theta_M)$ involves Λ', Λ'', and $\beta_{M,j}$, which are uniformly bounded at θ_M^* under the conditions established above.

By independence and the uniform boundedness of the weights and marginal-effect Jacobians,

$$\mathrm{E}\left[\left\|n^{-1}\sum_{i=1}^{n}\{w_i\nabla\mathrm{ME}_{ij}(\theta_M^*) - \mathrm{E}[w_i\nabla\mathrm{ME}_{ij}(\theta_M^*)]\}\right\|_2^2\right] = O(|M|/n),$$

and $\mathrm{E}\left[|n^{-1}\sum_{i=1}^{n}(w_i - \mathrm{E}[w_i])|^2\right] = O(n^{-1})$. Since $n^{-1}\sum_i w_i \geq C_l$ and $\|n^{-1}\sum_i \mathrm{E}[w_i\nabla\mathrm{ME}_{ij}(\theta_M^*)]\|_2 = O(\sqrt{|M|})$, a direct expansion of the sample ratio gives

$$\mathrm{E}\left[\left\|\nabla\hat{\rho}_M(\theta_M^*) - \frac{\sum_i \mathrm{E}[w_i\nabla\mathrm{ME}_{ij}(\theta_M^*)]}{\sum_i \mathrm{E}[w_i]}\right\|_2^2\right] = O(|M|/n).$$

As $\nabla\rho_M(\theta_M^*) = \mathrm{E}[\nabla\hat{\rho}_M(\theta_M^*)]$, centering at this expectation cannot increase the mean squared deviation. Hence $\mathrm{E}\left[\|\nabla\hat{\rho}_M(\theta_M^*) - \nabla\rho_M(\theta_M^*)\|_2^2\right] = O(|M|/n) = O(m_n/n)$, which by Markov's inequality implies $\|\nabla\hat{\rho}_M(\theta_M^*) - \nabla\rho_M(\theta_M^*)\|_2 = O_p(\sqrt{m_n/n})$, as required by Assumption 1(e).

The local Lipschitz conditions in Assumption 1(e) also hold for the weighted logit AME. Since Λ', Λ'', and Λ''' are uniformly bounded and $\|\theta_M^*\|_2 = O(1)$, the Hessian of the sample functional is bounded on any fixed neighborhood of θ_M^* by

$$\sup_{\|\theta_M - \theta_M^*\|_2 \leq \varepsilon} \|\nabla^2\hat{\rho}_M(\theta_M)\|_2 \leq C\left(1 + \|n^{-1}X'X\|_2\right) = O_p(1),$$

where the last equality follows from (3.31). For binary regressors, the same bound holds because setting one bounded regressor to zero or one preserves the order of the sample design-matrix bound. The corresponding population Jacobian is locally Lipschitz by the same derivative bounds and the uniform population second-moment bound.

Covariance estimation. To estimate the score covariance I_M, suppose that observations can be partitioned into groups $\mathcal{I}_h$, $h = 1, \ldots, J$, within which the population score mean is common. Let $n_h = |\mathcal{I}_h| \geq 2$ and define

$$\hat{s}_{i,M} \equiv -w_i x_i \dot{g}(y_i, x_{iM}'\hat{\theta}_M), \quad \bar{s}_{h,M} \equiv n_h^{-1}\sum_{i\in\mathcal{I}_h}\hat{s}_{i,M},$$

$$\hat{I}_M^{\mathrm{grp}} \equiv \frac{1}{n}\sum_{h=1}^{J}\frac{n_h}{n_h-1}\sum_{i\in\mathcal{I}_h}(\hat{s}_{i,M}-\bar{s}_{h,M})(\hat{s}_{i,M}-\bar{s}_{h,M})'. \tag{3.29}$$

The factor $n_h/(n_h-1)$ is the usual degrees-of-freedom correction. Group centering accounts for nonzero, group-specific score means without requiring identical distributions within groups. In particular, it accommodates i.n.i.d. observations whose score means need not vanish individually, although their average is zero in the active coordinates. This centering does not change the target parameter θ_M^*.

When $\mathrm{E}[s_{i,M}]=0$ for every i, no mean correction is needed, and the outer-product estimator $n^{-1}\sum_i \hat{s}_{i,M}\hat{s}_{i,M}'$ targets I_M. For a correctly specified logit model with survey weights fixed or measurable with respect to the regressors, the conditional information identity also yields

$$\hat{I}_M^{\mathrm{logit}} = n^{-1}\sum_{i=1}^{n} w_i^2 x_i x_i' \hat{p}_{i,M}(1-\hat{p}_{i,M}), \qquad \hat{p}_{i,M}=\Lambda(x_{iM}'\hat{\theta}_M).$$

This form is used in the simulations, but not in the empirical application, where the selected model may be misspecified.

For the population AME, a feasible influence-function contribution is

$$\hat{\tau}_{i,M} \equiv \frac{w_i\left(\mathrm{ME}_{ij}(\hat{\theta}_M)-\hat{\rho}_M(\hat{\theta}_M)\right)}{n^{-1}\sum_{k=1}^{n} w_k}.$$

A grouped estimator of the augmented covariance Ω_M is obtained by replacing $\hat{s}_{i,M}$ in (3.29) with $\hat{\omega}_{i,M}=(\hat{s}_{i,M}',\hat{\tau}_{i,M})'$. Its consistency under a common within-group mean condition is established in Lemma 3.5.

Auxiliary matrix results. We next establish the sample matrix rates used in the proof, for the design matrices X and X_M defined in Section 2.

Lemma 3.4 (Sample matrix concentration)**.** *Suppose that Assumption 1(a)–(b) holds. For every deterministic sequence $M=M_n\in\mathcal{M}_n(m_n)$, if $p/n\to 0$, then*

$$\|\hat{H}(\theta_M^*)-H(\theta_M^*)\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right), \tag{3.30}$$

$$\|n^{-1}X'X\|_2 = O_p(1). \tag{3.31}$$

The proof is given in Appendix A.1, using covariance concentration results for sub-Gaussian vectors (Vershynin, 2010, 2018).

Lemma 3.5 (Grouped covariance estimation)**.** *Suppose Assumption 1 holds, and let* $M = M_n \in \mathcal{M}_n(m_n)$ *be deterministic. Suppose the observations admit a partition* $\{\mathcal{I}_h\}_{h=1}^J$*, with* J *fixed, such that*

$$\mathrm{E}[s_{i,M}] = \mu_{h,M}, \qquad i \in \mathcal{I}_h, \tag{3.32}$$

and, for constants $c_s, c_\lambda > 0$, $\min_h n_h/n \geq c_s$ *and* $\lambda \geq c_\lambda n^{-1/2}$ *for all sufficiently large* n*. If* $p/n \to 0$ *and* $m_n(n^{-1/2} + \lambda) \to 0$*, then*

$$\|\hat{I}_M^{\mathrm{grp}} - I_M\|_2 = O_p\Big(\sqrt{\frac{p}{n}} + m_n\lambda\Big). \tag{3.33}$$

For a population AME, if additionally $\mathrm{E}[w_i\{\mathrm{ME}_{ij}(\theta_M^*) - \rho_M(\theta_M^*)\}]$ *is constant within each group, then*

$$\|\hat{\Omega}_M^{\mathrm{grp}} - \Omega_M\|_2 = O_p\Big(\sqrt{\frac{p}{n}} + m_n\lambda\Big). \tag{3.34}$$

The proof is given in Appendix A.3. Moreover, $\sqrt{\log 2p}(\sqrt{p/n} + m_n\lambda) = o(1)$ under the rate conditions of Proposition 3.2. Thus, the grouped estimators verify Assumption 2 for parameter functionals and, under the additional group-mean condition, for population AMEs.

4 Simulations

This section presents simulation evidence on the performance of the proposed selective inference procedure. We consider a logit model under two sampling schemes: standard stratified sampling and exogenous stratification. The empirical size of the tests is examined for different sample sizes and numbers of regressors.

Data-generating process. For each scheme, we consider four strata with population proportions $q = (0.1, 0.2, 0.3, 0.4)'$. An equal number of observations is drawn from each stratum, so that $n_h = n/4$, $h = 1, \ldots, 4$. The corresponding survey weights are $w_i = 4q_{H_i}$, where H_i denotes the stratum, and have sample mean one.

We set $d = 1$ and write $x_i = (1, \tilde{x}_i')'$, where $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})'$. The first non-constant regressor is binary, with $\tilde{x}_{i1} \mid \{H_i = h\} \sim \mathrm{Bernoulli}(\pi_h)$. Under standard stratified sampling, $\pi = (0.5, 0.5, 0.5, 0.5)'$, whereas under exogenous stratification, $\pi = (0.80, 0.65, 0.50, 0.35)'$. In both cases, $\sum_{h=1}^4 q_h\pi_h = 0.5$, so that the population proportion of observations with $\tilde{x}_{i1} = 1$ is the same across the two designs.

The remaining regressors are generated as follows:

$$\tilde{x}_{ij} = \frac{\sqrt{3}}{2}\left(2\Phi(Z_{ij}) - 1\right), \qquad j = 2, \ldots, p, \tag{4.1}$$

where $Z_i = (Z_{i2}, \ldots, Z_{ip})'$ is a mean-zero Gaussian vector with unit marginal variances and $\mathrm{Corr}(Z_{ij}, Z_{ik}) = 0.3^{|j-k|}$. Thus, in the population, $\tilde{x}_{ij} \sim U(-\sqrt{3}/2, \sqrt{3}/2)$, with mean zero and variance $1/4$, equal to the population variance of $\tilde{x}_{i1}$. All penalized regressors therefore have comparable population scales and are uniformly bounded by one.

Under standard stratified sampling, stratum membership is independent of the regressors. Under exogenous stratification, the strata are determined by $\tilde{x}_{i2}$. Specifically, an observation belongs to stratum h if

$$\sum_{\ell=1}^{h-1} q_\ell < \Phi(Z_{i2}) \leq \sum_{\ell=1}^{h} q_\ell, \qquad h = 1, \ldots, 4,$$

where the empty sum is zero. Observations are generated within each stratum by drawing Z_{i2} from the corresponding truncated normal distribution and generating the remaining Gaussian coordinates conditionally according to the specified correlation structure. Thus, stratum membership is a deterministic function of the regressors, as in Example 2.

Conditional on the regressors, the dependent variable is generated according to

$$y_i \,|\, x_i \sim \text{Bernoulli}\left(\Lambda(-0.5 + \tilde{x}_{i1})\right). \tag{4.2}$$

The true parameter is therefore $\theta_0 = (-0.5, 1, \mathbf{0}'_{p-1})'$. The corresponding average marginal effect is $\mathrm{AME}_2 = \Lambda(0.5) - \Lambda(-0.5) \approx 0.2449$. Since the intercept is always included and $\tilde{x}_{i1}$ is the only regressor with a nonzero slope coefficient, any selected model containing $\tilde{x}_{i1}$ also contains the true support, and hence $\theta^*_M = \theta_{0,M}$.

We consider sample sizes $n \in \{200, 400, 800\}$ and vary the number of regressors such that $p/n \in \{0.01, 0.025, 0.05, 0.10, 0.25\}$.[3] The empirical size of the tests is evaluated at the 5% significance level for the null hypotheses

$$H_0 : \theta_{(2)} = 1, \qquad H_0 : \mathrm{AME}_2 = \Lambda(0.5) - \Lambda(-0.5), \tag{4.3}$$

using 3,000 simulation replications for each combination of n, p, and sampling design.

[3]Sequences with p/n fixed at a positive value do not satisfy the growth condition of Proposition 3.2.

Tuning parameters. Proposition 3.2 is stated for a deterministic tuning sequence. To separate tuning parameter selection from inference, we calibrate the tuning parameter using independent auxiliary samples. For each (n, p, design) combination, we generate 49 independent auxiliary samples. On each auxiliary sample, we compute (i) the 10-fold cross-validation (CV) penalty minimizing out-of-fold binomial deviance and (ii) the bootstrapping-after-cross-validation (BCV) penalty of Chetverikov and Sørensen (2025). The medians of the 49 resulting CV and BCV penalties, denoted by λ_{CV} and λ_{BCV}, respectively, are held fixed throughout the 3,000 Monte Carlo replications for the corresponding combination. We refer to these procedures as SI–fixed CV and SI–fixed BCV.

For comparison, we also consider SI–adaptive CV and SI–adaptive BCV, where the tuning parameter is selected separately in each Monte Carlo replication using the same sample on which inference is conducted. For each simulated sample, we select the penalty $\hat{\lambda}$ using either 10-fold CV or BCV, estimate the corresponding ℓ_1-PML model, and apply the proposed SI procedure to the selected model. Adaptive CV and BCV use the same random 10-fold partition within each replication. These procedures reflect the common practice of selecting the tuning parameter from the observed data. Although the data-dependent tuning step is not covered by Proposition 3.2, we include these procedures to examine their finite-sample performance.

For BCV, the score residuals incorporate the survey weights, and the multiplier bootstrap uses 500 draws with $c_0 = 1.1$ and $\alpha_{\mathrm{BCV}} = 0.1/\log\{\max(p, n)\}$. Because all non-constant regressors have comparable population variances, no sample standardization is used in the simulations.

PoSI1. For comparison, we implement the restricted PoSI procedure of Berk et al. (2013), denoted PoSI1. The prespecified model family is

$$\mathcal{M}^{\mathrm{PoSI1}} = \{M \subseteq \{1, \ldots, p+1\} : 1 \in M,\ 1 \leq |M| - 1 \leq 2\}, \tag{4.4}$$

which contains all models with an intercept and either one or two non-constant regressors. For the target regressor $\tilde{x}_{i1}$, corresponding to the second coordinate of x_i, the calibration is restricted to $\mathcal{M}_{2.}^{\mathrm{PoSI1}} = \{M \in \mathcal{M}^{\mathrm{PoSI1}} : 2 \in M\}$.

The PoSI1 critical value is computed from an independent pilot sample using $B = 999$ stratified bootstrap replications, once for each (n, p, design) combination, and is then held fixed across the Monte Carlo replications. The standard survey-weighted t-test from the full model, t_{svy}, is also reported as a benchmark that ignores the selection step.

Results. Tables 1 and 2 report the empirical sizes of the tests under standard stratified sampling and exogenous stratification, respectively. The results under both sampling schemes are qualitatively similar. The SI procedures show reasonable size control for both hypotheses across different sample sizes and numbers of regressors. In particular, SI–fixed CV and SI–fixed BCV yield rejection frequencies generally close to the nominal 5% level.

The adaptive procedures also perform well, with results comparable to those obtained with fixed tuning. Some overrejection is observed at $n = 200$, but the empirical sizes are generally close to the nominal level for $n = 400$ and $n = 800$.

PoSI1 tends to underreject, particularly as p/n increases. In contrast, the rejection rate of the standard survey-weighted t-test, denoted by t_{svy}, deteriorates as p/n grows. At $p/n = 0.25$ and $n = 800$, its coefficient rejection frequencies reach 58.9% and 54.3% under standard and exogenous stratification, respectively. The size distortions for the AME test are less pronounced. Finally, BCV selects the regressor of interest less frequently in smaller samples, while at $n = 800$ it is selected in nearly every replication under all four SI procedures.

5 Empirical application

We apply the proposed methods to the corrected 2020 Canadian Internet Use Survey (CIUS) Public Use Microdata File (PUMF) to examine the demographic and socioeconomic correlates of accessing an online account for a government program or service.[4]

Data. The outcome is the response to "During the past 12 months, did you access an account for a government program or service on the Internet?" We code reported government-account access as one, reported nonaccess as zero and assign zero to the 2,115 *Valid skip* cases, corresponding to respondents who reported no personal Internet use in the preceding three months (see footnote 6 for a sensitivity analysis excluding these observations). We exclude 378 *Not stated* outcome responses, consistent with the nonrejection of the Hausman–McFadden-type independence of irrelevant alternatives (IIA) diagnostic reported below, leaving $n =$ 17,031 observations, of which 5,939 report access.

The specification contains 46 binary regressors for age, urban or rural residence, gender, Indigenous identity, knowledge of official languages, employment, education, visible-minority

[4]Statistics Canada released the 2020 CIUS PUMF on August 4, 2022 (catalogue no. 56M0003X) and corrected it on December 12, 2024, restoring omitted variables and incomplete bootstrap replicate weights. The corrections permit a more granular covariate specification and design-based inference using the complete replicate weights. See `https://www150.statcan.gc.ca/n1/pub/56m0003x/56m0003x2020001-eng.htm`.

Table 1: Empirical rejection frequencies (%) of the tests for $H_0 : \theta_{(2)} = 1$ and $H_0 : \mathrm{AME}_2 = 0.2449$ at the 5% level. Standard stratified sampling, bounded-continuous design.

	$H_0 : \theta_{(2)} = 1$					$H_0 : \mathrm{AME}_2 = 0.2449$				
	p/n					p/n				
	.01	.025	.05	.10	.25	.01	.025	.05	.10	.25
$n = 200$										
SI–fixed CV	4.8	5.3	5.4	5.4	6.1	4.5	4.9	5.0	4.8	5.5
SI–fixed BCV	4.9	5.4	5.4	4.7	5.0	5.1	5.4	5.4	4.4	5.4
SI–adaptive CV	4.4	4.8	5.6	6.7	7.4	4.3	4.5	5.3	6.0	6.6
SI–adaptive BCV	4.3	5.9	6.6	6.9	6.8	4.3	5.9	6.4	6.8	6.6
PoSI1	3.0	3.0	1.9	2.0	1.0	3.6	3.4	2.3	2.2	1.4
t_{svy}	4.3	4.9	6.1	8.2	25.1	4.9	5.2	6.2	7.0	15.0
n_{sel} fixed CV	2988	2968	2919	2876	2764	2988	2968	2919	2876	2764
n_{sel} fixed BCV	1006	885	811	767	483	1006	885	811	767	483
n_{sel} adaptive CV	2905	2814	2702	2616	2397	2905	2814	2702	2616	2397
n_{sel} adaptive BCV	1028	973	834	765	575	1028	973	834	765	575
$n = 400$										
SI–fixed CV	5.4	5.4	5.6	5.0	5.3	5.1	5.2	5.2	4.7	5.1
SI–fixed BCV	5.1	4.8	5.7	5.6	5.8	4.9	4.9	5.7	5.6	5.7
SI–adaptive CV	5.5	5.0	5.4	4.9	5.0	5.2	4.3	4.8	4.2	4.7
SI–adaptive BCV	5.7	5.0	5.5	5.8	6.5	5.4	4.9	5.2	5.7	6.5
PoSI1	3.9	4.7	2.3	2.2	2.4	4.0	4.9	2.6	2.6	2.7
t_{svy}	5.2	5.7	6.9	11.6	37.4	5.5	5.8	6.2	7.6	14.1
n_{sel} fixed CV	2999	3000	2994	2990	2977	2999	3000	2994	2990	2977
n_{sel} fixed BCV	2373	2327	2274	2107	1802	2373	2327	2274	2107	1802
n_{sel} adaptive CV	2994	2977	2973	2951	2911	2994	2977	2973	2951	2911
n_{sel} adaptive BCV	2325	2278	2183	2050	1780	2325	2278	2183	2050	1780
$n = 800$										
SI–fixed CV	5.4	5.4	5.4	5.1	5.2	5.3	4.9	4.9	4.6	4.9
SI–fixed BCV	5.9	5.5	4.7	5.1	5.5	5.3	5.0	4.1	4.7	5.1
SI–adaptive CV	5.6	5.4	5.7	4.9	5.9	5.2	5.2	5.0	4.5	5.4
SI–adaptive BCV	5.6	5.5	4.7	5.0	5.5	5.1	5.0	4.2	4.6	5.0
PoSI1	3.4	4.5	2.6	2.9	2.7	3.3	4.7	2.6	2.8	2.9
t_{svy}	5.5	5.8	7.4	15.4	58.9	5.5	5.5	5.0	6.5	13.6
n_{sel} fixed CV	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000
n_{sel} fixed BCV	2986	2984	2977	2972	2929	2986	2984	2977	2972	2929
n_{sel} adaptive CV	3000	3000	3000	3000	2999	3000	3000	3000	3000	2999
n_{sel} adaptive BCV	2975	2976	2971	2960	2913	2975	2976	2971	2960	2913

Notes: 3,000 Monte Carlo replications. For the fixed rows, the CV and BCV penalties are the medians from 49 independent auxiliary samples and are held fixed during inference. For the adaptive rows, $\hat{\lambda}$ is chosen on the same inference sample; adaptive CV and BCV use the same 10-fold split within a replication. These adaptive rows are not covered by Proposition 3.2. BCV uses 500 multiplier-bootstrap draws. PoSI1 uses $B = 999$ stratified bootstrap draws. n_{sel} is the number of replications in which the target regressor $\tilde{x}_{i1}$ is selected; SI rejection frequencies are conditional on target selection.

Table 2: Empirical rejection frequencies (%) of the tests for $H_0 : \theta_{(2)} = 1$ and $H_0 : \text{AME}_2 = 0.2449$ at the 5% level. Exogenous stratification, bounded-continuous design.

	$H_0 : \theta_{(2)} = 1$					$H_0 : \text{AME}_2 = 0.2449$				
	p/n					p/n				
	.01	.025	.05	.10	.25	.01	.025	.05	.10	.25
$n = 200$										
SI–fixed CV	5.6	6.3	6.0	5.3	4.7	5.4	5.9	5.6	4.9	4.3
SI–fixed BCV	5.9	4.8	6.0	6.8	5.5	5.3	4.7	6.0	6.5	5.3
SI–adaptive CV	5.3	6.2	7.2	5.5	7.0	5.4	6.1	6.8	5.1	6.3
SI–adaptive BCV	6.6	4.9	7.9	6.4	6.3	6.4	4.9	7.8	6.2	6.3
PoSI1	3.3	2.4	2.5	2.3	1.2	3.8	3.0	3.2	2.4	1.4
t_{svy}	5.4	5.5	6.9	7.6	23.2	6.0	6.0	7.5	7.0	13.9
n_{sel} fixed CV	2981	2970	2882	2846	2759	2981	2970	2882	2846	2759
n_{sel} fixed BCV	1313	1110	923	825	600	1313	1110	923	825	600
n_{sel} adaptive CV	2904	2806	2726	2588	2394	2904	2806	2726	2588	2394
n_{sel} adaptive BCV	1242	1108	963	801	646	1242	1108	963	801	646
$n = 400$										
SI–fixed CV	6.0	5.2	4.4	5.9	5.7	5.7	4.8	4.0	5.4	5.3
SI–fixed BCV	6.4	5.3	5.2	5.7	6.0	6.0	5.3	4.8	5.4	5.9
SI–adaptive CV	6.0	5.2	5.3	5.7	5.1	5.8	4.6	4.7	5.0	4.9
SI–adaptive BCV	6.6	5.2	5.1	4.8	6.3	6.2	5.0	4.9	4.8	6.4
PoSI1	3.6	2.2	1.8	2.2	1.3	3.8	2.5	2.0	2.2	1.6
t_{svy}	5.9	5.2	6.5	10.4	35.8	6.0	5.4	6.2	7.3	14.1
n_{sel} fixed CV	2999	2999	2997	2993	2972	2999	2999	2997	2993	2972
n_{sel} fixed BCV	2544	2450	2295	2244	1876	2544	2450	2295	2244	1876
n_{sel} adaptive CV	2996	2987	2970	2947	2910	2996	2987	2970	2947	2910
n_{sel} adaptive BCV	2443	2388	2269	2124	1857	2443	2388	2269	2124	1857
$n = 800$										
SI–fixed CV	5.1	4.8	5.7	6.2	5.6	4.8	4.5	5.3	5.7	5.3
SI–fixed BCV	5.4	5.2	5.3	5.6	5.9	4.7	4.8	5.0	5.3	5.3
SI–adaptive CV	5.0	5.7	5.3	5.3	6.0	4.7	5.2	5.0	4.7	5.6
SI–adaptive BCV	5.3	5.4	5.3	5.6	5.8	4.7	4.9	5.0	5.2	5.3
PoSI1	2.6	3.0	2.5	2.9	2.0	2.7	3.1	2.6	2.9	2.2
t_{svy}	5.0	5.6	7.5	14.5	54.3	5.0	5.6	5.4	6.8	13.3
n_{sel} fixed CV	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000
n_{sel} fixed BCV	2990	2985	2980	2964	2944	2990	2985	2980	2964	2944
n_{sel} adaptive CV	3000	3000	2999	3000	2998	3000	3000	2999	3000	2998
n_{sel} adaptive BCV	2985	2983	2975	2956	2917	2985	2983	2975	2956	2917

Notes: see Table 1.

status, household composition, family-income quintile, immigration status, disability, self-rated health, and province. Available *Not stated* covariate indicators are retained. Reference categories are age 45–54, large urban centre, male, non-Indigenous, neither official language, not employed, some post-secondary education, not a visible minority, a family with children under 18, income quintile 2, landed immigrant, no disability, good health, and Alberta. The rural/small-town indicator also includes the separately classified Prince Edward Island cases; a provincial PEI indicator enters independently.

The CIUS uses stratified probability sampling, but confidential strata identifiers are unavailable in the PUMF. We use the person-level survey weight and all 1,000 official bootstrap replicate weights for design-based covariance estimation.[5]

To assess the sensitivity of our results to excluding the 378 *Not stated* outcome responses, we implement a Hausman–McFadden-type IIA diagnostic using a survey-design analogue of `Stata`'s `suest` procedure. The test compares the *Yes*-versus-*No* coefficients from the three-category multinomial logit and the restricted binary logit, accounting for their joint covariance through first-order linearization using the CIUS bootstrap replicate weights. Among respondents with a *Not stated* outcome, none identified as Indigenous or reported poor health, resulting in quasi-separation and a nearly singular Hessian. We therefore exclude the two Indigenous-identity and five self-rated-health indicators in the primary test. The resulting 39-regressor specification does not reject coefficient equality, with a p-value of 0.488. A supplementary calculation using all 46 regressors and a Moore–Penrose generalized inverse of the multinomial Hessian in constructing the sandwich covariance estimator also yields nonrejection, with a p-value of 0.739.

Estimation and inference. We fit the survey-weighted logit model (2.5) with `glmnet`, standardizing regressors by survey-weighted means and standard deviations. The penalty parameter is selected using the survey-weighted adaptation of the BCV procedure described in Section 4, with ten-fold survey-weighted deviance cross-validation and 1,000 Gaussian multiplier draws. The procedure yields $\hat{\lambda}_{\mathrm{BCV}} = 0.023238$ and selects eight regressors.

Chetverikov and Sørensen (2025) establish their theoretical guarantees without survey weights, and we do not extend them to the survey-weighted adaptation. Moreover, because $\hat{\lambda}_{\mathrm{BCV}}$ is selected from the same sample, the empirical implementation is not formally covered by Proposition 3.2, which assumes a deterministic tuning parameter. The adaptive BCV results in Section 4 provide a finite-sample assessment of this implementation: their empirical

[5]See the CIUS 2020 Microdata User Guide, Section 10.

sizes are close to the nominal level for $n \geq 400$, well below the sample size used here.

SI inference conditions on the selected model and coefficient signs; under the conditions of Corollary 3.3, the same intervals are also valid conditional on the selected model alone. We estimate the pseudo-score covariance using the CIUS bootstrap replicate weights. For population AMEs, the covariance estimator additionally accounts for variation in the weighted average of individual marginal effects and its covariance with the pseudo-score. The t_{svy} columns report two-sided Wald p-values based on full-model survey estimates and their replicate-weight standard errors.

For comparison, we construct a computationally feasible PoSI1 model family comprising all $2^8 - 1 = 255$ nonempty subsets of the eight selected regressors and 1,000 randomly drawn nonempty models from the full pool of 46 regressors. After removing duplicates, the family contains 1,225 models, each including an intercept. Coefficient and AME critical values are computed separately from maxima over models containing each target regressor. For AMEs, the computation accounts for both the first-order estimation and empirical-averaging terms and their covariance. Of the 1,000 CIUS bootstrap replicates, 998 yield finite coefficient and AME estimates across all candidate models and are used to compute the critical values.

Since exhaustive enumeration of all $2^{46}-1$ nonempty models is computationally prohibitive, we use the partly data-dependent model family described above and report PoSI1 as an empirical benchmark rather than invoke the simultaneous-coverage guarantee for a prespecified model family.

Results. Table 3 presents full-model survey estimates, ℓ_1-PML estimates, and SI results for the candidate covariates. Nine *Not stated* covariate indicators are included in estimation but omitted from the table for brevity; none was selected by BCV. BCV selects Age 25–34, Age 35–44, Age 65+, French only, High school or less, University degree, Visible minority, and Single. BCV selects Age 25–34, Age 35–44, Age 65+, French, High school or less, University degree, Visible minority, and Single.[6] Seven have two-sided SI p-values below 0.05 for both coefficients and AMEs; the visible-minority AME has a p-value of 0.405.

The SI AMEs for Age 25–34, Age 35–44, and Age 65+ are 10.8, 8.9, and −15.9 percentage points, respectively. The university-degree AME is 8.2 percentage points, whereas the French-

[6] As a sensitivity check, we exclude the 2,115 *Valid skip* observations, reducing the sample to $n = 14{,}916$. Reapplying BCV selects seven of the eight regressors selected in the main analysis; Single is no longer selected. The SI AMEs of the seven retained regressors preserve their signs, with six significant at the 5% level. Visible minority remains insignificant (p-value 0.097). All seven variable-specific PoSI1 AME intervals exclude zero, whereas intervals based on a common critical value exclude zero for six.

only, high-school-or-less, and single indicators have AMEs of -11.8, -9.6, and -8.2 percentage points, respectively. For indicators within a categorical block, the AME changes one dummy while holding the others fixed, so it need not equal a contrast between two feasible category levels. The SI estimates target selected-model pseudo-true coefficients and their corresponding population AMEs. A causal interpretation of these targets would require additional identifying assumptions.

Table 4 compares SI and variable-specific PoSI1 intervals. All eight PoSI1 AME intervals exclude zero, although the lower endpoint for visible-minority status is just 0.00154; its SI interval includes zero. Taking the maximum additionally over all eight selected regressors yields common bootstrap critical values of 3.505 for coefficients and 3.516 for AMEs. Using these common critical values, seven AME intervals exclude zero, whereas the visible-minority interval $[-0.0107, 0.0896]$ includes zero, matching the SI significance classifications.

6 Conclusion

This paper has provided two main results. First, we have extended the ℓ_1-PML SI method to GLMs with survey weights and established its asymptotic validity for pseudo-true selected-model parameters under an i.n.i.d. triangular-array framework when the number of candidate regressors grows with the sample size. The validity result does not require sure screening: selected models may omit active full-model regressors, in which case the target is the corresponding restricted population minimizer. Second, we have established selective inference for nonlinear parameter functionals, including average marginal effects. Together, these results provide a framework for valid post-selection inference on both coefficients and economically interpretable nonlinear effects in survey-weighted GLMs.

Table 3: Logit estimates and average marginal effects for government online service access.

	Logit coefficients					Average marginal effects			
	Estimates			p-values		Estimates		p-values	
Variable	GLM	ℓ_1-PML	SI	t_{svy}	SI	GLM	SI	t_{svy}	SI
Intercept	−0.24	−0.32	−0.36	0.64	0.002	—	—	—	—
Rural/small town	−0.09	0.00	–	0.17	–	−0.018	–	0.164	–
Age (ref.: 45–54)									
Age 15–24	0.21	0.00	–	0.06	–	0.046	–	0.066	–
Age 25–34	0.48	0.24	0.47	< 0.001	0.007	0.106	0.108	< 0.001	0.008
Age 35–44	0.38	0.15	0.39	< 0.001	0.005	0.083	0.089	< 0.001	0.005
Age 55–64	−0.10	0.00	–	0.21	–	−0.021	–	0.204	–
Age 65+	−0.73	−0.65	−0.74	< 0.001	< 0.001	−0.154	−0.159	< 0.001	< 0.001
Female	−0.04	0.00	–	0.39	–	−0.009	–	0.386	–
Indigenous identity	0.06	0.00	–	0.64	–	0.014	–	0.643	–
English only	0.08	0.00	–	0.88	–	0.017	–	0.874	–
French only	−0.33	−0.23	−0.54	0.50	< 0.001	−0.070	−0.118	0.490	< 0.001
English and French	0.20	0.00	–	0.69	–	0.042	–	0.692	–
Employed	0.05	0.00	–	0.44	–	0.011	–	0.443	–
High school or less	−0.55	−0.32	−0.43	< 0.001	< 0.001	−0.119	−0.096	< 0.001	< 0.001
University degree	0.35	0.29	0.36	< 0.001	< 0.001	0.077	0.082	< 0.001	< 0.001
Visible minority	0.12	0.04	0.18	0.14	0.40	0.025	0.040	0.142	0.405
Family w/o 18	−0.05	0.00	–	0.43	–	−0.012	–	0.435	–
Single	−0.46	−0.12	−0.37	< 0.001	< 0.001	−0.095	−0.082	< 0.001	< 0.001
Other household type	−0.09	0.00	–	0.53	–	−0.020	–	0.527	–
Income (ref.: Q2, $44,120–$75,321)									
Q1: $44,119 or less	−0.03	0.00	–	0.73	–	−0.006	–	0.734	–
Q3: $75,322–$109,431	−0.07	0.00	–	0.35	–	−0.015	–	0.346	–
Q4: $109,432–$162,799	−0.05	0.00	–	0.50	–	−0.011	–	0.499	–
Q5: $162,800 or more	−0.23	0.00	–	0.008	–	−0.048	–	0.007	–
Non-landed immigrant	−0.01	0.00	–	0.89	–	−0.003	–	0.886	–
Disabled	−0.03	0.00	–	0.81	–	−0.005	–	0.812	–
Health: excellent	0.06	0.00	–	0.47	–	0.012	–	0.467	–
Health: very good	0.06	0.00	–	0.34	–	0.013	–	0.341	–
Health: fair	0.09	0.00	–	0.42	–	0.018	–	0.426	–
Health: poor	0.23	0.00	–	0.17	–	0.051	–	0.173	–
Province (ref.: Alberta)									
NL	0.19	0.00	–	0.10	–	0.042	–	0.105	–
PEI	0.07	0.00	–	0.55	–	0.016	–	0.549	–
NS	−0.17	0.00	–	0.13	–	−0.036	–	0.122	–
NB	−0.25	0.00	–	0.03	–	−0.054	–	0.023	–
QC	−0.23	0.00	–	0.04	–	−0.050	–	0.037	–
ON	−0.13	0.00	–	0.14	–	−0.028	–	0.135	–
MB	−0.20	0.00	–	0.10	–	−0.041	–	0.099	–
SK	0.02	0.00	–	0.88	–	0.004	–	0.880	–
BC	0.09	0.00	–	0.37	–	0.020	–	0.370	–

Notes: $n = 17{,}031$. GLM and t_{svy} denote full-model survey estimates and t/Wald p-values. ℓ_1-PML is the BCV-penalized estimator; SI reports selection-conditional one-step estimates and two-sided p-values. BCV selects eight of 46 regressors; nine "not stated" indicators enter the fitted model but are omitted from the display. AME inference incorporates estimation and empirical-averaging uncertainty. "−" denotes nonselection; — denotes an inapplicable intercept AME. The location indicator also includes PEI; province indicators enter separately. For categorical regressors, AMEs change one dummy at a time, holding others fixed.

Table 4: SI and PoSI1 confidence intervals for logit coefficients and average marginal effects, government online service access.

	SI		PoSI1		
Variable	1-step est.	95% CI	Estimate	s.e.	95% CI
Panel A: Logit coefficients					
Intercept	−0.36	[−0.45, −0.16]	—	—	—
Age 25–34	0.47	[0.16, 0.61]	0.48	0.07	[0.28, 0.67]
Age 35–44	0.39	[0.15, 0.51]	0.39	0.06	[0.22, 0.57]
Age 65+	−0.74	[−0.98, −0.64]	−0.76	0.05	[−0.92, −0.60]
French only	−0.54	[−0.67, −0.41]	−0.59	0.08	[−0.83, −0.35]
High school or less	−0.43	[−0.55, −0.29]	−0.44	0.06	[−0.61, −0.28]
University degree	0.36	[0.25, 0.47]	0.36	0.06	[0.21, 0.51]
Visible minority	0.18	[−0.24, 0.29]	0.18	0.06	[0.01, 0.35]
Single	−0.37	[−0.46, −0.27]	−0.40	0.05	[−0.54, −0.25]
Panel B: Average marginal effects					
Age 25–34	0.108	[0.036, 0.141]	0.107	0.017	[0.062, 0.152]
Age 35–44	0.089	[0.034, 0.117]	0.088	0.014	[0.048, 0.128]
Age 65+	−0.159	[−0.214, −0.138]	−0.160	0.011	[−0.193, −0.128]
French only	−0.118	[−0.146, −0.089]	−0.122	0.015	[−0.169, −0.075]
High school or less	−0.096	[−0.122, −0.065]	−0.097	0.013	[−0.132, −0.061]
University degree	0.082	[0.057, 0.108]	0.081	0.013	[0.047, 0.115]
Visible minority	0.040	[−0.056, 0.067]	0.039	0.014	[0.002, 0.077]
Single	−0.082	[−0.104, −0.060]	−0.084	0.011	[−0.113, −0.055]

Notes: The 95% SI intervals invert the truncated-Gaussian pivot conditional on the selected model and coefficient signs. Population-AME inference includes the empirical weighted-average term and its covariance with the pseudo-score. PoSI1 estimates use the unpenalized survey-weighted selected-model fit and target-specific bootstrap critical values over the 1,225-model empirical candidate family. In row order, the coefficient critical values are $(2.658, 2.810, 2.964, 3.087, 2.626, 2.680, 2.645, 2.783)$; the AME-specific values are $(2.679, 2.812, 2.931, 3.103, 2.652, 2.713, 2.655, 2.761)$. The bootstrap maxima use 998 common finite replicates. The observed model family partly depends on selection, so these PoSI1 intervals are empirical benchmarks rather than claims of formal simultaneous coverage. The intercept is excluded from PoSI1.

References

Ali, A. and Tibshirani, R. J. (2019), 'The Generalized Lasso Problem and Uniqueness', *Electronic Journal of Statistics* **13**(2), 2307 – 2347.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'Inference on Treatment Effects after Selection among High-Dimensional Controls', *Review of Economic Studies* **81**(2), 608–650.

Belloni, A., Chernozhukov, V. and Wei, Y. (2016), 'Post-Selection Inference for Generalized

Linear Models with Many Controls', *Journal of Business & Economic Statistics* **34**(4), 606–619.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013), 'Valid Post-Selection Inference', *The Annals of Statistics* **41**(2), 802–837.

Cameron, A. C. and Trivedi, P. K. (2009), *Microeconometrics: Methods and Evaluations*, Cambridge University Press.

Cattaneo, M. D., Masini, R. P. and Underwood, W. G. (2025), 'Yurinskii's Coupling for Martingales', *The Annals of Statistics* **53**(5), 2179–2203.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017), 'Central Limit Theorems and Bootstrap in High Dimensions', *The Annals of Probability* **45**(4), 2309–2352.

Chetverikov, D. and Sørensen, J. R.-V. (2025), 'Selecting Penalty Parameters of High-Dimensional M-Estimators Using Bootstrapping after Cross Validation', *Journal of Political Economy* **133**(10), 3208–3248.

Chiang, H. D. (2022), 'Many Average Partial Effects: with An Application to Text Regression', *arXiv preprint arXiv:1812.09397* .

Eaton, M. L. and Tyler, D. E. (1991), 'On Wielandt's Inequality and its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix', *The Annals of Statistics* **19**(1), 260–271.

Fuller, W. A. (2011), *Sampling Statistics*, John Wiley & Sons.

Hirshberg, D. A. and Wager, S. (2020), 'Debiased Inference of Average Partial Effects in Single-Index Models: Comment on Wooldridge and Zhu', *Journal of Business & Economic Statistics* **38**(1), 19–24.

Javanmard, A. and Montanari, A. (2014), 'Confidence Intervals and Hypothesis Testing for High-Dimensional Regression', *The Journal of Machine Learning Research* **15**(1), 2869–2909.

Kuchibhotla., A. K., Kolassa., J. E. and Kuffner., T. A. (2022), 'Post-Selection Inference', *Annual Review of Statistics and Its Application* **9**(1), 505–527.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016), 'Exact Post-Selection Inference, with Application to the Lasso', *The Annals of Statistics* **44**(3), 907–927.

Manski, C. F. and Lerman, S. R. (1977), 'The Estimation of Choice Probabilities from Choice Based Samples', *Econometrica* **45**(8), 1977–1988.

McConville, K. S., Breidt, F. J., Lee, T. and Moisen, G. G. (2017), 'Model-Assisted Survey Regression Estimation with the Lasso', *Journal of Survey Statistics and Methodology* **5**(2), 131–158.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers', *Statistical Science* **27**(4), 538 – 557.

Newey, W. K. and McFadden, D. (1994), Large Sample Estimation and Hypothesis Testing, *in* R. F. Engle and D. L. McFadden, eds, 'Handbook of Econometrics, Volume 4', Amsterdam, chapter 36, pp. 2111–2245.

Taylor, J. and Tibshirani, R. (2018), 'Post-Selection Inference for l1-Penalized Likelihood Models', *Canadian Journal of Statistics* **46**(1), 41–61.

Thompson, S. K. (2012), *Sampling*, Vol. 755, John Wiley & Sons.

Tian, X. and Taylor, J. (2017), 'Asymptotics of Selective Inference', *Scandinavian Journal of Statistics* **44**(2), 480–499.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models', *The Annals of Statistics* **42**(3), 1166 – 1202.

Vershynin, R. (2010), 'Introduction to the Non-Asymptotic Analysis of Random Matrices', *arXiv preprint arXiv:1011.3027* .

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47, Cambridge University Press.

Wooldridge, J. M. (2001), 'Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples', *Econometric Theory* **17**(2), 451–470.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT press.

Wooldridge, J. M. and Zhu, Y. (2020), 'Inference in Approximately Sparse Correlated Random Effects Probit Models with Panel Data', *Journal of Business & Economic Statistics* **38**(1), 1–18.

Xia, L., Nan, B. and Li, Y. (2023), 'Debiased Lasso for Generalized Linear Models with a Diverging Number of Covariates', *Biometrics* **79**(1), 344–357.

Zhang, C.-H. and Zhang, S. S. (2014), 'Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

A Proofs

A.1 Proof of Lemma 3.4

Let $k \equiv d + p$ and define $A_{i,M} \equiv (w_i \ddot{g}(y_i, x_{iM}'\theta_M^*))^{1/2} x_i \in \mathbb{R}^k$. Convexity of $g(y,\cdot)$ implies that $\ddot{g}(y,t) \geq 0$. The upper bound C_u on the survey weights (Assumption 1(a)) and the bound $|\ddot{g}(y, x_{iM}'\theta_M^*)| \leq C_g$ (Assumption 1(b)) together give, for every $u \in \mathbb{R}^k$ with $\|u\|_2 = 1$, $|u'A_{i,M}| \leq (C_u C_g)^{1/2} |u'x_i|$ a.s. Consequently, $\|u'A_{i,M}\|_{\psi_2} \leq (C_u C_g)^{1/2} \|u'x_i\|_{\psi_2} \leq K$, where K is independent of i, M, and n, using $\|x_i\|_{\psi_2} \leq C_u$ from Assumption 1(a). Thus, the independent vectors $A_{i,M}$ are uniformly sub-Gaussian with common norm bound K. Recall that

$$\hat{H}(\theta_M^*) = n^{-1} \sum_{i=1}^n A_{i,M} A_{i,M}', \qquad H(\theta_M^*) = n^{-1} \sum_{i=1}^n \mathrm{E}[A_{i,M} A_{i,M}'],$$

and let

$$B_M \equiv \hat{H}(\theta_M^*) - H(\theta_M^*) = n^{-1} \sum_{i=1}^n \left(A_{i,M} A_{i,M}' - \mathrm{E}[A_{i,M} A_{i,M}'] \right).$$

Step 1: Covering-net reduction. Let $\mathcal{N}$ be a 1/4-net of the unit sphere $\mathbb{S}^{k-1}$. By Corollary 4.2.13 of Vershynin (2018), $\mathcal{N}$ may be chosen with cardinality $|\mathcal{N}| \leq 9^k$. Since B_M is symmetric, Lemma 5.4 of Vershynin (2010) gives

$$\|B_M\|_2 \leq 2 \max_{u \in \mathcal{N}} |u' B_M u|. \tag{A.1}$$

For fixed $u \in \mathcal{N}$, set $Z_{i,u} = u'A_{i,M}$, so that

$$u' B_M u = n^{-1} \sum_{i=1}^n \left(Z_{i,u}^2 - \mathrm{E}[Z_{i,u}^2] \right).$$

Step 2: Sub-exponential norm of the summands. The summands $\{Z_{i,u}^2 - \mathrm{E}[Z_{i,u}^2]\}_{i=1}^n$ are independent and centered by construction. Since $Z_{i,u}$ is sub-Gaussian with $\|Z_{i,u}\|_{\psi_2} \leq K$, Lemma 5.14 of Vershynin (2010) (equivalently, Lemma 2.7.6 of Vershynin (2018)) gives

$$\|Z_{i,u}^2\|_{\psi_1} \leq 2 \|Z_{i,u}\|_{\psi_2}^2 \leq 2K^2,$$

and hence, by the triangle inequality for the sub-exponential norm,

$$\|Z_{i,u}^2 - \mathrm{E}[Z_{i,u}^2]\|_{\psi_1} \leq 2 \|Z_{i,u}^2\|_{\psi_1} \leq 4K^2 \equiv CK^2. \tag{A.2}$$

Step 3: Bernstein's inequality. By (A.2), the summands are independent, centered, and uniformly sub-exponential with norm bound CK^2. Bernstein's inequality (Corollary 5.17 of Vershynin (2010), or Corollary 2.8.3 of Vershynin (2018)) therefore gives an absolute constant $c_1 > 0$, not depending on u, i, M, or n, such that for every $v > 0$

$$P[|u' B_M u| > v] \le 2 \exp\left\{ -c_1 n \min\left(\frac{v^2}{K^4}, \frac{v}{K^2} \right) \right\}. \tag{A.3}$$

Step 4: Union bound. Combining (A.1) and (A.3) with a union bound over the $|\mathcal{N}| \le 9^k$ points of the net gives constants $c_0, C_0 > 0$, depending only on K, such that for every $t \ge 0$,

$$P\left[\|B_M\|_2 > C_0 \left(\sqrt{\frac{k+t}{n}} + \frac{k+t}{n} \right) \right] \le 2 \exp(-c_0 t). \tag{A.4}$$

Since d is fixed, $k = d + p$, and $p/n \to 0$,

$$\|\hat{H}(\theta_M^*) - H(\theta_M^*)\|_2 = O_p\left(\sqrt{\frac{p}{n}} \right).$$

Design bound. Steps 1–4 applied with $A_{i,M}$ replaced by x_i (for which $\|x_i\|_{\psi_2} \le C_u$ by Assumption 1(a) plays the role of K) give, by the identical argument,

$$\left\| n^{-1} X'X - \mathrm{E}[n^{-1} X'X] \right\|_2 = O_p\left(\sqrt{\frac{p}{n}} \right).$$

Moreover, since the sub-Gaussian norm controls the second moment up to an absolute constant (Lemma 5.5 of Vershynin (2010)),

$$\left\| \mathrm{E}[n^{-1} X'X] \right\|_2 = \sup_{\|u\|_2 = 1} n^{-1} \sum_{i=1}^{n} \mathrm{E}[(u' x_i)^2] \le C \sup_{\substack{1 \le i \le n \\ \|u\|_2 = 1}} \|u' x_i\|_{\psi_2}^2 \le C.$$

It follows that

$$\|n^{-1} X'X\|_2 = O_p(1),$$

which proves the lemma. □

A.2 A rate for the restricted estimator

Lemma A.1 (Restricted model rate)**.** *Suppose that Assumption 1 holds, and let $M = M_n \in \mathcal{M}_n(m_n)$ be deterministic. If $p/n \to 0$ and $m_n(n^{-1/2} + \lambda) \to 0$, then*

$$\|\hat{\theta}_M - \theta^*_M\|_2 = O_p\left(\sqrt{\frac{|M|}{n}} + \sqrt{|M| - d}\,\lambda\right), \tag{A.5}$$

$$\|\hat{\theta}_M - \theta^*_M\|_1 = O_p\left(\frac{|M|}{\sqrt{n}} + \sqrt{|M|(|M| - d)}\,\lambda\right). \tag{A.6}$$

Under the rate conditions of Proposition 3.2, these bounds reduce to

$$\|\hat{\theta}_M - \theta^*_M\|_2 = O_p(\sqrt{m_n}\lambda), \qquad \|\hat{\theta}_M - \theta^*_M\|_1 = O_p(m_n\lambda). \tag{A.7}$$

Proof. By the population first-order condition, $\mathrm{E}[S_M(\theta^*_M)] = 0$. Write

$$S_M(\theta^*_M) = n^{-1}\sum_{i=1}^n \psi_{i,M}, \qquad \psi_{i,M} \equiv -w_i x_{iM}\dot{g}(y_i, x'_{iM}\theta^*_M), \qquad \mu_{i,M} \equiv \mathrm{E}[\psi_{i,M}].$$

The individual means need not vanish, but $\sum_i \mu_{i,M} = 0$. By independence, $\mathrm{E}[\psi'_{i,M}\psi_{k,M}] = \mu'_{i,M}\mu_{k,M}$ for $i \neq k$, and

$$\sum_{i\neq k} \mu'_{i,M}\mu_{k,M} = \left\|\sum_{i=1}^n \mu_{i,M}\right\|_2^2 - \sum_{i=1}^n \|\mu_{i,M}\|_2^2 = -\sum_{i=1}^n \|\mu_{i,M}\|_2^2 \leq 0.$$

Hence, by the uniform moment bounds, for some constant $\bar{C}$,

$$\mathrm{E}\,\|S_M(\theta^*_M)\|_2^2 \leq n^{-2}\sum_{i=1}^n \mathrm{E}\left[w_i^2\|x_{iM}\|_2^2\,\dot{g}(y_i, x'_{iM}\theta^*_M)^2\right] \leq \bar{C}\frac{|M|}{n}.$$

Thus,

$$\|S_M(\theta^*_M)\|_2 = O_p\left(\sqrt{\frac{|M|}{n}}\right). \tag{A.8}$$

Lemma 3.4 and Assumption 1(d) imply that, wpa 1,

$$\|\hat{H}_M(\theta^*_M) - H_M\|_2 = o_p(1), \qquad \lambda_{\min}\left(\hat{H}_M(\theta^*_M)\right) \geq c_H/2. \tag{A.9}$$

Let $r_{n,M} \equiv A_0\left(\sqrt{\frac{|M|}{n}} + \sqrt{|M| - d}\,\lambda\right)$, where $A_0 > 0$ is a constant. For some constant C,

uniformly over $\|\delta\|_2 \le r_{n,M}$

$$\max_i |x'_{iM}\delta| \le C_u\|\delta\|_1 \le C_u\sqrt{|M|}\, r_{n,M} \le CA_0 m_n(n^{-1/2}+\lambda) = o(1).$$

Thus, this ball is contained in the neighborhood specified in Assumption 1(b). On this ball, the local Lipschitz condition and Lemma 3.4 imply

$$\sup_{\|\delta\|_2 \le r_{n,M}} \|\hat{H}_M(\theta^*_M+\delta) - \hat{H}_M(\theta^*_M)\|_2 \le C\sqrt{|M|}\, r_{n,M}\|n^{-1}X'_M X_M\|_2 = o_p(1). \tag{A.10}$$

It follows from (A.9), (A.10), and Weyl's inequality (Eaton and Tyler, 1991, Lemma 2.1) that wpa 1

$$\inf_{\|\delta\|_2 \le r_{n,M}} \lambda_{\min}\left(\hat{H}_M(\theta^*_M+\delta)\right) \ge c_H/4. \tag{A.11}$$

Write $\delta = (\delta'_\alpha, \delta'_\beta)'$, where $\delta_\alpha \in \mathbb{R}^d$ and $\delta_\beta \in \mathbb{R}^{|M|-d}$ correspond to the unpenalized and penalized components, respectively. By the second-order Taylor expansion with integral remainder,

$$-L_M(\theta^*_M+\delta) + L_M(\theta^*_M) = -S_M(\theta^*_M)'\delta + \int_0^1 (1-t)\,\delta'\hat{H}_M(\theta^*_M+t\delta)\delta\, dt.$$

Since, by (A.11), uniformly over $\|\delta\|_2 \le r_{n,M}$, $\lambda_{\min}\left(\hat{H}_M(\theta^*_M+\delta)\right) \ge \frac{c_H}{4}$ wpa 1, for every $t \in [0,1]$, $\delta'\hat{H}_M(\theta^*_M+t\delta)\delta \ge \frac{c_H}{4}\|\delta\|_2^2$ wpa 1. It follows that wpa 1

$$\int_0^1 (1-t)\,\delta'\hat{H}_M(\theta^*_M+t\delta)\delta\, dt \ge \frac{c_H}{4}\|\delta\|_2^2\int_0^1(1-t)\,dt = \frac{c_H}{8}\|\delta\|_2^2.$$

Therefore, by the Cauchy–Schwarz inequality, wpa 1

$$-L_M(\theta^*_M+\delta) + L_M(\theta^*_M) \ge -S_M(\theta^*_M)'\delta + \frac{c_H}{8}\|\delta\|_2^2 \ge -\|S_M(\theta^*_M)\|_2\|\delta\|_2 + \frac{c_H}{8}\|\delta\|_2^2.$$

For the penalty term, the reverse triangle inequality and the Cauchy–Schwarz inequality yield

$$\|\beta^*_M+\delta_\beta\|_1 - \|\beta^*_M\|_1 \ge -\|\delta_\beta\|_1 \ge -\sqrt{|M|-d}\,\|\delta_\beta\|_2 \ge -\sqrt{|M|-d}\,\|\delta\|_2.$$

Consequently,

$$\begin{aligned} &-L_M(\theta^*_M+\delta) + \lambda\|\beta^*_M+\delta_\beta\|_1 + L_M(\theta^*_M) - \lambda\|\beta^*_M\|_1 \\ &\quad\ge -\left(\|S_M(\theta^*_M)\|_2 + \sqrt{|M|-d}\,\lambda\right)\|\delta\|_2 + \frac{c_H}{8}\|\delta\|_2^2. \end{aligned}$$

By (A.8), the right-hand side is positive uniformly on the sphere $\|\delta\|_2 = r_{n,M}$ with probability arbitrarily close to one when A_0 is sufficiently large. By convexity, the minimizer is inside the sphere, proving (A.5). Equation (A.6) follows from $\|v\|_1 \le \sqrt{|M|}\,\|v\|_2$. Finally, since d is fixed and $\lambda \asymp \sqrt{(\log 2p)/n} \ge cn^{-1/2}$, (A.7) follows. □

A.3 Proof of Lemma 3.5

Fix a deterministic sequence $M = M_n \in \mathcal{M}_n(m_n)$. Define the infeasible group means and covariance estimator

$$\bar{s}^*_{h,M} \equiv n_h^{-1} \sum_{i \in \mathcal{I}_h} s_{i,M}, \quad \bar{I}^*_M \equiv n^{-1} \sum_{h=1}^{J} \frac{n_h}{n_h - 1} \sum_{i \in \mathcal{I}_h} (s_{i,M} - \bar{s}^*_{h,M})(s_{i,M} - \bar{s}^*_{h,M})'.$$

By (3.32), the centered vectors $s_{i,M} - \mu_{h,M}$ have mean zero within group h, although their distributions and covariance matrices may differ across i. Define $I_{h,M} \equiv n_h^{-1} \sum_{i \in \mathcal{I}_h} \mathrm{Var}(s_{i,M})$. Independence and common within-group means imply

$$\mathrm{E}\left[\frac{1}{n_h - 1} \sum_{i \in \mathcal{I}_h} (s_{i,M} - \bar{s}^*_{h,M})(s_{i,M} - \bar{s}^*_{h,M})'\right] = I_{h,M}.$$

The vectors $s_{i,M} - \mu_{h,M}$ are uniformly sub-Gaussian by Assumption 1(a)–(b). Applying the covariance concentration argument from the proof of Lemma 3.4 to the i.n.i.d. vectors $s_{i,M} - \mu_{h,M}$ in each group, and using $\|\bar{s}^*_{h,M} - \mu_{h,M}\|_2^2 = O_p((d+p)/n_h)$ to account for centering at the sample mean, yields

$$\left\| \frac{1}{n_h - 1} \sum_{i \in \mathcal{I}_h} (s_{i,M} - \bar{s}^*_{h,M})(s_{i,M} - \bar{s}^*_{h,M})' - I_{h,M} \right\|_2 = O_p\left(\sqrt{\frac{d+p}{n_h}} + \frac{d+p}{n_h} \right).$$

Because J is fixed, $\min_h n_h/n \ge c_s$, and $I_M = \sum_h (n_h/n) I_{h,M}$,

$$\|\bar{I}^*_M - I_M\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right). \tag{A.12}$$

Assumption 1(d) then gives

$$\|\bar{I}^*_M\|_2 = O_p(1). \tag{A.13}$$

It remains to control the replacement of θ_M^* by $\hat{\theta}_M$. Let

$$e_{i,M} \equiv \hat{s}_{i,M} - s_{i,M}, \qquad \bar{e}_{h,M} \equiv n_h^{-1} \sum_{i \in \mathcal{I}_h} e_{i,M}.$$

By the mean value theorem, for some $\tilde{t}_{i,M}$ between $x_{iM}'\theta_M^*$ and $x_{iM}'\hat{\theta}_M$, we have $e_{i,M} = -w_i x_i \ddot{g}(y_i, \tilde{t}_{i,M}) x_{iM}'(\hat{\theta}_M - \theta_M^*)$. From Lemma A.1, $\max_i |x_{iM}'(\hat{\theta}_M - \theta_M^*)| \leq C_u \|\hat{\theta}_M - \theta_M^*\|_1 = O_p\left(m_n(n^{-1/2} + \lambda)\right) = o_p(1)$. Thus, the intermediate predictors lie in the neighborhood in Assumption 1(b) wpa 1. We may therefore write

$$e_{i,M} = a_{i,M} x_i, \qquad \max_i |a_{i,M}| = O_p\left(m_n(n^{-1/2} + \lambda)\right). \tag{A.14}$$

Define

$$R_{I,M} \equiv \frac{1}{n} \sum_{h=1}^{J} \frac{n_h}{n_h - 1} \sum_{i \in \mathcal{I}_h} (e_{i,M} - \bar{e}_{h,M})(e_{i,M} - \bar{e}_{h,M})'.$$

For each group, $\sum_{i \in \mathcal{I}_h} (e_{i,M} - \bar{e}_{h,M})(e_{i,M} - \bar{e}_{h,M})' = \sum_{i \in \mathcal{I}_h} e_{i,M} e_{i,M}' - n_h \bar{e}_{h,M} \bar{e}_{h,M}'$, so (in matrix sense)

$$\sum_{i \in \mathcal{I}_h} (e_{i,M} - \bar{e}_{h,M})(e_{i,M} - \bar{e}_{h,M})' \leq \sum_{i \in \mathcal{I}_h} e_{i,M} e_{i,M}' = \sum_{i \in \mathcal{I}_h} a_{i,M}^2 x_i x_i' \leq \{\max_i |a_{i,M}|\}^2 \sum_{i \in \mathcal{I}_h} x_i x_i'.$$

Summing over h, using J fixed and $\min_h n_h/n \geq c_s$ to bound $n_h/(n_h - 1)$, together with (3.31) and (A.14), gives

$$\|R_{I,M}\|_2 \leq C\{\max_i |a_{i,M}|\}^2 \|n^{-1} X'X\|_2 = O_p\Big(m_n^2(n^{-1/2} + \lambda)^2\Big). \tag{A.15}$$

Since $\hat{s}_{i,M} - \bar{s}_{h,M} = (s_{i,M} - \bar{s}_{h,M}^*) + (e_{i,M} - \bar{e}_{h,M})$, expanding $\hat{I}_M^{\text{grp}} - \bar{I}_M^*$ gives two cross terms and the quadratic term $R_{I,M}$. By the Cauchy–Schwarz inequality for sums of outer products, $\|\sum_i a_i b_i'\|_2 \leq \|\sum_i a_i a_i'\|_2^{1/2} \|\sum_i b_i b_i'\|_2^{1/2}$,

$$\|\hat{I}_M^{\text{grp}} - \bar{I}_M^*\|_2 \leq 2\|\bar{I}_M^*\|_2^{1/2} \|R_{I,M}\|_2^{1/2} + \|R_{I,M}\|_2.$$

It follows from (A.13), (A.15), and $m_n(n^{-1/2} + \lambda) \to 0$ that

$$\|\hat{I}_M^{\text{grp}} - \bar{I}_M^*\|_2 = O_p\Big(m_n(n^{-1/2} + \lambda)\Big).$$

Since $\lambda \geq c_\lambda n^{-1/2}$, the last bound is $O_p(m_n \lambda)$. Combining it with (A.12) proves (3.33).

For a population AME, let $u_{i,M} = w_i\{\mathrm{ME}_{ij}(\theta^*_M) - \rho_M(\theta^*_M)\}$ and $\bar{w}_n = n^{-1}\sum_i \mathrm{E}[w_i]$. By (3.27) and the additional group-mean condition, group centering $(s'_{i,M}, u_{i,M}/\bar{w}_n)'$ yields an unbiased infeasible estimator of Ω_M. Since the marginal effects and weights are bounded, the preceding concentration argument gives a spectral norm error of $O_p(\sqrt{p/n})$. Furthermore, bounded logit derivatives, (A.6), and (3.28) imply $\max_i |\hat{\tau}_{i,M} - u_{i,M}/\bar{w}_n| = O_p(m_n\lambda)$. Together with (A.14) and (3.31), this implies that the grouped covariance matrix of the augmented plug-in errors has spectral norm $O_p(m_n^2\lambda^2)$. Since the infeasible augmented covariance estimator has bounded spectral norm in probability, the covariance perturbation is $O_p(m_n\lambda)$ by Cauchy–Schwarz. Combining the two bounds proves (3.34). □

A.4 Proof of Proposition 3.2

We prove the proposition in five steps: auxiliary bounds, linear representations, Gaussian coupling, estimation of the covariance and threshold, and transfer of the Gaussian pivot to the selected sample.

Step 1: Auxiliary bounds. On the event $E_{M,\mathtt{s}_M} = \{\hat{M} = M, \hat{\mathtt{s}}_M = \mathtt{s}_M\}$, the estimator $\hat{\theta}_M$ equals the active block of the full ℓ_1-PML estimator. Thus, every fixed M statistic below coincides with its feasible counterpart on this event. By the triangle inequality,

$$\|\hat{H}(\hat{\theta}_M) - H(\theta^*_M)\|_2 \le \|\hat{H}(\hat{\theta}_M) - \hat{H}(\theta^*_M)\|_2 + \|\hat{H}(\theta^*_M) - H(\theta^*_M)\|_2. \tag{A.16}$$

The second term is $O_p(\sqrt{p/n})$ by (3.30) in Lemma 3.4. For the first term, the local Lipschitz condition on $\ddot{g}$ (Assumption 1(b)) gives

$$\|\hat{H}(\hat{\theta}_M) - \hat{H}(\theta^*_M)\|_2 \le C \max_i |x'_{iM}(\hat{\theta}_M - \theta^*_M)| \|n^{-1}X'X\|_2.$$

Lemma A.1 gives $\max_i |x'_{iM}(\hat{\theta}_M - \theta^*_M)| \le C_u\|\hat{\theta}_M - \theta^*_M\|_1 = O_p\left(m_n(n^{-1/2} + \lambda)\right)$, and (3.31) in Lemma 3.4 gives $\|n^{-1}X'X\|_2 = O_p(1)$. Hence, $\|\hat{H}(\hat{\theta}_M) - \hat{H}(\theta^*_M)\|_2 = O_p\left(m_n(n^{-1/2} + \lambda)\right) = O_p(m_n\lambda)$, where the last equality follows from $\lambda = C\sqrt{(\log 2p)/n}$ in Proposition 3.2, which implies $n^{-1/2} = O(\lambda)$ since $p \ge 1$. Combining this with (A.16) and the bound on the second term gives

$$\|\hat{H}(\hat{\theta}_M) - H(\theta^*_M)\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_n\lambda\right) \equiv O_p(r_{H,n}) = o_p(1). \tag{A.17}$$

Assumption 1(d), Weyl's inequality, and the resolvent identity ((A.23) below) now imply

$$\|\hat{H}_M(\hat{\theta}_M)^{-1}\|_2 = O_p(1), \tag{A.18}$$

$$\|\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1}\|_2 = O_p(r_{H,n}), \tag{A.19}$$

$$\|\hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} - H_{-M,M}H_M^{-1}\|_2 = O_p(r_{H,n}). \tag{A.20}$$

Indeed, since $\hat{H}_M(\hat{\theta}_M)$ and H_M are the (M, M) principal submatrices of $\hat{H}(\hat{\theta}_M)$ and $H(\theta_M^*)$, the spectral norm of a principal submatrix is bounded by that of the full matrix, so (A.17) gives

$$\|\hat{H}_M(\hat{\theta}_M) - H_M\|_2 \le \|\hat{H}(\hat{\theta}_M) - H(\theta_M^*)\|_2 = O_p(r_{H,n}) = o_p(1). \tag{A.21}$$

By Weyl's inequality, $|\lambda_{\min}\left(\hat{H}_M(\hat{\theta}_M)\right) - \lambda_{\min}(H_M)| \le \|\hat{H}_M(\hat{\theta}_M) - H_M\|_2$, so (A.21) and Assumption 1(d) ($\lambda_{\min}(H_M) \ge c_H$) give, wpa 1, $\lambda_{\min}\left(\hat{H}_M(\hat{\theta}_M)\right) \ge c_H/2$, hence

$$\|\hat{H}_M(\hat{\theta}_M)^{-1}\|_2 = O_p(1), \qquad \|H_M^{-1}\|_2 \le 1/c_H = O(1), \tag{A.22}$$

which is (A.18). For (A.19), the resolvent identity

$$\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1} = \hat{H}_M(\hat{\theta}_M)^{-1}\left(H_M - \hat{H}_M(\hat{\theta}_M)\right)H_M^{-1} \tag{A.23}$$

and the Cauchy–Schwarz inequality for the spectral norm give

$$\|\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1}\|_2 \le \|\hat{H}_M(\hat{\theta}_M)^{-1}\|_2\|\hat{H}_M(\hat{\theta}_M) - H_M\|_2\|H_M^{-1}\|_2 = O_p(1)O_p(r_{H,n})O(1) = O_p(r_{H,n}),$$

using (A.21) and the two bounds in (A.22). For (A.20), write

$$\begin{aligned}
&\hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1} - H_{-M,M}H_M^{-1} \\
&= \left(\hat{H}_{-M,M}(\hat{\theta}_M) - H_{-M,M}\right)\left(\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1}\right) + \left(\hat{H}_{-M,M}(\hat{\theta}_M) - H_{-M,M}\right)H_M^{-1} \\
&\quad + H_{-M,M}\left(\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1}\right) \equiv \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

As in (A.21), the submatrix bound gives $\|\hat{H}_{-M,M}(\hat{\theta}_M) - H_{-M,M}\|_2 \le \|\hat{H}(\hat{\theta}_M) - H(\theta_M^*)\|_2 = O_p(r_{H,n})$, while $\|H_{-M,M}\|_2 \le \|H(\theta_M^*)\|_2 \le C_H = O(1)$ by Assumption 1(d). Hence,

$$\|\mathrm{I}\|_2 = O_p(r_{H,n}^2) = o_p(r_{H,n}), \qquad \|\mathrm{II}\|_2 = O_p(r_{H,n}), \qquad \|\mathrm{III}\|_2 = O_p(r_{H,n}),$$

using (A.19) for II and III, and $r_{H,n} = o(1)$ for I. Summing the three terms proves (A.20).

Also, (A.8) gives

$$\|n^{1/2}\, S_M(\theta^*_M)\|_2 = O_p(\sqrt{|M|}) = O_p(\sqrt{m_n}). \tag{A.24}$$

Step 2: Linear representations. Let $\Delta_M \equiv \hat{\theta}_M - \theta^*_M$. By the mean value expansion, for each i there exists $\tilde{t}_{i,M}$ between $x'_{iM}\theta^*_M$ and $x'_{iM}\hat{\theta}_M$ such that

$$\dot{g}(y_i, x'_{iM}\hat{\theta}_M) = \dot{g}(y_i, x'_{iM}\theta^*_M) + \ddot{g}(y_i, \tilde{t}_{i,M})\, x'_{iM}\Delta_M. \tag{A.25}$$

Multiplying by $-w_i[x_i]_A$, where $[x_i]_A$ denotes the subvector of x_i indexed by A, and averaging over i gives, for $A \in \{M, -M\}$,

$$S_A(\hat{\theta}_M) = S_A(\theta^*_M) - \left(n^{-1}\sum_{i=1}^n w_i[x_i]_A \ddot{g}(y_i, \tilde{t}_{i,M}) x'_{iM}\right)\Delta_M.$$

The matrix in the bracket involves the intermediate term $\ddot{g}(y_i, \tilde{t}_{i,M})$, rather than the plug-in term $\ddot{g}(y_i, x'_{iM}\hat{\theta}_M)$ defining $\hat{H}_{A,M}(\hat{\theta}_M)$. Adding and subtracting $\hat{H}_{A,M}(\hat{\theta}_M)\Delta_M$ therefore gives

$$S_M(\hat{\theta}_M) = S_M(\theta^*_M) - \hat{H}_M(\hat{\theta}_M)\Delta_M + R_M, \tag{A.26}$$

$$S_{-M}(\hat{\theta}_M) = S_{-M}(\theta^*_M) - \hat{H}_{-M,M}(\hat{\theta}_M)\Delta_M + R_{-M}, \tag{A.27}$$

where $R_A \equiv -n^{-1}\sum_{i=1}^n w_i[x_i]_A \left(\ddot{g}(y_i, \tilde{t}_{i,M}) - \ddot{g}(y_i, x'_{iM}\hat{\theta}_M)\right) x'_{iM}\Delta_M, A \in \{M, -M\}$. We next bound R_A. By Assumption 1(a) and (A.7),

$$\max_i |x'_{iM}\Delta_M| \le C_u\|\Delta_M\|_1 = O_p(m_n\lambda) = o_p(1).$$

Hence, wpa 1, $x'_{iM}\hat{\theta}_M$ and every intermediate point $\tilde{t}_{i,M}$ lie in the neighborhood on which Assumption 1(b) applies. In consequence,

$$|\ddot{g}(y_i, \tilde{t}_{i,M}) - \ddot{g}(y_i, x'_{iM}\hat{\theta}_M)| \le L_0|\tilde{t}_{i,M} - x'_{iM}\hat{\theta}_M| \le L_0|x'_{iM}\Delta_M| \le C\|\Delta_M\|_1.$$

Thus, the remainder is quadratic in Δ_M. To make the rate explicit, define

$$D_M \equiv \mathrm{diag}\Big[w_i\left(\ddot{g}(y_i, \tilde{t}_{i,M}) - \ddot{g}(y_i, x'_{iM}\hat{\theta}_M)\right)\Big]_{i=1}^n.$$

The stacked remainder can be rewritten as (up to a permutation of coordinates which does not affect the Euclidean norm) $(R'_M, R'_{-M})' = -n^{-1}X'D_M X_M \Delta_M$. Since D_M is diagonal, boundedness of the weights and the preceding Lipschitz bound imply $\|D_M\|_2 = \max_i |(D_M)_{ii}| \leq C\|\Delta_M\|_1 = O_p(m_n\lambda)$. Therefore,

$$\begin{aligned}\left\|n^{-1}X'D_M X_M \Delta_M\right\|_2 &\leq \|D_M\|_2 \|n^{-1}X'X\|_2^{1/2} \|n^{-1}X'_M X_M\|_2^{1/2} \|\Delta_M\|_2 \\ &= O_p(m_n\lambda)\, O_p(1)\, O_p(1)\, O_p(\sqrt{m_n}\lambda) \\ &= O_p(m_n^{3/2}\lambda^2),\end{aligned}$$

where we use $\|n^{-1}X'X\|_2 = O_p(1)$ (Lemma 3.4) and $\|n^{-1}X'_M X_M\|_2 \leq \|n^{-1}X'X\|_2$. Since $\lambda = C\sqrt{(\log 2p)/n}$,

$$n^{1/2}\left\|(R'_M, R'_{-M})'\right\|_2 = O_p\left(\frac{m_n^{3/2}\log 2p}{n^{1/2}}\right) = o_p\Big((\log 2p)^{-1/2}\Big), \tag{A.28}$$

because $\frac{m_n^{3/2}(\log 2p)^{3/2}}{n^{1/2}} \to 0$ by the rate conditions of Proposition 3.2. Combining (3.5), (A.26), and (A.28) gives

$$n^{1/2}(\tilde{\theta}_M - \theta^*_M) = \hat{H}_M(\hat{\theta}_M)^{-1} n^{1/2} S_M(\theta^*_M) + o_p\Big((\log 2p)^{-1/2}\Big).$$

By (A.19) and (A.24),

$$\left\|\Big(\hat{H}_M(\hat{\theta}_M)^{-1} - H_M^{-1}\Big)\, n^{1/2} S_M(\theta^*_M)\right\|_2 = O_p(r_{H,n} m_n^{1/2}) = O_p\left(\sqrt{\frac{pm_n}{n}} + m_n^{3/2}\lambda\right) = o_p\left(\sqrt{\frac{1}{\log 2p}}\right),$$

where the last equality follows from the rate conditions of Proposition 3.2. Hence,

$$n^{1/2}(\tilde{\theta}_M - \theta^*_M) = H_M^{-1} n^{1/2} S_M(\theta^*_M) + o_p\Big((\log 2p)^{-1/2}\Big). \tag{A.29}$$

For the inactive score, let $\hat{B}_M \equiv \hat{H}_{-M,M}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}$ and $B_M \equiv H_{-M,M}H_M^{-1}$. Substituting (A.26) and (A.27) into (3.6), the terms involving Δ_M cancel exactly:

$$\tilde{S}_{-M}(\hat{\theta}_M) = S_{-M}(\theta^*_M) - \hat{B}_M S_M(\theta^*_M) + R_{-M} - \hat{B}_M R_M.$$

Therefore,

$$n^{1/2}\left(\tilde{S}_{-M}(\hat{\theta}_M)-\nu_{-M}\right)=n^{1/2}\left(S_{-M}(\theta_M^*)-\nu_{-M}\right)-B_M n^{1/2}S_M(\theta_M^*) \\ -(\hat{B}_M-B_M)n^{1/2}S_M(\theta_M^*)+n^{1/2}(R_{-M}-\hat{B}_M R_M).$$

By (A.20), (A.24), and (A.28),

$$\left\|(\hat{B}_M-B_M)n^{1/2}S_M(\theta_M^*)\right\|_2=O_p(r_{H,n}\sqrt{m_n})=o_p\Big((\log 2p)^{-1/2}\Big),$$

while $\|\hat{B}_M\|_2=O_p(1)$ implies $n^{1/2}\left\|R_{-M}-\hat{B}_M R_M\right\|_2=o_p\Big((\log 2p)^{-1/2}\Big)$. It follows that

$$n^{1/2}\left(\tilde{S}_{-M}(\hat{\theta}_M)-\nu_{-M}\right)=n^{1/2}\left(S_{-M}(\theta_M^*)-\nu_{-M}\right)-H_{-M,M}H_M^{-1}n^{1/2}S_M(\theta_M^*)+o_p\Big((\log 2p)^{-1/2}\Big). \tag{A.30}$$

For the general functional, Taylor expansion of the sample functional and its locally Lipschitz Jacobian give

$$n^{1/2}\left|\hat{\rho}_M(\hat{\theta}_M)-\hat{\rho}_M(\theta_M^*)-\nabla\hat{\rho}_M(\hat{\theta}_M)'(\hat{\theta}_M-\theta_M^*)\right|\le Cn^{1/2}\|\hat{\theta}_M-\theta_M^*\|_2^2 \\ =O_p\left(\frac{m_n\log 2p}{n^{1/2}}\right)=o_p\Big((\log 2p)^{-1/2}\Big).$$

Moreover, the empirical Jacobian condition in Assumption 1(e), its local Lipschitz property, (A.7), and (A.24) imply

$$\left|\left(\nabla\hat{\rho}_M(\hat{\theta}_M)-\nabla\rho_M(\theta_M^*)\right)'H_M^{-1}n^{1/2}S_M(\theta_M^*)\right|=O_p(m_n\lambda)=o_p\Big((\log 2p)^{-1/2}\Big).$$

By Assumption 1(e) and the influence-function representation (3.11),

$$n^{1/2}\left(\hat{\rho}_M(\theta_M^*)-\rho_M(\theta_M^*)\right)=n^{-1/2}\sum_{i=1}^n\tau_{i,M}+o_p\left((\log 2p)^{-1/2}\right).$$

For functionals depending only on the parameter vector, $\hat{\rho}_M=\rho_M$ and $\tau_{i,M}=0$. Combining this representation with the preceding Taylor expansion, the gradient bounds, and (A.29) yields

$$n^{1/2}\left(\tilde{\rho}_M-\rho_M(\theta_M^*)\right)=\nabla\rho_M(\theta_M^*)'H_M^{-1}n^{1/2}S_M(\theta_M^*)+n^{-1/2}\sum_{i=1}^n\tau_{i,M}+o_p\left((\log 2p)^{-1/2}\right). \tag{A.31}$$

Since $\mathrm{E}[S_M(\theta_M^*)] = 0$ and $\mathrm{E}[S_{-M}(\theta_M^*)] = \nu_{-M}$, (A.29), (A.30), and (A.31) give

$$Z - \mu_M = Q_M n^{-1/2} \sum_{i=1}^{n} \omega_{i,M} + r_{M,n}, \qquad \|r_{M,n}\|_\infty = o_p\Big((\log 2p)^{-1/2}\Big). \tag{A.32}$$

Step 3: Gaussian coupling. Set $W_M \equiv Q_M n^{-1/2} \sum_{i=1}^n \omega_{i,M}$. By Assumption 1(a)–(b), the weights w_i and score derivatives $\dot{g}(y_i, x_{iM}'\theta_M^*)$ are uniformly bounded. Since $s_{i,M} = -w_i \dot{g}(y_i, x_{iM}'\theta_M^*) x_i$ and x_i is uniformly sub-Gaussian, $\|s_{i,M}\|_{\psi_2} \le C$. Centering preserves this bound, so $\|\xi_{i,M}\|_{\psi_2} \le C$. Assumption 1(e) gives the same bound for $\tau_{i,M}$, and hence $\|\omega_{i,M}\|_{\psi_2} \le C$.

We first bound the population transformation matrix Q_M. By (3.14), Assumption 1(d), and Assumption 1(e), $\|H_M^{-1}\|_2 \le c_H^{-1}$ and $\|H_{-M,M} H_M^{-1}\|_2 \le \frac{C_H}{c_H}$, because $\|H_{-M,M}\|_2 \le \|H(\theta_M^*)\|_2 \le C_H$. Moreover,

$$\|\nabla \rho_M(\theta_M^*)' H_M^{-1}\|_2 \le \frac{\|\nabla \rho_M(\theta_M^*)\|_2}{c_H} \le \frac{\sqrt{C_\rho}}{c_H}.$$

Since the remaining nonzero blocks of Q_M are identity matrices, it follows that

$$\|Q_M\|_2 = O(1). \tag{A.33}$$

We next apply Theorem B.1 to $U_{i,M} \equiv n^{-1/2} Q_M \omega_{i,M} \in \mathbb{R}^{d+p+1}$ and $V_{i,M} \equiv \mathrm{Var}(U_{i,M})$. By the uniform sub-Gaussian bounds established above and (A.33),

$$\|U_{i,M}\|_{\psi_2} \le C n^{-1/2}, \qquad \max_j (V_{i,M})_{jj} \le C n^{-1}, \qquad \mathrm{tr}(V_{i,M}) \le Cp/n.$$

The sub-Gaussian moment bounds imply $\mathrm{E}\left[\|U_{i,M}\|_2^4\right] \le Cp^2/n^2$ and $\mathrm{E}\left[\|U_{i,M}\|_\infty^2\right] \le C\frac{\log 2p}{n}$. Hence, by Cauchy–Schwarz,

$$\sum_{i=1}^n \mathrm{E}\left[\|U_{i,M}\|_2^2 \|U_{i,M}\|_\infty\right] \le \sum_{i=1}^n \left(\mathrm{E}\left[\|U_{i,M}\|_2^4\right]\right)^{1/2} \left(\mathrm{E}\left[\|U_{i,M}\|_\infty^2\right]\right)^{1/2} = O\left(\frac{p\sqrt{\log 2p}}{\sqrt{n}}\right).$$

The Gaussian inequality in Lemma SA.3 of Cattaneo et al. (2025) gives

$$\sum_{i=1}^n \mathrm{E}\left[\|V_{i,M}^{1/2} Z_i\|_2^2 \|V_{i,M}^{1/2} Z_i\|_\infty\right] \le C \sum_{i=1}^n \sqrt{\frac{\log 2p}{n}}\, \mathrm{tr}(V_{i,M}) = O\left(\frac{p\sqrt{\log 2p}}{\sqrt{n}}\right).$$

Consequently,

$$\beta_{\infty,2} = O\left(\frac{p\sqrt{\log 2p}}{\sqrt{n}}\right). \tag{A.34}$$

Since $\phi_\infty(d+p+1) = O(\sqrt{\log 2p})$ and $\sum_{i=1}^n V_{i,M} = Q_M \Omega_M Q_M' = \Sigma_M$, Theorem B.1 implies that, for any sequence $R_n \to \infty$, there exists a coupling with $T_M \sim N(0, \Sigma_M)$ such that

$$\|W_M - T_M\|_\infty = O_p\left(\frac{p^{1/3}\sqrt{\log 2p}}{n^{1/6}} R_n\right). \tag{A.35}$$

By the rate conditions of Proposition 3.2, $\frac{p^{1/3}\log 2p}{n^{1/6}} \to 0$. Hence, $R_n \to \infty$ can be chosen sufficiently slowly that $\frac{p^{1/3}\sqrt{\log 2p}}{n^{1/6}} R_n = o\Big((\log 2p)^{-1/2}\Big)$. Therefore, $\|W_M - T_M\|_\infty = o_p\Big((\log 2p)^{-1/2}\Big)$. Combining this coupling with (A.32) yields

$$\|Z - \mu_M - T_M\|_\infty = o_p\Big((\log 2p)^{-1/2}\Big). \tag{A.36}$$

The possibly nonzero inactive mean ν_{-M} enters only through the deterministic centering μ_M; the Gaussian coupling itself is applied to the centered pseudo-score. Write $G_M \equiv \mu_M + T_M$ for the coupled Gaussian vector.

Step 4: Estimated covariance and threshold. Equation (A.17), the empirical gradient condition in Assumption 1(e), and (A.7) imply

$$\|\hat{Q}_M - Q_M\|_2 = O_p\Big(\sqrt{\frac{p}{n}} + m_n\lambda\Big) = o_p(1). \tag{A.37}$$

By Assumption 1(d) and the uniform second-moment bound on $\tau_{i,M}$, $\|\Omega_M\|_2 = O(1)$, while Assumption 2 implies $\|\hat{\Omega}_M\|_2 = O_p(1)$. Together with (A.33) and (A.37),

$$\begin{aligned}\|\hat{\Sigma}_M - \Sigma_M\|_2 \le{}& \|\hat{Q}_M - Q_M\|_2\|\hat{\Omega}_M\|_2\|\hat{Q}_M\|_2 + \|Q_M\|_2\|\hat{\Omega}_M - \Omega_M\|_2\|\hat{Q}_M\|_2 \\ &+ \|Q_M\|_2\|\Omega_M\|_2\|\hat{Q}_M - Q_M\|_2 \\ ={}& O_p\Big(\sqrt{\frac{p}{n}} + m_n\lambda + r_{I,n}\Big) = o_p(1).\end{aligned} \tag{A.38}$$

The variance of the target coordinate is bounded away from zero by Assumption 1(e):

$$\sigma_M^2 = \boldsymbol{e}_1'\Sigma_M\boldsymbol{e}_1 \ge c_\sigma \tag{A.39}$$

for some constant $c_\sigma > 0$ and all sufficiently large n. This condition also covers the possibility of covariance between the empirical AME and the parameter estimation component; for a

parameter functional it follows from the original score and curvature bounds. Consequently,

$$\left|\boldsymbol{e}_1'\hat{\Sigma}_M\boldsymbol{e}_1 - \sigma_M^2\right| \le \|\hat{\Sigma}_M - \Sigma_M\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_n\lambda + r_{I,n}\right).$$

Moreover, since $c_M = \Sigma_M\boldsymbol{e}_1\left(\boldsymbol{e}_1'\Sigma_M\boldsymbol{e}_1\right)^{-1}$, and the denominator is bounded away from zero by (A.39), the map $\Sigma \mapsto \frac{\Sigma\boldsymbol{e}_1}{\boldsymbol{e}_1'\Sigma\boldsymbol{e}_1}$ is locally Lipschitz at Σ_M. Hence,

$$\|\hat{c}_M - c_M\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_n\lambda + r_{I,n}\right) = o_p(1). \tag{A.40}$$

Let b_M denote the population version of $\hat{b}$ in (3.9), obtained by replacing $\hat{H}_M(\hat{\theta}_M)$ and $\hat{H}_{-M,M}(\hat{\theta}_M)$ with H_M and $H_{-M,M}$. Since $\|(\mathbf{0}_d', \mathbf{s}_M')'\|_2 = \sqrt{|M| - d}$ and $n^{1/2}\lambda = C\sqrt{\log 2p}$, (A.19) and (A.20) give

$$\|\hat{b} - b_M\|_\infty = O_p\left(r_{H,n}\sqrt{m_n \log 2p}\right) = O_p\left(\sqrt{\frac{pm_n \log 2p}{n}} + \frac{m_n^{3/2}\log 2p}{n^{1/2}}\right) = o_p(1). \tag{A.41}$$

Finally, the rate conditions of Proposition 3.2, the covariance rate in Assumption 2, and (A.36) imply

$$\sqrt{\log 2p}\,\max\{\|Z - G_M\|_\infty, \|\hat{\Sigma}_M - \Sigma_M\|_2, \|\hat{b} - b_M\|_\infty, \|\hat{c}_M - c_M\|_2\} = o_p(1). \tag{A.42}$$

Step 5: Transfer of the Gaussian pivot. Write $E^G_{M,\mathbf{s}_M} \equiv \{AG_M \le b_M\}$. For the truncation limits in (3.19)–(3.21), write $\mathcal{V}^{\pm}_{b,c}(r)$ and $\mathcal{V}^{0}_{b,c}(r)$ when the threshold and projection vector need to be shown explicitly. Recall that $G_M = \mu_M + T_M$. By Lemma 3.1,

$$\mathcal{P}^G_{M,n} \equiv F\left(\boldsymbol{e}_1'G_M; \boldsymbol{e}_1'\mu_M, \sigma_M^2, \mathcal{V}^-_{b_M,c_M}\left(r(G_M,\Sigma_M,\boldsymbol{e}_1)\right), \mathcal{V}^+_{b_M,c_M}\left(r(G_M,\Sigma_M,\boldsymbol{e}_1)\right)\right) \tag{A.43}$$

satisfies $\mathcal{P}^G_{M,n} \,|\, E^G_{M,\mathbf{s}_M} \sim U(0,1)$.

Substep 5.1: Selection-event approximation. We first establish an anti-concentration bound to show that the feasible and Gaussian selection events agree asymptotically. Let $a_{j,M}'$ be the row j of A. Since the first column of A is zero and the last column of Q_M is $\boldsymbol{e}_1$, the empirical-averaging component does not enter the selection constraints. Writing Q^s_M for the first $d+p$ columns of Q_M, we have

$$a_{j,M}'\Sigma_M a_{j,M} = a_{j,M}'Q^s_M I_M {Q^s_M}' a_{j,M} \ge c_I\|a_{j,M}'Q^s_M\|_2^2.$$

For an active constraint, $a'_{j,M}Q_M$ is, up to sign, a row of PH_M^{-1}. Since $1 \le \|e'_j PH_M^{-1}\|_2 \, \|H_M\|_2$ and $\|H_M\|_2 \le \|H(\theta_M^*)\|_2 \le C_H$, $\|e'_j PH_M^{-1}\|_2 \ge C_H^{-1}$. For an inactive constraint, $a'_{j,M}Q_M$ contains a unit coordinate in its inactive block, so its Euclidean norm is at least one. Thus, there is a constant $c_A > 0$, independent of M and n, such that $a'_{j,M}\Sigma_M a_{j,M} \ge c_A$. Assumption 1 (d) and (A.33) give $a'_{j,M}\Sigma_M a_{j,M} \le C_I \|Q_M\|_2^2 \|a_{j,M}\|_2^2 \le C_A$, since each row of A has unit Euclidean norm. Hence, for constants $0 < c_A < C_A < \infty$,

$$c_A \le a'_{j,M}\Sigma_M a_{j,M} \le C_A, \qquad 1 \le j \le 2(d+p) - |M| - d. \tag{A.44}$$

Set

$$U_{M,n}^G \equiv \max_{1 \le j \le 2(d+p)-|M|-d} \left(a'_{j,M} G_M - b_{M,j} \right), \qquad \hat{U}_{M,n} \equiv \max_{1 \le j \le 2(d+p)-|M|-d} \left(a'_{j,M} Z - \hat{b}_j \right).$$

Then $E_{M,\mathtt{s}_M}^G = \{U_{M,n}^G \le 0\}$, and, up to the boundary of the polyhedral representation, $E_{M,\mathtt{s}_M} = \{\hat{U}_{M,n} \le 0\}$. To control that boundary, define the centered Gaussian vector

$$Y_M \equiv \left(a'_{1,M} T_M, \ldots, a'_{2(d+p)-|M|-d,M} T_M \right)',$$

where T_M is as in (A.35). By (A.44), every coordinate of Y_M has variance bounded below by c_A. Nazarov's Gaussian anti-concentration inequality, in the hyperrectangle form of Chernozhukov et al. (2017, Lemma A.1), therefore implies that, uniformly over $y \in \mathbb{R}^{2(d+p)-|M|-d}$ and $\varepsilon > 0$,

$$P\left[Y_M \le y + \varepsilon \mathbf{1}\right] - P[Y_M \le y] \le C\varepsilon\sqrt{\log 2p},$$

because $2(d+p) - |M| - d = O(p)$. Taking $y = b_M - A\mu_M - \varepsilon \mathbf{1}$ and applying the preceding inequality with 2ε in place of ε yields

$$P\left[|U_{M,n}^G| \le \varepsilon\right] \le C\varepsilon\sqrt{\log 2p}, \tag{A.45}$$

where the factor of two is absorbed into C. Remark that every row of A has exactly one nonzero entry, equal to 1 or -1, hence $|\hat{U}_{M,n} - U_{M,n}^G| \le \|Z - G_M\|_\infty + \|\hat{b} - b_M\|_\infty$. From (A.42), there exists a deterministic sequence $\gamma_n \downarrow 0$ such that $\gamma_n\sqrt{\log 2p} \to 0$,

$$\max\{\|Z - G_M\|_\infty, \|\hat{\Sigma}_M - \Sigma_M\|_2, \|\hat{b} - b_M\|_\infty, \|\hat{c}_M - c_M\|_2\} = o_p(\gamma_n). \tag{A.46}$$

Therefore, letting $\triangle$ denote the symmetric difference of sets, we have

$$P\Big[E_{M,\mathbf{s}_M} \triangle E^G_{M,\mathbf{s}_M}\Big] \le P\Big[|\hat{U}_{M,n} - U^G_{M,n}| > \gamma_n\Big] + P\Big[|U^G_{M,n}| \le \gamma_n\Big] \to 0. \tag{A.47}$$

Thus, the feasible and Gaussian selection events agree asymptotically. Since $P[E_{M,\mathbf{s}_M}]$ is bounded away from zero by Assumption 3(b), (A.47) also gives

$$\liminf_{n\to\infty} \; P[E^G_{M,\mathbf{s}_M}] > 0, \qquad \frac{P[E^G_{M,\mathbf{s}_M}]}{P[E_{M,\mathbf{s}_M}]} \to 1. \tag{A.48}$$

Substep 5.2: Truncation nondegeneracy. We next show that, conditional on Gaussian selection, the truncation interval has sufficient probability mass and the constraints with zero population slopes have sufficient slack. These properties ensure that the truncated Gaussian pivot is stable under the approximation errors. The Gaussian mass of the truncation interval is

$$\mathcal{D}_M(r) \equiv \Phi\left(\frac{\mathcal{V}^+_{b_M,c_M}(r) - \boldsymbol{e}_1'\mu_M}{\sigma_M}\right) - \Phi\left(\frac{\mathcal{V}^-_{b_M,c_M}(r) - \boldsymbol{e}_1'\mu_M}{\sigma_M}\right). \tag{A.49}$$

Write $r^G_M = r(G_M, \Sigma_M, \boldsymbol{e}_1)$ and take $\mathcal{V}^0_{b_M,c_M}(r) = +\infty$ when no row j has $(Ac_M)_j = 0$. Conditional on r^G_M, Lemma 3.1 implies $P\left[E^G_{M,\mathbf{s}_M} \,|\, r^G_M\right] = 1\Big(\mathcal{V}^0_{b_M,c_M}(r^G_M) \ge 0\Big)\,\mathcal{D}_M(r^G_M)$. Consequently,

$$P\Big[\mathcal{D}_M(r^G_M) < \gamma_n,\, E^G_{M,\mathbf{s}_M}\Big] = \mathrm{E}\Big[1\Big(\mathcal{D}_M(r^G_M) < \gamma_n\Big)\,1\Big(\mathcal{V}^0_{b_M,c_M}(r^G_M) \ge 0\Big)\,\mathcal{D}_M(r^G_M)\Big] \le \gamma_n.$$

Hence,

$$P\Big[\mathcal{D}_M(r^G_M) < \gamma_n \,\Big|\, E^G_{M,\mathbf{s}_M}\Big] \le \frac{\gamma_n}{P[E^G_{M,\mathbf{s}_M}]} \to 0. \tag{A.50}$$

It remains to control the rows with zero population slope. Let $\mathcal{B} \equiv \{j : (Ac_M)_j = 0\}$. If $\mathcal{B} \neq \varnothing$, then, because $a_{j,M}'c_M = 0$ for $j \in \mathcal{B}$,

$$-\mathcal{V}^0_{b_M,c_M}(r^G_M) = \max_{j\in\mathcal{B}}\{a_{j,M}'r^G_M - b_{M,j}\} = \max_{j\in\mathcal{B}}\{a_{j,M}'G_M - b_{M,j}\}.$$

The subvector $\{a_{j,M}'T_M : j \in \mathcal{B}\}$ again has coordinate variances bounded below by c_A. Applying Nazarov's inequality again gives

$$P\Big[0 \le \mathcal{V}^0_{b_M,c_M}(r^G_M) < \gamma_n,\, E^G_{M,\mathbf{s}_M}\Big] \le C\gamma_n\sqrt{\log 2p}.$$

Dividing by $P[E^G_{M,\mathbf{s}_M}]$ and using (A.48),

$$P\Big[0 \le \mathcal{V}^0_{b_M,c_M}(r^G_M) < \gamma_n \,\Big|\, E^G_{M,\mathbf{s}_M}\Big] \to 0. \tag{A.51}$$

When $\mathcal{B} = \varnothing$, the same conclusion is automatic under the convention $\mathcal{V}^0_{b_M,c_M} = +\infty$. Combining (A.50) and (A.51) yields

$$P\Big[\mathcal{D}_M(r^G_M) \ge \gamma_n,\ \mathcal{V}^0_{b_M,c_M}(r^G_M) \ge \gamma_n \,\Big|\, E^G_{M,\mathbf{s}_M}\Big] \to 1. \tag{A.52}$$

Substep 5.3: Pivot transfer. Let $\mathcal{P}_{M,n}$ denote the feasible pivot evaluated at $\rho_M(\theta^*_M)$. By (A.46), Assumption 3(a) implies that, for every j with $(Ac_M)_j \neq 0$, $\operatorname{sign}((A\hat{c}_M)_j) = \operatorname{sign}((Ac_M)_j)$ and $|(A\hat{c}_M)_j| \ge \kappa/2$ wpa 1. For these rows,

$$\frac{b_{M,j} - (Ar^G_M)_j}{(Ac_M)_j} = \boldsymbol{e}_1' G_M + \frac{b_{M,j} - (AG_M)_j}{(Ac_M)_j}, \tag{A.53}$$

and the analogous identity holds for $\hat{b}$, $\hat{r}$, $\hat{c}_M$, and Z. Since $\boldsymbol{e}_1'(G_M - \mu_M) = O_p(1)$, the Gaussian density is bounded, and $\sup_{t\in\mathbb{R}} |t|\phi(t) < \infty$, the Gaussian CDF evaluated at these truncation bounds is uniformly Lipschitz in the target coordinate, constraint slack, and nonzero slope, with an $O_p(1)$ Lipschitz bound. Thus, the corresponding Gaussian CDF values differ by $o_p(1)$.

For rows with $(Ac_M)_j = 0$, on the event in (A.52), $b_{M,j} - (AG_M)_j \ge \gamma_n$. By (A.46), $\hat{b}_j - (AZ)_j \ge \gamma_n - o_p(\gamma_n)$ and $(A\hat{c}_M)_j = o_p(\gamma_n)$. Consequently, whenever $(A\hat{c}_M)_j \neq 0$,

$$\left|\frac{\hat{b}_j - (A\hat{r})_j}{(A\hat{c}_M)_j} - \boldsymbol{e}_1' Z\right| = \left|\frac{\hat{b}_j - (AZ)_j}{(A\hat{c}_M)_j}\right| \longrightarrow_p \infty.$$

These rows therefore contribute only asymptotically infinite truncation bounds. Together with $\|\hat{\Sigma}_M - \Sigma_M\|_2 = o_p(\gamma_n)$, these results imply that the numerator and normalizing probabilities of the feasible and Gaussian truncated CDFs differ by $o_p(1)$, conditionally on $E^G_{M,\mathbf{s}_M}$. For any fixed $\varepsilon > 0$, the argument underlying (A.50) gives

$$P\left[\mathcal{D}_M(r^G_M) < \varepsilon \,\middle|\, E^G_{M,\mathbf{s}_M}\right] \le \frac{\varepsilon}{P[E^G_{M,\mathbf{s}_M}]}.$$

On $\mathcal{D}_M(r^G_M) \ge \varepsilon$, the feasible normalizing probability is at least $\varepsilon/2$ wpa 1. The $o_p(1)$ differences in the numerator and denominator therefore imply $|\mathcal{P}_{M,n} - \mathcal{P}^G_{M,n}| = o_p(1)$ on this event.

Letting $\varepsilon \downarrow 0$ and using (A.48) and (A.52) yields

$$|\mathcal{P}_{M,n} - \mathcal{P}^G_{M,n}| = o_p(1) \quad \text{conditionally on } E^G_{M,\mathbf{s}_M}. \tag{A.54}$$

By (A.47), (A.48), and the nonvanishing selection probability, the same conclusion holds conditional on the feasible selection event:

$$P\Big[|\mathcal{P}_{M,n} - \mathcal{P}^G_{M,n}| > \epsilon \,\Big|\, E_{M,\mathbf{s}_M}\Big] \to 0 \tag{A.55}$$

for every $\epsilon > 0$. By (A.47) and the nonvanishing selection probability, for any fixed $t \in (0,1)$,

$$\Big|P[\mathcal{P}^G_{M,n} \le t \,|\, E_{M,\mathbf{s}_M}] - P[\mathcal{P}^G_{M,n} \le t \,|\, E^G_{M,\mathbf{s}_M}]\Big| \le 2\frac{P[E_{M,\mathbf{s}_M} \triangle E^G_{M,\mathbf{s}_M}]}{P[E_{M,\mathbf{s}_M}]} = o(1). \tag{A.56}$$

Therefore, for fixed $t \in (0,1)$ and $\epsilon > 0$,

$$|P[\mathcal{P}_{M,n} \le t \,|\, E_{M,\mathbf{s}_M}] - t| \le P\Big[|\mathcal{P}_{M,n} - \mathcal{P}^G_{M,n}| > \epsilon \,|\, E_{M,\mathbf{s}_M}\Big] + 2\epsilon + 2\,\frac{P[E_{M,\mathbf{s}_M} \triangle E^G_{M,\mathbf{s}_M}]}{P[E_{M,\mathbf{s}_M}]}.$$

The right-hand side converges to 2ϵ by (A.43), (A.47), and (A.55). Letting $\epsilon \downarrow 0$ gives $P[\mathcal{P}_{M,n} \le t \,|\, E_{M,\mathbf{s}_M}] \to t$. Thus, the feasible pivot is asymptotically uniform conditional on $\{\hat{M} = M, \hat{\mathbf{s}}_M = \mathbf{s}_M\}$. Inverting the equal-tailed pivot yields

$$P\Big[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \,\Big|\, \hat{M} = M, \hat{\mathbf{s}}_M = \mathbf{s}_M\Big] \to 1 - \zeta.$$

A.5 Proof of Corollary 3.3

Let $q_M = |M| - d$ and $E_{M,s} \equiv \{\hat{M} = M, \hat{\mathbf{s}}_M = s\}$. Define $\Delta_{n,s} \equiv P\left[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \,|\, E_{M,s}\right] - (1-\zeta)$ for sign configurations with $P[E_{M,s}] > 0$, and set $\Delta_{n,s} = 0$ otherwise. Also, let $p_{n,s} \equiv P[E_{M,s} \mid \hat{M} = M]$. By the law of total probability,

$$P\left[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \,\middle|\, \hat{M} = M\right] - (1-\zeta) = \sum_{s \in \{-1,1\}^{q_M}} p_{n,s}\Delta_{n,s}.$$

Let $\bar{q} \equiv \sup_n q_M < \infty$. Changing the sign vector only changes the signs of some rows of A, so $|(Ac_M)_j|$ and hence Assumption 3(a) do not depend on the sign vector. For any $\varepsilon > 0$, since

$\sum_s p_{n,s} = 1$ and $|\Delta_{n,s}| \leq 1$, we have

$$\left|\sum_s p_{n,s}\Delta_{n,s}\right| \leq \sum_{s:p_{n,s}>\varepsilon} p_{n,s}|\Delta_{n,s}| + \sum_{s:p_{n,s}\leq\varepsilon} p_{n,s}|\Delta_{n,s}| \leq \sup_{s:p_{n,s}>\varepsilon} |\Delta_{n,s}| + 2^{\bar{q}}\varepsilon,$$

where the supremum over an empty set is taken to be zero. We show that the supremum converges to zero. Otherwise, there exist a subsequence and corresponding sign configurations s_n such that $p_{n,s_n} > \varepsilon$ and $|\Delta_{n,s_n}|$ remains bounded away from zero. Since $q_M \leq \bar{q}$, only finitely many sign configurations are possible, so at least one sign vector, say s, must occur infinitely often along this subsequence. Restricting to those terms yields a subsequence with a fixed sign vector s. By the assumption on $P[\hat{M} = M]$, $P[E_{M,s}] = p_{n,s}P[\hat{M} = M]$ is bounded away from zero along this sequence. The proof of Proposition 3.2, applied along the same sequence, then gives $\Delta_{n,s} \to 0$, which is a contradiction. Consequently,

$$\limsup_{n\to\infty} \left|P\left[\rho_M(\theta^*_M) \in \mathrm{CI}_{\hat{M}} \,\middle|\, \hat{M} = M\right] - (1-\zeta)\right| \leq 2^{\bar{q}}\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the result follows.

B Supplementary result

We use the following result from Cattaneo et al. (2025).

Theorem B.1 (Yurinskii coupling for independent vectors, Corollary 2.3 of Cattaneo et al. (2025))**.** *Let $U_1, \ldots, U_n$ be independent, mean-zero random vectors in $\mathbb{R}^k$, and let $V_i = \mathrm{Var}(U_i)$ and $\Sigma = \sum_{i=1}^n V_i$. Define*

$$\beta_{\infty,2} \equiv \sum_{i=1}^n \mathrm{E}\left[\|U_i\|_2^2\|U_i\|_\infty + \|V_i^{1/2}Z_i\|_2^2\|V_i^{1/2}Z_i\|_\infty\right],$$

where $Z_1, \ldots, Z_n$ are independent standard Gaussian vectors in $\mathbb{R}^k$. For every $\eta > 0$, there exists a Gaussian vector $T \sim N(0, \Sigma)$ such that

$$P\left[\left\|\sum_{i=1}^n U_i - T\right\|_\infty > \eta\right] \leq 24\left(\frac{\beta_{\infty,2}\phi_\infty(k)^2}{\eta^3}\right)^{1/3},$$

where $\phi_\infty(k) = \sqrt{2\log 2k}$, provided $\beta_{\infty,2} < \infty$.