VALORIS: A privacy-aware logistic regression method for vertically partitioned data within a novel privacy risk assessment framework **- Supplementary Information**

Supplementary Tables 1

Fig. S1 Glossary for general notation conventions

Random variable in $\mathbb{R}$	A	Uppercase Non-italic
Random vector in $\mathbb{R}^p$	$\mathbf{A}$	Uppercase Non-italic bold
Scalar in $\mathbb{R}$	a	Lowercase Italic
Vector in $\mathbb{R}^p$	$\boldsymbol{a}$	Lowercase Italic bold
Vector in $\mathbb{R}^p$ with all components equal to 1	$\mathbf{1}_p$	-
Matrix in $\mathbb{R}^{n\times p}$	$\boldsymbol{A}$	Uppercase Italic bold
Identity matrix in $\mathbb{R}^{n\times n}$	$\boldsymbol{I_n}$	-
Gradient of $f(\boldsymbol{\theta})$ (column vector)	$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$	$\nabla^2_{\boldsymbol{\theta}}$ for Hessian
$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\vert_{\boldsymbol{\theta}=\boldsymbol{a}}$	$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{a})$	$\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{a})$ for Hessian
$\max_{1\leq j\leq p} \vert a_j\vert$	$\vert\vert\boldsymbol{a}\vert\vert_\infty$	Infinite norm
$\sum_{j=1}^p \vert a_j\vert$	$\vert\vert\boldsymbol{a}\vert\vert_1$	ℓ1-norm
$\sqrt{\sum_{j=1}^p a_i^2}$	$\vert\vert\boldsymbol{a}\vert\vert_2$	ℓ2-norm
Diagonal matrix with entries of $\boldsymbol{a}$ on diagonal	$\mathrm{diag}(\boldsymbol{a})$	Dimension $p\times p$ for $\boldsymbol{a}\in\mathbb{R}^p$
Quantity $\cdot$ at iteration t (step count)	$\cdot_{(t)}$	Starts with $\cdot_{(0)}$

Fig. S2 Glossary for quantities that pertain to the regression settings

Analytical dataset with sample size n	$\mathcal{D}=\{\ldots\}_{i=1}^n$
Covariate vector for *ith* individual	$\boldsymbol{x}_i=[x_{i1},\ldots,x_{ip}]^\top$
Covariate vector for *ith* individual with intercept	$[1,\boldsymbol{x}_i^\top]^\top$
Covariate matrix in $\mathbb{R}^{n\times p}$	$\boldsymbol{X}=\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np}\end{bmatrix}=\begin{bmatrix}\boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top\end{bmatrix}$
Gram matrix	$\boldsymbol{\mathcal{K}}=[\boldsymbol{X}\ \mathbf{1}_n]\,[\boldsymbol{X}\ \mathbf{1}_n]^\top$
True (unknown) parameters	$\beta_{0\star},\boldsymbol{\beta}_\star$
Exact MLE of the parameter	$\widehat{\beta}_0,\widehat{\boldsymbol{\beta}}$
Exact penalized estimate of the parameter	$\widehat{\beta}_0^\lambda,\widehat{\boldsymbol{\beta}}^\lambda$
Estimate obtained via numerical approximation	$\tilde{\beta}_0,\tilde{\boldsymbol{\beta}},\tilde{\beta}_0^\lambda,\tilde{\boldsymbol{\beta}}^\lambda$
Log-likelihood	$\ell_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n \log(\cdot)$
Penalized log-likelihood	$l_n^\lambda(\boldsymbol{\beta})$
Mean of the jth column in covariate matrix	$u_{n,j}$
Standard deviation of the jth column in covariate matrix	$s_{n,j}$
Fisher information matrix	$\boldsymbol{\mathcal{I}}(\beta_0,\boldsymbol{\beta})$

Fig. S3 Glossary for quantities specific to the vertical setting

Number of covariate-nodes	K
Number of covariates at covariate-node k	$p^{(k)}$
Covariate matrix at covariate-node k	$\boldsymbol{X}^{(k)}$
Centered and scaled covariate matrix at covariate-node k	$\boldsymbol{X}_{\mathrm{cs}}^{(k)}$
Mean and s.d. of the jth column in $\boldsymbol{X}^{(k)}$	$u_{n,j}^{(k)}$,$s_{n,j}^{(k)}$
Gram matrix at covariate-node k	$\boldsymbol{\mathcal{K}}^{(k)} := \boldsymbol{X}_{\mathrm{cs}}^{(k)}(\boldsymbol{X}_{\mathrm{cs}}^{(k)})^\top$
Dual parameter estimates (numerical approx.)	$\tilde{\boldsymbol{\alpha}}^{\lambda}$
Penalized estimate associated with covariate-node k (numerical approx.)	$\tilde{\beta}_j^{\lambda(k)}$
Standard errors associated with covariate-node k (numerical approx.)	$\tilde{\sigma}_j^{\lambda(k)}$
Matrix in null-space of $\boldsymbol{\mathcal{K}}^{(k)}$	$\boldsymbol{\mathcal{N}}^{(k)}$
Intermediary quantities	$\tilde{\boldsymbol{c}}^{\lambda(k)}$,$\tilde{\boldsymbol{S}}$

Supplementary Notes 1

The code for the implementation of the algorithm using R is available at: https://github.com/OpenLHS/Distrib_analysis/tree/main/Vertically_distributed_analysis/logistic_regression_nonpenalized. It includes an automated example with simulated data. The folder also includes a basic implementation of the tool that supports the privacy assessment to verify if an infinite number of solutions exists in some settings.

Supplementary Methods 1 Theoretical derivations for the optimization problem

In the followings, for any $(\beta_0, \boldsymbol{\beta})$ let

$$\begin{aligned}\boldsymbol{\mathcal{I}}(\beta_0, \boldsymbol{\beta}) &= -\nabla^2_{\beta_0,\boldsymbol{\beta}} \ell_n(\beta_0, \boldsymbol{\beta}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\{y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\}}{[1 + \exp\{y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\}]^2} \begin{bmatrix} 1 & \boldsymbol{x}_i^\top \\ \boldsymbol{x}_i & \boldsymbol{x}_i \boldsymbol{x}_i^\top \end{bmatrix} . \end{aligned} \tag{S1}$$

In this notation, $\boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ corresponds to the observed Fisher information matrix introduced in Equation (10) in the manuscript.

The following lemma establishes that the unique solution to the ridge-penalized log-likelihood maximization problem for logistic regression can be obtained by solving its dual formulation, which is a minimization problem over a compact search space. This result implies that, for a given sample and a fixed λ, the solution to the dual minimization problem cannot lie arbitrarily close to the boundary of the domain $(0, 1)^n$.

Lemma S1. *For any $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \in \{-1, 1\} \times \mathbb{R}^p$, the unique maximizer $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ of the maximization problem*

$$\max_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \left(\check{l}_n^\lambda(\beta_0, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \log \left[\frac{1}{1 + \exp\{-y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\}} \right] - \frac{\lambda}{2} \sum_{j=0}^{p} \beta_j^2 \right) \tag{S2}$$

satisfies

$$\begin{bmatrix} \check{\beta}_0^\lambda \\ \check{\boldsymbol{\beta}}^\lambda \end{bmatrix} = (n\lambda)^{-1} \sum_{i=1}^{n} y_i \check{\alpha}_i^\lambda \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix} ,$$

where $\check{\boldsymbol{\alpha}}^\lambda = (\check{\alpha}_1^\lambda, \ldots, \check{\alpha}_n^\lambda)^\top \in (0, 1)^n$ is the unique solution to the following minimization problem:

$$\min_{\boldsymbol{\alpha} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda})^n} \left(\frac{1}{n} \sum_{i=1}^{n} \Big\{ (1 - \alpha_i) \log(1 - \alpha_i) + \alpha_i \log(\alpha_i) \Big\} + \frac{1}{2\lambda n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^\top \boldsymbol{x}_j + 1) \right), \tag{S3}$$

with $\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda} = \left[\{1+\exp(b^\lambda_{\boldsymbol{X}})\}^{-1}, \{1+\exp(-b^\lambda_{\boldsymbol{X}})\}^{-1}\right]$,

where $b^\lambda_{\boldsymbol{X}} = (p+1)(n\lambda)^{-1}\{\sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 1)\}^2$. *Moreover,* $\check{\boldsymbol{\alpha}}^\lambda$ *is the unique stationary point of the objective function in* (S3)*, and the set* $\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda}$ *can be replaced by* $(0,1)$.

Proof of Lemma S1. We begin by showing that the search space $\mathbb{R}\times\mathbb{R}^p$ in the maximization program on the first line at (S2) can be replaced by a suitably chosen compact set. To this end, we first note that the function $(\beta_0, \boldsymbol{\beta}) \mapsto \check{l}^\lambda_n(\beta_0, \boldsymbol{\beta})$ is strongly concave. This follows upon observing that as $\check{l}^\lambda_n(\beta_0, \boldsymbol{\beta}) = \ell_n(\beta_0, \boldsymbol{\beta}) - (\lambda/2)\sum_{j=0}^p \beta_j^2$, from (S1) we have

$$\nabla^2_{\beta_0,\boldsymbol{\beta}} \check{l}^\lambda_n(\beta_0, \boldsymbol{\beta}) = -\left(\frac{1}{n}\sum_{i=1}^n \frac{\exp\{y_i(\beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta})\}}{[1+\exp\{y_i(\beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta})\}]^2}\begin{bmatrix} 1 & \boldsymbol{x}_i^\top \\ \boldsymbol{x}_i & \boldsymbol{x}_i\boldsymbol{x}_i^\top\end{bmatrix} + \lambda \boldsymbol{I}_{p+1}\right).$$

Since the first term inside the parentheses is a weighted sum of positive semi-definite matrices with strictly positive weights, and the second term is a diagonal matrix with strictly positive entries, their sum is positive definite. This implies that the Hessian $\nabla^2_{\beta_0,\boldsymbol{\beta}} \check{l}^\lambda_n(\beta_0, \boldsymbol{\beta})$ is negative definite, and therefore, the penalized log-likelihood function $\check{l}^\lambda_n(\beta_0, \boldsymbol{\beta})$ is strongly concave.

As $\check{l}^\lambda_n(\beta_0, \boldsymbol{\beta})$ is strongly concave, its maximum is unique and is achieved at the point $(\check{\beta}^\lambda_0, \check{\boldsymbol{\beta}}^\lambda)$ that satisfies

$$n^{-1}\sum_{i=1}^n \frac{y_i}{1+\exp\{y_i(\check{\beta}^\lambda_0 + \boldsymbol{x}_i^\top \check{\boldsymbol{\beta}}^\lambda)\}}\begin{bmatrix} 1 \\ \boldsymbol{x}_i\end{bmatrix} = \lambda \begin{bmatrix} \check{\beta}^\lambda_0 \\ \check{\boldsymbol{\beta}}^\lambda\end{bmatrix}.$$

By the triangle inequality, the latter equation implies

$$\max_{0\le j\le p} |\check{\beta}^\lambda_j| \le \lambda^{-1}\Big(n^{-1}\sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 1)\Big). \tag{S4}$$

Therefore, letting $\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda} = \{\beta\in\mathbb{R} : |\beta| \le (n\lambda)^{-1}\sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 1)\}$, it holds for any $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in\mathbb{R}^p$ that

$$\max_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p} n^{-1}\sum_{i=1}^n \log\left\{\frac{1}{1+\exp\{-y_i(\beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta})\}}\right\} - \frac{\lambda}{2}\sum_{j=0}^p \beta_j^2$$

$$= \max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} n^{-1} \sum_{i=1}^n \log \left\{ \frac{1}{1+\exp\{-y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\}} \right\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2 .$$

Next, we show that

$$\max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} n^{-1} \sum_{i=1}^n \log \left\{ \frac{1}{1+\exp\{-y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\}} \right\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2$$
$$= \max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} \min_{\boldsymbol{\alpha} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda})^n} n^{-1} \sum_{i=1}^n \Big\{ y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})(1-\alpha_i) \Big\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2$$
$$+ n^{-1} \sum_{i=1}^n \Big\{ (1-\alpha_i)\log(1-\alpha_i) + \alpha_i \log(\alpha_i) \Big\}, \qquad \text{(S5)}$$

with $\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda}$ as in (S3).

To establish this result, we begin by noting that for any $x \in [a, b]$, we have

$$\log\left(\frac{1}{1+e^{-x}}\right) = \min_{\alpha \in [(1+e^{-a})^{-1}, (1+e^{-b})^{-1}]} \alpha x + (1-\alpha)\log(1-\alpha) + \alpha \log(\alpha) . \qquad \text{(S6)}$$

To see why (S6) holds, it suffices to first verify that the function $\alpha \mapsto \alpha x + (1-\alpha)\log(1-\alpha) + \alpha\log(\alpha)$ attains its minimum at $\alpha = (1+e^x)^{-1}$. Substituting $\alpha = (1+e^x)^{-1}$, we obtain

$$\min_{\alpha \in (0,1)} \alpha x + (1-\alpha)\log(1-\alpha) + \alpha \log(\alpha)$$
$$= \frac{x}{1+e^x} + \left(\frac{e^x}{1+e^x}\right)\log\left(\frac{e^x}{1+e^x}\right) + \left(\frac{1}{1+e^x}\right)\log\left(\frac{1}{1+e^x}\right)$$
$$= \log\left(\frac{e^x}{1+e^x}\right) = \log\left(\frac{1}{1+e^{-x}}\right) .$$

From this, (S6) follows directly from the fact that $x \in [a, b]$.

Since

$$\max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} \max_{1 \le i \le n} |y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})|$$
$$\le (p+1)\Big(\max_{1\le i \le n} \|\boldsymbol{x}_i\|_\infty + 1\Big)\Big((n\lambda)^{-1} \sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 1)\Big)$$

$$\leq (p+1)(n\lambda)^{-1}\Big(\sum_{i=1}^{n}(\|\boldsymbol{x}_i\|_\infty+1)\Big)^2 = b_{\boldsymbol{X}}^{\lambda}\,,$$

where $b_{\boldsymbol{X}}^{\lambda}$ is defined in the statement of the lemma, the proof that (S5) holds follows from (S6), which ensures that

$$\begin{aligned}
&\sum_{i=1}^{n}\log\left[\frac{1}{1+\exp\{-y_i(\beta_0+\boldsymbol{x}_i^\top\boldsymbol{\beta})\}}\right]\\
&=\sum_{i=1}^{n}\min_{\alpha_i\in\Theta_{\boldsymbol{\alpha},\lambda}^{\boldsymbol{X}}}\Big\{y_i(\beta_0+\boldsymbol{x}_i^\top\boldsymbol{\beta})\alpha_i+(1-\alpha_i)\log(1-\alpha_i)+\alpha_i\log(\alpha_i)\Big\}\\
&=\min_{\boldsymbol{\alpha}\in(\Theta_{\boldsymbol{\alpha},\lambda}^{\boldsymbol{X}})^n}\sum_{i=1}^{n}\Big\{y_i(\beta_0+\boldsymbol{x}_i^\top\boldsymbol{\beta})\alpha_i+(1-\alpha_i)\log(1-\alpha_i)+\alpha_i\log(\alpha_i)\Big\}\,.
\end{aligned}$$

To conclude the proof of the lemma, it remains to show that in the optimization problem given on the second line of (S5), we can interchange the maximum and minimum operations and then solve the inner maximization problem explicitly.

To prove that we can swap the max and the min, we apply Sion's minimax theorem [1, 2]. To justify the application of this theorem and conclude that the order of the minimum and maximum operators can be interchanged, we must verify that the function

$$\begin{aligned}
g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha}) :=&\sum_{i=1}^{n}\Big[y_i(\beta_0+\boldsymbol{x}_i^\top\boldsymbol{\beta})\alpha_i+\big\{(1-\alpha_i)\log(1-\alpha_i)+\alpha_i\log(\alpha_i)\big\}\Big]\\
&-\frac{\lambda}{2}\sum_{j=0}^{p}\beta_j^2
\end{aligned}$$

is such that for any fixed $(\beta_0,\boldsymbol{\beta})$, $\boldsymbol{\alpha}\mapsto g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha})$ is convex, and for any fixed $\boldsymbol{\alpha}$, $(\beta_0,\boldsymbol{\beta})\mapsto g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha})$ is concave. Once these conditions are established, Sion's theorem guarantees that the maximum and minimum operators can be interchanged.

The convexity of $\boldsymbol{\alpha}\mapsto g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha})$ follows from the fact $\nabla_{\boldsymbol{\alpha}}^2 g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha})$ is a diagonal matrix, with diagonal entries given by the following, for $j\in\{1,\ldots,n\}$:

$$[\nabla_{\boldsymbol{\alpha}}^2 g(\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha})]_{jj}=\{\alpha_j(1-\alpha_j)\}^{-1}\geq 1/4\,.$$

Since each diagonal entry is positive for all $\boldsymbol{\alpha} \in (0,1)^n$ and $(\beta_0, \boldsymbol{\beta}) \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^{p+1}$, the Hessian is positive definite, which ensures that $\boldsymbol{\alpha} \mapsto g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is convex in $\boldsymbol{\alpha}$ for all $(\beta_0, \boldsymbol{\beta})$.

The concavity of $(\beta_0, \boldsymbol{\beta}) \mapsto g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ follows from the fact that $\nabla^2_{\beta_0,\boldsymbol{\beta}}(\beta_0, \boldsymbol{\beta}) = -\lambda \boldsymbol{I}_{p+1}$, with $\boldsymbol{I}_{p+1}$ the identity matrix of size $(p+1) \times (p+1)$. As the Hessian of $g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $(\beta_0, \boldsymbol{\beta})$ is a diagonal matrix with negative entries, it is therefore negative definite, which ensures that $(\beta_0, \boldsymbol{\beta}) \mapsto g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is concave for all $(\beta_0, \boldsymbol{\beta})$ and $\boldsymbol{\alpha}$.

Since we have just proved that for any fixed $(\beta_0, \boldsymbol{\beta})$, $\boldsymbol{\alpha} \mapsto g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is convex, and for any fixed $\boldsymbol{\alpha}$, $(\beta_0, \boldsymbol{\beta}) \mapsto g(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is concave, we can apply Sion's minimax theorem, which, from (S5), ensures that

$$\begin{aligned}
&\max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} \min_{\boldsymbol{\alpha} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda})^n} n^{-1} \sum_{i=1}^n \left\{ y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\alpha_i \right\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2 \\
&\qquad\qquad + n^{-1} \sum_{i=1}^n \left\{ (1-\alpha_i)\log(1-\alpha_i) + \alpha_i \log(\alpha_i) \right\} \\
&= \min_{\boldsymbol{\alpha} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda})^n} \left(\max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} n^{-1} \sum_{i=1}^n \left\{ y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\alpha_i \right\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2 \right) \\
&\qquad\qquad + n^{-1} \sum_{i=1}^n \left\{ (1-\alpha_i)\log(1-\alpha_i) + \alpha_i \log(\alpha_i) \right\}.
\end{aligned}$$

Now the inner maximization program can be solved exactly, since, for any $\boldsymbol{\alpha} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda})^n$, the maximum of $n^{-1} \sum_{i=1}^n \{y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\alpha_i\} - (\lambda/2) \sum_{j=0}^p \beta_j^2$ is achieved at

$$\begin{bmatrix} \check{\beta}_0^\lambda(\boldsymbol{\alpha}) \\ \check{\boldsymbol{\beta}}^\lambda(\boldsymbol{\alpha}) \end{bmatrix} = (n\lambda)^{-1} \sum_{i=1}^n y_i \alpha_i \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix}.$$

Since it can readily be verified that $(\check{\beta}_0^\lambda(\boldsymbol{\alpha}), \check{\boldsymbol{\beta}}^\lambda(\boldsymbol{\alpha}))$ lies in the interior of $(\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^{p+1}$ we therefore have

$$\begin{aligned}
&\max_{\beta_0 \in \Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda}, \boldsymbol{\beta} \in (\Theta^{\boldsymbol{X}}_{\boldsymbol{\beta},\lambda})^p} n^{-1} \sum_{i=1}^n \left\{ y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})(1-\alpha_i) \right\} - \frac{\lambda}{2} \sum_{j=0}^p \beta_j^2 \\
&\qquad\qquad = \frac{\lambda}{2} \sum_{j=0}^p \{\check{\beta}_j^\lambda(\boldsymbol{\alpha})\}^2 = \frac{1}{2\lambda n^2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^\top \boldsymbol{x}_j + 1).
\end{aligned}$$

This concludes the proof of the lemma. □

Recall the maximization problem defined in (2) in the manuscript, namely,

$$\max_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}\left(l_n^\lambda(\beta_0,\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \log\left[\frac{1}{1+\exp\{-y_i(\beta_0+\boldsymbol{x}_i^\top\boldsymbol{\beta})\}}\right]- \frac{\lambda}{2}\left[\Big(\beta_0+\sum_{j=1}^p\beta_j\mu_{n,j}\Big)^2+\sum_{j=1}^p\beta_j^2 s_{n,j}^2\right]\right),$$

and recall from the manuscript that its solution is denoted by $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$. Also, recall the definition of $J^\lambda(\boldsymbol{\alpha})$ in (3) in the manuscript, i.e., that

$$J^\lambda(\boldsymbol{\alpha}) = \frac{1}{2\lambda n^2}\boldsymbol{\alpha}^\top \operatorname{diag}(\boldsymbol{y})\Big(\sum_{k=1}^K \boldsymbol{\mathcal{K}}^{(k)} + \mathbf{1}_n\mathbf{1}_n^\top\Big)\operatorname{diag}(\boldsymbol{y})\boldsymbol{\alpha} + \frac{1}{n}\sum_{i=1}^n\Big\{(1-\alpha_i)\log(1-\alpha_i)+\alpha_i\log(\alpha_i)\Big\},$$

where $\boldsymbol{\mathcal{K}}^{(k)} = \boldsymbol{X}_{\text{cs}}^{(k)}(\boldsymbol{X}_{\text{cs}}^{(k)})^\top$.

The following proposition proves the assertion in Methods 4.2.2 that the solution $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ can be computed using Equation (11) in the manuscript.

Proposition S1. *Assume that $s_{n,j} > 0$ for all $j \in \{1,\dots,p\}$. Then, $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ satisfies*

$$\widehat{\beta}_0^\lambda = \frac{1}{n\lambda}\sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i - \sum_{j=1}^p \mu_{n,j}\widehat{\beta}_j^\lambda,$$

$$\widehat{\boldsymbol{\beta}}^\lambda = \operatorname{diag}(s_{n,1},\dots,s_{n,p})^{-1}\left(\frac{1}{n\lambda}\sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i \boldsymbol{x}_{i,cs}\right),$$

where $\widehat{\boldsymbol{\alpha}}^\lambda := [\widehat{\alpha}_1^\lambda,\dots,\widehat{\alpha}_n^\lambda]^\top \in (0,1)^n$ is the unique minimizer of $J^\lambda(\boldsymbol{\alpha})$ over $(0,1)^n$.

Proof. Since, for all $i \in \{1, \ldots, n\}$ we have

$$\begin{aligned}
\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} &= \beta_0 + \sum_{j=1}^p x_{ij}\beta_j \\
&= \left(\beta_0 + \sum_{j=1}^p \mu_{n,j}\beta_j\right) + \sum_{j=1}^p \left(\frac{x_{ij} - \mu_{n,j}}{s_{n,j}}\right)(\beta_j s_{n,j}) \\
&= \left(\beta_0 + \widehat{\boldsymbol{\mu}}^\top \boldsymbol{\beta}\right) + \boldsymbol{x}_{i,\mathrm{cs}}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}\,,
\end{aligned}$$

where we have introduced

$$\widehat{\boldsymbol{\mu}} = [\mu_{n,1}, \ldots, \mu_{n,p}]^\top\,, \qquad \widehat{\boldsymbol{\Sigma}} = \mathrm{diag}([s_{n,1}, \ldots, s_{n,p}]^\top)\,, \tag{S7}$$

it follows upon adopting the re-parametrization $(\beta_0^\circ, \boldsymbol{\beta}^\circ) \equiv (\beta_0 + \widehat{\boldsymbol{\mu}}^\top \boldsymbol{\beta}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta})$ that

$$\begin{aligned}
&\max_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} l_n^\lambda(\beta_0, \boldsymbol{\beta}) \\
&= \max_{\beta_0^\circ \in \mathbb{R}, \boldsymbol{\beta}^\circ \in \mathbb{R}^p} \left(n^{-1} \sum_{i=1}^n \log\left[\frac{1}{1 + \exp\{-y_i(\beta_0^\circ + \boldsymbol{x}_{i,\mathrm{cs}}^\top \boldsymbol{\beta}^\circ)\}}\right] - \frac{\lambda}{2}\sum_{j=0}^p (\beta_j^\circ)^2\right).
\end{aligned}$$

The maximization problem on the last line of the previous equation fits the framework of Lemma S1, which implies that its unique maximizer, denoted by $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$, satisfies

$$\begin{bmatrix} \check{\beta}_0^\lambda \\ \check{\boldsymbol{\beta}}^\lambda \end{bmatrix} = (n\lambda)^{-1} \sum_{i=1}^n y_i \widehat{\alpha}_i^\lambda \begin{bmatrix} 1 \\ \boldsymbol{x}_{i,\mathrm{cs}} \end{bmatrix}\,,$$

where $\widehat{\boldsymbol{\alpha}}^\lambda = (\widehat{\alpha}_1^\lambda, \ldots, \widehat{\alpha}_n^\lambda)^\top \in (0,1)^n$ is the unique solution to the following minimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha} \in (0,1)^n} \left(\frac{1}{n}\sum_{i=1}^n \Big\{(1 - \alpha_i)\log(1 - \alpha_i) + \alpha_i \log(\alpha_i)\Big\}\right. \\
&\qquad\qquad\qquad\qquad \left. + \frac{1}{2\lambda n^2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_{i,\mathrm{cs}}^\top \boldsymbol{x}_{j,\mathrm{cs}} + 1)\right) \\
&= \min_{\boldsymbol{\alpha} \in (0,1)^n} J^\lambda(\boldsymbol{\alpha})\,,
\end{aligned}$$

where, in applying Lemma S1, we replaced the set $\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda}$ by $(0,1)$.

The proof of the proposition follows from the fact that since the bijective nature of the reparametrization implies

$$\begin{bmatrix} \widehat{\beta}_0^{\lambda} \\ \widehat{\boldsymbol{\beta}}^{\lambda} \end{bmatrix} = \begin{bmatrix} \check{\beta}_0^{\lambda} - \widehat{\boldsymbol{\mu}}^{\top}\widehat{\boldsymbol{\beta}}^{\lambda} \\ \widehat{\boldsymbol{\Sigma}}^{-1}\check{\boldsymbol{\beta}}^{\lambda} \end{bmatrix}, \tag{S8}$$

we have

$$\widehat{\beta}_0^{\lambda} = (n\lambda)^{-1}\sum_{i=1}^{n} y_i\widehat{\alpha}_i^{\lambda} - \widehat{\boldsymbol{\mu}}^{\top}\widehat{\boldsymbol{\beta}}^{\lambda}, \qquad \widehat{\boldsymbol{\beta}}^{\lambda} = \widehat{\boldsymbol{\Sigma}}^{-1}\left((n\lambda)^{-1}\sum_{i=1}^{n} y_i\widehat{\alpha}_i^{\lambda}\boldsymbol{x}_{i,\mathrm{cs}}\right).$$

□

The following lemma establishes that, if the unpenalized maximum likelihood estimate $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ exists and is unique, then it is close to the penalized estimate $(\check{\beta}_0^{\lambda}, \check{\boldsymbol{\beta}}^{\lambda})$ defined in Lemma S1 for sufficiently small λ. It is well known [3] that if the columns of the matrix $\boldsymbol{X}$ are linearly independent, and also linearly independent of the vector $\mathbf{1}_n$, then the Hessian $\nabla^2_{\beta_0,\boldsymbol{\beta}}\ell_n(\beta_0,\boldsymbol{\beta})$ is strictly negative definite, implying that the log-likelihood function $\ell_n(\beta_0,\boldsymbol{\beta})$ is strictly concave. In this case, if a maximizer exists for the problem $\max_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}\ell_n(\beta_0,\boldsymbol{\beta})$ then it must be unique, and it must be a stationary point of $\nabla_{\beta_0,\boldsymbol{\beta}}\ell_n(\beta_0,\boldsymbol{\beta})$. The existence of such a solution is guaranteed when the response vector $\boldsymbol{y}$ is not separable [4]. Specifically, $\boldsymbol{y}$ is said to be separable if there exists $(\beta_0,\boldsymbol{\beta})$ such that $y_i(\beta_0 + \boldsymbol{x}_i^{\top}\boldsymbol{\beta}) > 0$ for all $i \in \{1,\ldots,n\}$. In the presence of separability, the log-likelihood function increases indefinitely, and a finite maximum likelihood estimate does not exist.

In what follows, for any positive definite matrix $\boldsymbol{A}$, let $\iota_{\min}(\boldsymbol{A})$ denote its smallest eigen value.

Lemma S2. *Let $(y_1,\boldsymbol{x}_1),\ldots,(y_n,\boldsymbol{x}_n) \in \{-1,1\}\times\mathbb{R}^p$ be such that the matrix $[\mathbf{1}_n,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]^{\top}$ has full column rank and such that $\boldsymbol{y}$ is not separable. Then, the unique solution $(\check{\beta}_0^{\lambda},\check{\boldsymbol{\beta}}^{\lambda})$ to the maximization problem $\max_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}\check{l}_n^{\lambda}(\beta_0,\boldsymbol{\beta})$, with $\check{l}_n^{\lambda}$ as defined in Lemma S1 , satisfies*

$$\left\|\begin{bmatrix} \check{\beta}_0^{\lambda} - \widehat{\beta}_0 \\ \check{\boldsymbol{\beta}}^{\lambda} - \widehat{\boldsymbol{\beta}} \end{bmatrix}\right\|_{\infty} \\ \le \iota_{\min}\{\boldsymbol{\mathcal{I}}(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}}) + \lambda\boldsymbol{I}_{p+1}\}^{-1}$$

$$\times \left\{ \left\| \begin{bmatrix} \check{\beta}_0^\lambda - \widehat{\beta}_0 \\ \check{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty^2 n^{-1} \sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 1)^3 + \lambda \left\| \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \right\}.$$

Proof of Lemma S2. Since $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ is a stationary point of $\check{l}_n^\lambda(\beta_0, \boldsymbol{\beta})$, we have $0 = \nabla_{\beta_0,\boldsymbol{\beta}} \check{l}_n^\lambda(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda) = \nabla_{\beta_0,\boldsymbol{\beta}} \ell_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda) - \lambda[\check{\beta}_0^\lambda, (\check{\boldsymbol{\beta}}^\lambda)^\top]^\top$, and therefore

$$\|\nabla_{\beta_0,\boldsymbol{\beta}} \ell_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)\|_2^2 = \lambda^2 \sum_{j=0}^p (\check{\beta}_j^\lambda)^2 \leq \lambda^2 \sum_{j=0}^p (\widehat{\beta}_j)^2 .$$

Since $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ does not depend on λ and is finite, we conclude that as $\lambda \to 0$, $\|\nabla_{\beta_0,\boldsymbol{\beta}} \ell_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)\|_2^2 \to 0$, which implies that $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda) \to (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ as $\lambda \to 0$ since the maximum of $\ell_n(\beta_0, \boldsymbol{\beta})$ is unique.

Now since $\nabla_{\beta_0,\boldsymbol{\beta}} \check{l}_n^\lambda(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda) = 0$ we have

$$\begin{aligned} \lambda \begin{bmatrix} \check{\beta}_0^\lambda \\ \check{\boldsymbol{\beta}}^\lambda \end{bmatrix} &= n^{-1} \sum_{i=1}^n \frac{y_i}{1 + \exp\{y_i(\check{\beta}_0^\lambda + \boldsymbol{x}_i^\top \check{\boldsymbol{\beta}}^\lambda)\}} \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix} \\ &= n^{-1} \sum_{i=1}^n y_i \Big[\frac{1}{1 + \exp\{y_i(\check{\beta}_0^\lambda + \boldsymbol{x}_i^\top \check{\boldsymbol{\beta}}^\lambda)\}} \\ &\qquad - \frac{1}{1 + \exp\{y_i(\widehat{\beta}_0 + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})\}} \Big] \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix} , \end{aligned} \tag{S9}$$

where, to obtain the second line, we used the fact that $\nabla_{\beta_0,\boldsymbol{\beta}} \ell_n(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = 0$.

As for any $x, y \in \mathbb{R}$ a Taylor expansion of order two shows that

$$\begin{aligned} &\Big| \frac{1}{1+e^x} - \frac{1}{1+e^y} + \frac{e^y}{(1+e^y)^2}(x-y) \Big| \\ &\qquad \leq (x-y)^2 \sup_{z \in \mathbb{R}} \Big| \Big\{ \frac{e^z}{(1+e^z)^2} \Big\} \Big(\frac{e^z - 1}{1+e^z} \Big) \Big| \leq (x-y)^2 , \end{aligned}$$

and since (S9) implies

$$\begin{aligned} &\lambda \begin{bmatrix} \check{\beta}_0^\lambda \\ \check{\boldsymbol{\beta}}^\lambda \end{bmatrix} + n^{-1} \sum_{i=1}^n \frac{\exp\{y_i(\widehat{\beta}_0 + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})\}}{[1 + \exp\{y_i(\widehat{\beta}_0 + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})\}]^2} \begin{bmatrix} 1 & \boldsymbol{x}_i^\top \\ \boldsymbol{x}_i & \boldsymbol{x}_i \boldsymbol{x}_i^\top \end{bmatrix} \begin{bmatrix} \check{\beta}_0^\lambda - \widehat{\beta}_0 \\ \check{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \\ &= \lambda \begin{bmatrix} \check{\beta}_0^\lambda \\ \check{\boldsymbol{\beta}}^\lambda \end{bmatrix} \end{aligned}$$

$$+ n^{-1}\sum_{i=1}^{n}\frac{\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}}{[1+\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}]^2}\Big[(\check{\beta}_0^\lambda-\widehat{\beta}_0)+\boldsymbol{x}_i^\top(\check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}})\Big]\begin{bmatrix}1\\ \boldsymbol{x}_i\end{bmatrix}$$

$$= n^{-1}\sum_{i=1}^{n} y_i\Bigg\{\frac{1}{1+\exp\{y_i(\check{\beta}_0^\lambda+\boldsymbol{x}_i^\top\check{\boldsymbol{\beta}}^\lambda)\}}-\frac{1}{1+\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}} + \frac{\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}}{[1+\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}]^2}\Big[y_i\{(\check{\beta}_0^\lambda-\widehat{\beta}_0)+\boldsymbol{x}_i^\top(\check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}})\}\Big]\Bigg\}\begin{bmatrix}1\\ \boldsymbol{x}_i\end{bmatrix},$$

we conclude that

$$\left\|\lambda\begin{bmatrix}\check{\beta}_0^\lambda\\ \check{\boldsymbol{\beta}}^\lambda\end{bmatrix}+n^{-1}\sum_{i=1}^{n}\frac{\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}}{[1+\exp\{y_i(\widehat{\beta}_0+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})\}]^2}\begin{bmatrix}1 & \boldsymbol{x}_i^\top\\ \boldsymbol{x}_i & \boldsymbol{x}_i\boldsymbol{x}_i^\top\end{bmatrix}\begin{bmatrix}\check{\beta}_0^\lambda-\widehat{\beta}_0\\ \check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\end{bmatrix}\right\|_\infty$$

$$\leq n^{-1}\sum_{i=1}^{n}\Big[(\check{\beta}_0^\lambda-\widehat{\beta}_0)+\boldsymbol{x}_i^\top(\check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}})\Big]^2(\|\boldsymbol{x}_i\|_\infty+1)$$

$$\leq \left\|\begin{bmatrix}\check{\beta}_0^\lambda-\widehat{\beta}_0\\ \check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\end{bmatrix}\right\|_\infty^2 n^{-1}\sum_{i=1}^{n}(\|\boldsymbol{x}_i\|_\infty+1)^3.$$

Recalling the definition of $\boldsymbol{\mathcal{I}}(\beta_0,\boldsymbol{\beta})$ above the statement of the lemma, by rearranging the terms and applying the triangle inequality, we obtain from the last equation that

$$\left\|\Big(\boldsymbol{\mathcal{I}}(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}})+\lambda\boldsymbol{I}_{p+1}\Big)\begin{bmatrix}\check{\beta}_0^\lambda-\widehat{\beta}_0\\ \check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\end{bmatrix}\right\|_\infty \leq \left\|\begin{bmatrix}\check{\beta}_0^\lambda-\widehat{\beta}_0\\ \check{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\end{bmatrix}\right\|_\infty^2 n^{-1}\sum_{i=1}^{n}(\|\boldsymbol{x}_i\|_\infty+1)^3+\lambda\left\|\begin{bmatrix}\widehat{\beta}_0\\ \widehat{\boldsymbol{\beta}}\end{bmatrix}\right\|_\infty.$$

The result follows from the fact that for any positive matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\boldsymbol{x}\|\geq \iota_{\min}(\boldsymbol{A})\|\boldsymbol{x}\|$. □

Remark S1. *Consider $(\widehat{\beta}_0^\lambda,\widehat{\boldsymbol{\beta}}^\lambda)$, and let $(\check{\beta}_0^\lambda,\check{\boldsymbol{\beta}}^\lambda)$ be defined as in the proof of Proposition S1. Also, let $(\check{\beta}_0,\check{\boldsymbol{\beta}})$ and $(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}})$ be the maximum likelihood estimates that solves the log-likelihood maximization problem based on, respectively, the centered and scaled covariate data $\boldsymbol{x}_{1,cs},\ldots,\boldsymbol{x}_{n,cs}$, and the covariate data in their original scale $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$. Using arguments that are identical to the ones used in the proof of Proposition S1, it is* *straightforward to deduce that*

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \check{\beta}_0 - \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\Sigma}}^{-1} \check{\boldsymbol{\beta}} \end{bmatrix} .$$

Combining the last equation to (S8) *implies*

$$\begin{aligned} \max_{j \in \{0,\ldots,p\}} |\widehat{\beta}_j^\lambda - \widehat{\beta}_j| &\leq \left(1 + \frac{p(\|\widehat{\boldsymbol{\mu}}\|_\infty + 1)}{\min_{j\in\{1,\ldots,p\}} s_{n,j}}\right) \max_{j \in \{0,\ldots,p\}} |\check{\beta}_j^\lambda - \check{\beta}_j| \\ &:= \kappa_{\boldsymbol{X}} \max_{j \in \{0,\ldots,p\}} |\check{\beta}_j^\lambda - \check{\beta}_j| \, . \end{aligned}$$

Since $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ *and* $(\check{\beta}_0, \check{\boldsymbol{\beta}})$ *fit the framework of Lemma* S2*, we have, under the lemma's assumptions that*

$$\begin{aligned} &\left\| \begin{bmatrix} \check{\beta}_0^\lambda - \check{\beta}_0 \\ \check{\boldsymbol{\beta}}^\lambda - \check{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \\ &\leq \iota_{\min}\{\check{\boldsymbol{\mathcal{I}}}(\check{\beta}_0, \check{\boldsymbol{\beta}}) + \lambda \boldsymbol{I}_{p+1}\}^{-1} \\ &\qquad \times \left\{ \left\| \begin{bmatrix} \check{\beta}_0^\lambda - \check{\beta}_0 \\ \check{\boldsymbol{\beta}}^\lambda - \check{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty^2 n^{-1} \sum_{i=1}^n (\|\boldsymbol{x}_{i,cs}\|_\infty + 1)^3 + \lambda \left\| \begin{bmatrix} \check{\beta}_0 \\ \check{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \right\}, \end{aligned}$$

with $\check{\boldsymbol{\mathcal{I}}}(\beta_0, \boldsymbol{\beta})$ *a version of* $\boldsymbol{\mathcal{I}}(\beta_0, \boldsymbol{\beta})$ *(see* (S1)*) where the original data are replaced by the centered and scaled data, given by*

$$\check{\boldsymbol{\mathcal{I}}}(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \frac{\exp\{y_i(\beta_0 + \boldsymbol{x}_{i,cs}^\top \boldsymbol{\beta})\}}{[1 + \exp\{y_i(\beta_0 + \boldsymbol{x}_{i,cs}^\top \boldsymbol{\beta})\}]^2} \begin{bmatrix} 1 & \boldsymbol{x}_{i,cs}^\top \\ \boldsymbol{x}_{i,cs} & \boldsymbol{x}_{i,cs}\boldsymbol{x}_{i,cs}^\top \end{bmatrix} .$$

Since the inequality $\|\boldsymbol{x}_{i,cs}\|_\infty \leq \kappa_{\boldsymbol{X}} \|\boldsymbol{x}\|_i + \kappa_{\boldsymbol{X}}$ *implies* $n^{-1}\sum_{i=1}^n (\|\boldsymbol{x}_{i,cs}\|_\infty + 1)^3 \leq \kappa_{\boldsymbol{X}}^3 n^{-1} \sum_{i=1}^n (\|\boldsymbol{x}_i\|_\infty + 2)^3$*, and as* $\boldsymbol{T}_n \boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \boldsymbol{T}_n^\top = \check{\boldsymbol{\mathcal{I}}}(\check{\beta}_0, \check{\boldsymbol{\beta}})$*, with*

$$\boldsymbol{T}_n = \begin{bmatrix} 1 & \boldsymbol{0}^\top \\ -\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}} & \widehat{\boldsymbol{\Sigma}}^{-1} \end{bmatrix} , \quad \text{which ensures} \quad \boldsymbol{T}_n \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix} = \begin{bmatrix} 1 \\ \boldsymbol{x}_{i,cs} \end{bmatrix} ,$$

then, we conclude that

$$\left\| \begin{bmatrix} \widehat{\beta}_0^\lambda - \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \leq \kappa_{\boldsymbol{X}}^5 \iota_{\min}\{\boldsymbol{T}_n \boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \boldsymbol{T}_n^\top + \lambda \boldsymbol{I}_{p+1}\}^{-1}$$

$$\times \left\{ \left\| \left(n^{-1}\sum_{i=1}^{n}(\|\boldsymbol{x}_i\|_\infty + 2)^3\right) \begin{bmatrix} \widehat{\beta}_0^\lambda - \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty^2 + \lambda \left\| \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \right\}.$$

Supplementary Methods 2 Auxiliary results related to the asymptotic normality of $\widehat{\boldsymbol{\beta}}^\lambda$ and computation of standard errors

2.1 Asymptotic normality of $\widehat{\boldsymbol{\beta}}^\lambda$ and consistency of standard error estimates

Recall from Results 2.1 that we assume a binary random variable $Y \in \{-1, 1\}$ and a random vector of covariates $\mathbf{X} = [X_1, \ldots, X_p]^\top \in \mathbb{R}^p$ following a logistic regression model. In this model, there exists an unknown parameter vector $\beta_{0\star} \in \mathbb{R}, \boldsymbol{\beta}_\star \in \mathbb{R}^p$ such that

$$\mathbb{P}(Y = y \mid \mathbf{X} = \boldsymbol{x}) = \frac{1}{1 + \exp\{-y(\beta_{0\star} + \boldsymbol{x}^\top \boldsymbol{\beta}_\star)\}}. \tag{S10}$$

Let $(\mathrm{Y}_1, \mathbf{X}_1), \ldots, (\mathrm{Y}_n, \mathbf{X}_n)$ be i.i.d. random variables satisfying the model in (S10). Throughout this section, we use $\ell_n(\boldsymbol{\beta})$ and $l_n^\lambda(\boldsymbol{\beta})$, as defined in (9) and (2), respectively, where the $(y_i, \boldsymbol{x}_i)$'s in their definitions are replaced here by the random variables $(\mathrm{Y}_i, \mathbf{X}_i)$. Specifically, we consider

$$\ell_n(\beta_0, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \log\left(\frac{1}{1 + \exp\{-\mathrm{Y}_i(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta})\}}\right). \tag{S11}$$

and

$$l_n^\lambda(\beta_0, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \log\left(\frac{1}{1 + \exp\{-\mathrm{Y}_i(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta})\}}\right) - \frac{\lambda}{2}\left[\left(\beta_0 + \sum_{j=1}^{p}\beta_j \mu_{n,j}\right)^2 + \sum_{j=1}^{p}\beta_j^2 s_{n,j}^2\right], \tag{S12}$$

with $\mu_{n,j} = n^{-1}\sum_{i=1}^{n} \mathrm{X}_{ij}$ and $s_{n,j}^2 = (n-1)^{-1}\sum_{i=1}^{n}(\mathrm{X}_{ij} - \mu_{n,j})^2$.

Throughout this section, we also consider a version of $\widehat{\boldsymbol{\beta}}^\lambda$ computed with the random variables $(\mathrm{Y}_1, \mathbf{X}_1), \ldots, (\mathrm{Y}_n, \mathbf{X}_n)$. That is, we define the estimator $\widehat{\boldsymbol{\beta}}^\lambda = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}} l_n^\lambda(\boldsymbol{\beta})$, with $l_n^\lambda(\boldsymbol{\beta})$ as in (S12) (recall from Lemma S1 that when $\lambda > 0$ the function $l_n^\lambda(\boldsymbol{\beta})$ is strongly concave and has a unique maximizer).

We also consider a version of $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ computed from the random variables $(\mathrm{Y}_1, \mathbf{X}_1), \ldots, (\mathrm{Y}_n, \mathbf{X}_n)$. That is, we define the maximum likelihood estimator as $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \arg\max_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p} \ell_n(\beta_0, \boldsymbol{\beta})$. For sufficiently large n, this estimator exists with probability one, since for any finite $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$, the response vector $\boldsymbol{Y}$ will be non-separable with probability one when n is large enough.

Likewise, we consider of version of $\boldsymbol{\mathcal{I}}(\beta_0, \boldsymbol{\beta})$ at (S1), where the $(y_i, \boldsymbol{x}_i)$'s are replaced here by the random variables $(\mathrm{Y}_i, \mathbf{X}_i)$.

The following lemma establishes that, if $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ converges in probability to the true $(\beta_{0\star}, \boldsymbol{\beta}_\star)$, then $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) + O_{\mathbb{P}}(\lambda)$.

Lemma S3. *Let* $(\mathrm{Y}_1, \mathbf{X}_1), \ldots, (\mathrm{Y}_n, \mathbf{X}_n)$ *be i.i.d. random variables satisfying the model in* (S10)*. Assume that the matrix*

$$\mathbb{E}\left(\begin{bmatrix} 1 & \mathbf{X}_1^\top \\ \mathbf{X}_1 & \mathbf{X}_1\mathbf{X}_1^\top \end{bmatrix}\right)$$

is invertible, and that $\mathbb{E}\{\|\mathbf{X}_1\|_\infty^2\} < \infty$*. Then, if* $\lambda \to 0$ *and* $\max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty) = o_{\mathbb{P}}(1)$ *as* $n \to \infty$*, it follows that* $\max(|\widehat{\beta}_0^\lambda - \widehat{\beta}_0|, \|\widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}}\|_\infty) = O_{\mathbb{P}}(\lambda)$ *as* $n \to \infty$*.*

Proof of Lemma S3. We start by showing that

$$\max(|\widehat{\beta}_0^\lambda - \widehat{\beta}_0|, \|\widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}}\|_\infty) = o_{\mathbb{P}}(1) \quad \text{as } n \to \infty. \tag{S13}$$

To this end, first note that under our conditions, $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ exists and is unique with probability one. Since $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ and $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ are respectively the maximizers of $l_n^\lambda(\beta_0, \boldsymbol{\beta})$ and $\ell_n(\beta_0, \boldsymbol{\beta})$, we have

$$l_n^\lambda(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) > l_n^\lambda(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \quad \text{and} \quad \ell_n(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) < \ell_n(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})\,. \tag{S14}$$

As

$$l_n^\lambda(\beta_0, \boldsymbol{\beta}) = \ell_n(\beta_0, \boldsymbol{\beta}) - \frac{\lambda}{2}\left[\left(\beta_0 + \widehat{\boldsymbol{\mu}}^\top\boldsymbol{\beta}\right)^2 + \boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}^2\boldsymbol{\beta}\right],$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are defined as in (S7) with the $\boldsymbol{x}_i$'s in the definition of the quantities $\mu_{n,j}$ and $s_{n,j}$ replaced here by the random $\mathbf{X}_i$'s here, we have

$$\begin{aligned}
0 &< l_n^\lambda(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) - l_n^\lambda(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \\
&= \{\ell_n(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) - \ell_n(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})\} \\
&\quad - \frac{\lambda}{2}\Big(\big[\big(\widehat{\beta}_0^\lambda + \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}^\lambda\big)^2 + (\widehat{\boldsymbol{\beta}}^\lambda)^\top \widehat{\boldsymbol{\Sigma}}^2 \widehat{\boldsymbol{\beta}}^\lambda\big] - \big[\big(\widehat{\beta}_0 + \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}\big)^2 + \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}^2 \widehat{\boldsymbol{\beta}}\big] \Big).
\end{aligned}$$

This implies that

$$\begin{aligned}
&\frac{\lambda}{2}\Big(\big[\big(\widehat{\beta}_0 + \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}\big)^2 + \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}^2 \widehat{\boldsymbol{\beta}}\big] - \big[\big(\widehat{\beta}_0^\lambda + \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}^\lambda\big)^2 + (\widehat{\boldsymbol{\beta}}^\lambda)^\top \widehat{\boldsymbol{\Sigma}}^2 \widehat{\boldsymbol{\beta}}^\lambda\big] \Big) \\
&\qquad\qquad > \ell_n(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) - \ell_n(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) > 0\,.
\end{aligned}$$

Since we have assumed that $\max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty) = o_{\mathbb{P}}(1)$ as $n \to \infty$, as under our conditions the weak law of large numbers ensures $\mu_{n,j} = \mathbb{E}(\mathrm{X}_j) + o_{\mathbb{P}}(1)$, and since the weak law of large numbers combined with the continuous mapping theorem implies $s_{n,j} = \sqrt{\mathrm{Var}(\mathrm{X}_j)} + o_{\mathbb{P}}(1)$, we conclude from the last equation display that as $n \to \infty$,

$$\big[\big(\widehat{\beta}_0^\lambda + \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}^\lambda\big)^2 + (\widehat{\boldsymbol{\beta}}^\lambda)^\top \widehat{\boldsymbol{\Sigma}}^2 \widehat{\boldsymbol{\beta}}^\lambda\big] \le \big[\big(\beta_{0\star} + \boldsymbol{\mu}^\top \boldsymbol{\beta}_\star\big)^2 + (\boldsymbol{\beta}_\star)^\top \boldsymbol{\Sigma}^2 \boldsymbol{\beta}_\star\big] + o_{\mathbb{P}}(1)\,,$$

where $\boldsymbol{\mu} = [\mathbb{E}(\mathrm{X}_1), \ldots, \mathbb{E}(\mathrm{X}_p)]^\top$ and $\boldsymbol{\Sigma} = \mathrm{diag}([\sqrt{\mathrm{Var}(\mathrm{X}_1)}, \ldots, \sqrt{\mathrm{Var}(\mathrm{X}_p)}]^\top)$.

As $\lambda \to 0$ when $n \to \infty$, and since $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ is bounded in probability as $n \to \infty$, we conclude from (S14) that $\ell_n(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) = l_n^\lambda(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) + o_{\mathbb{P}}(1) \ge l_n^\lambda(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) + o_{\mathbb{P}}(1) = \ell_n(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) + o_{\mathbb{P}}(1)$. Therefore, $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ is a near-maximizer of ℓ_n (see e.g. [5] chapter 5), and we conclude that (S13) holds.

Next we show that the term on the right-hand side of the equality at (S13) can be replaced by $O_{\mathbb{P}}(\lambda)$. To this end, note that we have from Remark S1 that

$$\begin{aligned}
\left\| \begin{bmatrix} \widehat{\beta}_0^\lambda - \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty &\le \kappa_{\boldsymbol{X}}^5 \iota_{\min}\{\boldsymbol{T}_n \boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \boldsymbol{T}_n^\top + \lambda \boldsymbol{I}_{p+1}\}^{-1} \\
&\quad \times \left\{ \left\| \Big(n^{-1} \sum_{i=1}^n (\|\mathbf{X}_i\|_\infty + 2)^3\Big) \begin{bmatrix} \hat{\beta}_0^\lambda - \hat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty^2 + \lambda \left\| \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} \right\|_\infty \right\},
\end{aligned}$$

where

$$\kappa_{\boldsymbol{X}} = \left(1 + \frac{p(\|\widehat{\boldsymbol{\mu}}\|_\infty + 1)}{\min_{j\in\{1,\dots,p\}} s_{n,j}}\right), \quad \text{and} \quad \boldsymbol{T}_n = \begin{bmatrix} 1 & \mathbf{0}^\top \\ -\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}} & \widehat{\boldsymbol{\Sigma}}^{-1} \end{bmatrix}.$$

As for any $x, y \in \mathbb{R}$ the mean value theorem ensures

$$\left|\frac{e^x}{(1+e^x)^2} - \frac{e^y}{(1+e^y)^2}\right| \le |x-y| \sup_{z\in\mathbb{R}} \left|\left\{\frac{e^z}{(1+e^z)^2}\right\}\left(\frac{e^z-1}{1+e^z}\right)\right| \le |x-y|,$$

and since

$$\begin{aligned}
&\boldsymbol{T}_n\{\boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) - \boldsymbol{\mathcal{I}}(\beta_{0\star}, \boldsymbol{\beta}_\star)\}\boldsymbol{T}_n^\top \\
&= \frac{1}{n}\sum_{i=1}^n \left\{\frac{\exp\{\mathrm{Y}_i(\widehat{\beta}_0 + \mathbf{X}_i^\top\widehat{\boldsymbol{\beta}})\}}{[1+\exp\{\mathrm{Y}_i(\widehat{\beta}_0 + \mathbf{X}_i^\top\widehat{\boldsymbol{\beta}})\}]^2} - \frac{\exp\{\mathrm{Y}_i(\beta_{0\star} + \mathbf{X}_i^\top\boldsymbol{\beta}_\star)\}}{[1+\exp\{\mathrm{Y}_i(\beta_{0\star} + \mathbf{X}_i^\top\boldsymbol{\beta}_\star)\}]^2}\right\} \\
&\qquad \times \begin{bmatrix} 1 & \mathbf{X}_{i,\mathrm{cs}}^\top \\ \mathbf{X}_{i,\mathrm{cs}} & \mathbf{X}_{i,\mathrm{cs}}\mathbf{X}_{i,\mathrm{cs}}^\top \end{bmatrix},
\end{aligned}$$

we deduce that

$$\begin{aligned}
&\left\|\boldsymbol{T}_n\boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})\boldsymbol{T}_n^\top - \boldsymbol{T}_n\boldsymbol{\mathcal{I}}(\beta_{0\star}, \boldsymbol{\beta}_\star)\boldsymbol{T}_n^\top\right\|_\infty \\
&\le \frac{1}{n}\sum_{i=1}^n \left|\mathrm{Y}_i\big\{(\widehat{\beta}_0 - \beta_{0\star}) + \mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\big\}\right|(1 + \|\mathbf{X}_{i,\mathrm{cs}}\|_\infty^2) \\
&\le \max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty)\frac{1}{n}\sum_{i=1}^n(1 + \|\mathbf{X}_i\|_\infty)(1 + \|\mathbf{X}_{i,\mathrm{cs}}\|_\infty^2) \\
&\le \max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty) \\
&\qquad \times \left(1 + p + \frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty + \frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty\|\mathbf{X}_{i,\mathrm{cs}}\|_\infty^2\right) \\
&\le \max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty) \\
&\qquad \left(1 + p + \frac{1+\kappa_{\mathbf{X}}^2}{n}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty + \frac{1}{n\,\min_{j\in\{1,\dots,p\}} s_{n,j}^2}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty^3\right).
\end{aligned}$$

Since $n^{-1}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty \le \sum_{j=1}^p (n^{-1}\sum_{i=1}^n |\mathrm{X}_{ij}|)$ and $n^{-1}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty^3 \le \sum_{j=1}^p (n^{-1}\sum_{i=1}^n |\mathrm{X}_{ij}|^3)$, and as p is finite, we obtain from the weak law of large numbers that $n^{-1}\sum_{i=1}^n \|\mathbf{X}_i\|_\infty = O_{\mathbb{P}}(1)$ and

$n^{-1}\sum_{i=1}^{n}\|\mathbf{X}_i\|_\infty^3 = O_{\mathbb{P}}(1)$ (recall that we have assumed $\mathbb{E}(|\mathrm{X}_{ij}|^3) < \infty$ for all $j \in \{1,\dots,p\}$). Furthermore, as we have established above that $\mu_{n,j} = \mathbb{E}(\mathrm{X}_j) + o_{\mathbb{P}}(1)$ and that $s_{n,j} = \sqrt{\mathrm{Var}(\mathrm{X}_j)} + o_{\mathbb{P}}(1)$, then, as p is finite, we get $\kappa_{\mathbf{X}} = O_{\mathbb{P}}(1)$. Therefore, we conclude from the previous equation that

$$\left\|\boldsymbol{T}_n\boldsymbol{\mathcal{I}}(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}})\boldsymbol{T}_n^\top - \boldsymbol{T}_n\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\boldsymbol{T}_n^\top\right\|_\infty = O_{\mathbb{P}}\Big\{\max(|\widehat{\beta}_0-\beta_{0\star}|, \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_\star\|_\infty)\Big\} = o_{\mathbb{P}}(1)\,,$$

where the last equality followed from the fact that we have assumed $\max(|\widehat{\beta}_0-\beta_{0\star}|, \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_\star\|_\infty) = o_{\mathbb{P}}(1)$ as $n\to\infty$.

As the weak law of large numbers ensures, under our assumptions, that $\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star) = \mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\} + o_{\mathbb{P}}(1)$ as $n\to\infty$, and since $\boldsymbol{T}_n = \boldsymbol{T} + o_{\mathbb{P}}(1)$ with

$$\boldsymbol{T} = \begin{bmatrix} 1 & \mathbf{0}^\top \\ -\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} & \boldsymbol{\Sigma}^{-1}\end{bmatrix}\,,$$

the assumption that

$$\mathbb{E}\left(\begin{bmatrix} 1 & \mathbf{X}_1^\top \\ \mathbf{X}_1 & \mathbf{X}_1\mathbf{X}_1^\top\end{bmatrix}\right)$$

is invertible implies $\iota_{\min}\{\boldsymbol{T}\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\}\boldsymbol{T}^\top + \lambda\boldsymbol{I}_{p+1}\} \geq \iota_{\min}\{\boldsymbol{T}\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\}\boldsymbol{T}^\top\} > 0$. Therefore, we conclude from Slutsky's lemma and the continuous mapping theorem that as $n\to\infty$,

$$\iota_{\min}\{\boldsymbol{T}_n\boldsymbol{\mathcal{I}}(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}})\boldsymbol{T}_n^\top + \lambda\boldsymbol{I}_{p+1}\}^{-1} = \iota_{\min}[\boldsymbol{T}\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\}\boldsymbol{T}^\top]^{-1} + o_{\mathbb{P}}(1)\,.$$

Therefore, as $n\to\infty$,

$$\begin{aligned}&\max(|\widehat{\beta}_0^\lambda-\widehat{\beta}_0|, \|\widehat{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\|_\infty)\\ &\leq \Big(\iota_{\min}\{\boldsymbol{T}\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star},\boldsymbol{\beta}_\star)\}\boldsymbol{T}^\top\}^{-1} + o_{\mathbb{P}}(1)\Big)\\ &\qquad\times\Big[O_{\mathbb{P}}\Big\{\max(|\widehat{\beta}_0^\lambda-\widehat{\beta}_0|, \|\widehat{\boldsymbol{\beta}}^\lambda-\widehat{\boldsymbol{\beta}}\|_\infty)^2\Big\} + \lambda\max_{0\leq j\leq p}|\widehat{\beta}_j|\Big]\,.\end{aligned}$$

The proof follows from the fact that, as we have assumed $\max(|\widehat{\beta}_0 - \beta_{0\star}|, \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_\star\|_\infty) = o_{\mathbb{P}}(1)$ as $n\to\infty$, it follows that $\max_{0\leq j\leq p}|\widehat{\beta}_j| \leq \max_{0\leq j\leq p}|\beta_{j\star}| + o_{\mathbb{P}}(1)|$. □

The preceding lemma implies that if $\lambda = o(n^{-1/2})$, and if $\sqrt{n}\{(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) - (\beta_{0\star}, \boldsymbol{\beta}_\star)\}$ is asymptotically a mean-zero normal, then the penalized estimators used in our procedure can replace the unpenalized ones without affecting the asymptotic normality result. In other words, $\sqrt{n}\{(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) - (\beta_{0\star}, \boldsymbol{\beta}\star)\}$ has the same asymptotic distribution than $\sqrt{n}\{(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) - (\beta_{0\star}, \boldsymbol{\beta}_\star)\}$.

Under our conditions, using arguments that are similar to those used in e.g. [5] chapter 5, under our conditions, the asymptotic variance-covariance matrix of $\sqrt{n}\{(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) - (\beta_{0\star}, \boldsymbol{\beta}_\star)\}$ is given by

$$
\begin{aligned}
&(\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star}, \boldsymbol{\beta}_\star)\})^{-1} \\
&\qquad = \left[\mathbb{E}\left\{\frac{\exp\{\mathrm{Y}_1(\beta_{0\star} + \mathbf{X}_1^\top \boldsymbol{\beta}_\star)\}}{[1 + \exp\{\mathrm{Y}_1(\beta_{0\star} + \mathbf{X}_i^\top \boldsymbol{\beta}_\star)\}]^2} \begin{bmatrix} 1 & \mathbf{X}_i^\top \\ \mathbf{X}_i & \mathbf{X}_i \mathbf{X}_i^\top \end{bmatrix}\right\}\right]^{-1} .
\end{aligned}
$$

Inference on $(\beta_{0\star}, \boldsymbol{\beta}_\star)$ based on the maximum likelihood estimator $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ (for example, constructing confidence intervals or performing hypothesis tests) is typically carried out by combining these estimators with their corresponding standard errors. The standard deviation of $\widehat{\beta}_j$, given by $\sqrt{\mathrm{Var}(\widehat{\beta}_j)} = n^{-1/2}[(\mathbb{E}\{\boldsymbol{\mathcal{I}}(\beta_{0\star}, \boldsymbol{\beta}_\star)\})^{-1}]_{jj}$ is commonly estimated using the following quantity:

$$
\widehat{\sigma}_j := \frac{1}{\sqrt{n}} \left(\left[\left(\boldsymbol{\mathcal{I}}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) \right)^{-1} \right]_{jj} \right)^{1/2} .
$$

Based on Lemma S3, which implies that $\max(|\widehat{\beta}_0^\lambda - \widehat{\beta}_0|, \|\widehat{\boldsymbol{\beta}}^\lambda - \widehat{\boldsymbol{\beta}}\|_\infty) = O_{\mathbb{P}}(\lambda)$ as $n \to \infty$, and using theoretical arguments similar to those employed therein, it is straightforward to show that, if $\lambda = o(n^{-1/2})$, as $n \to \infty$,

$$
\frac{1}{\sqrt{n}} \left(\left[\left(\boldsymbol{\mathcal{I}}(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda) \right)^{-1} \right]_{jj} \right)^{1/2} = \widehat{\sigma}_j \{1 + o_{\mathbb{P}}(1)\} .
$$

The proof of the consistency of our standard error computation procedure follows from the derivations provided in the following section, which show that $[(\boldsymbol{\mathcal{I}}(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda))^{-1}]_{jj} = [(\boldsymbol{\mathcal{I}}^\lambda)^{-1}]_{jj}/s_{n,j}^2$, with $\boldsymbol{\mathcal{I}}^\lambda$ defined at the beginning of Methods 4.2.3 in the manuscript.

2.2 Computation of standard errors

Let $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ to denote n independent realizations of the random pair $(\mathbf{X}, \mathrm{Y})$. We next describe the derivation of the expression for the estimates of standard errors.

As in the proof of Proposition S1, let $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ to denote the solutions of the following maximization problem:

$$\max_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \left(\frac{1}{n} \sum_{i=1}^n \log \left[\frac{1}{1 + \exp\{-y_i(\beta_0 + \boldsymbol{x}_{i,\mathrm{cs}}^\top \boldsymbol{\beta})\}} \right] + \lambda \sum_{j=0}^p \beta_j^2 \right).$$

Then, for $j \in \{1, \ldots, p\}$, one obtains from the relationship between $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ and $(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ that

$$[\{\boldsymbol{\mathcal{I}}(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)\}^{-1}]_{jj} = [\{-\nabla^2_{\beta_0, \boldsymbol{\beta}} \check{\ell}_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)\}^{-1}]_{jj} / s_{n,j}^2 ,$$

where

$$\begin{aligned} -\nabla^2_{\beta_0, \boldsymbol{\beta}} \check{\ell}_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda) &= \frac{1}{n} \sum_{i=1}^n \frac{\exp\{y_i(\check{\beta}_0^\lambda + \boldsymbol{x}_{i,\mathrm{cs}}^\top \check{\boldsymbol{\beta}}^\lambda)\}}{[1 + \exp\{y_i(\check{\beta}_0^\lambda + \boldsymbol{x}_{i,\mathrm{cs}}^\top \check{\boldsymbol{\beta}}^\lambda)\}]^2} \begin{bmatrix} 1 & \boldsymbol{x}_{i,\mathrm{cs}}^\top \\ \boldsymbol{x}_{i,\mathrm{cs}} & \boldsymbol{x}_{i,\mathrm{cs}} \boldsymbol{x}_{i,\mathrm{cs}}^\top \end{bmatrix} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\exp\{y_i(\widehat{\beta}_0^\lambda + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^\lambda)\}}{[1 + \exp\{y_i(\widehat{\beta}_0^\lambda + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^\lambda)\}]^2} \begin{bmatrix} 1 & \boldsymbol{x}_{i,\mathrm{cs}}^\top \\ \boldsymbol{x}_{i,\mathrm{cs}} & \boldsymbol{x}_{i,\mathrm{cs}} \boldsymbol{x}_{i,\mathrm{cs}}^\top \end{bmatrix} . \end{aligned}$$

Now, recall that, for each $k \in \{1, \ldots, K\}$, the vector $\widehat{\boldsymbol{c}}^{\lambda(k)}$ defined in (12) satisfies $\widehat{\boldsymbol{c}}^{\lambda(k)} = \boldsymbol{X}_{\mathrm{cs}}^{(k)} \operatorname{diag}(s_{n,1}^{(k)}, \ldots, s_{n,p^{(k)}}^{(k)}) \widehat{\boldsymbol{\beta}}^{\lambda(k)} = \boldsymbol{X}^{(k)} \widehat{\boldsymbol{\beta}}^{\lambda(k)} - (\sum_{j=1}^{p^{(k)}} \widehat{\beta}_j^{\lambda(k)} \mu_j^{(k)}) \mathbf{1}_n$, and that the response-node has access to $\widehat{\boldsymbol{c}}^{\lambda(1)}, \ldots, \widehat{\boldsymbol{c}}^{\lambda(K)}$. Since the response-node can also compute $(n\lambda)^{-1} \sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i = \widehat{\beta}_0^\lambda + \sum_{j=1}^n \widehat{\beta}_j^\lambda \mu_{n,j}$ (recall the expression given in (11)), it is therefore able to compute

$$\widehat{\beta}_0^\lambda \mathbf{1}_n + \boldsymbol{X} \widehat{\boldsymbol{\beta}}^\lambda = \widehat{\beta}_0^\lambda \mathbf{1}_n + \sum_{k=1}^K \boldsymbol{X}^{(k)} \widehat{\boldsymbol{\beta}}^{\lambda(k)} = ((n\lambda)^{-1} \sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i) \mathbf{1}_n + \sum_{k=1}^K \widehat{\boldsymbol{c}}^{\lambda(k)} .$$

Then, upon defining a version $\widehat{\boldsymbol{V}}^\lambda$ introduced above (7) with $\tilde{\boldsymbol{\alpha}}^\lambda$ there replaced by $\widehat{\boldsymbol{\alpha}}^\lambda$, whose diagonal entries $[\widehat{\boldsymbol{V}}^\lambda]_{jj}$ satisfying

$$[\widehat{\boldsymbol{V}}^\lambda]_{jj} = \frac{\exp\left[y_j\left\{(n\lambda)^{-1}\sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i + \sum_{k=1}^K \widehat{c}_j^{\lambda(k)}\right\}\right]}{\left(1+\exp\left[y_j\left\{(n\lambda)^{-1}\sum_{i=1}^n \widehat{\alpha}_i^\lambda y_i + \sum_{k=1}^K \widehat{c}_j^{\lambda(k)}\right\}\right]\right)^2}$$
$$= \frac{\exp\{y_j(\widehat{\beta}_0^\lambda + \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}^\lambda)\}}{[1+\exp\{y_j(\widehat{\beta}_0^\lambda + \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}^\lambda)\}]^2},$$

the matrix $-\nabla^2_{\beta_0,\boldsymbol{\beta}}\check{\ell}_n(\check{\beta}_0^\lambda, \check{\boldsymbol{\beta}}^\lambda)$ can be computed as

$$\boldsymbol{\mathcal{I}}^\lambda = n^{-1}\left[\mathbf{1}_n\ \boldsymbol{X}_{\text{cs}}^{(1)}\ \ldots\ \boldsymbol{X}_{\text{cs}}^{(k)}\right]^\top \widehat{\boldsymbol{V}}^\lambda \left[\mathbf{1}_n\ \boldsymbol{X}_{\text{cs}}^{(1)}\ \ldots\ \boldsymbol{X}_{\text{cs}}^{(k)}\right]$$
$$= n^{-1}\begin{bmatrix} \mathbf{1}_n^\top \widehat{\boldsymbol{V}}^\lambda \mathbf{1}_n & \mathbf{1}_n^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(1)} & \cdots & \mathbf{1}_n^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(K)} \\ (\boldsymbol{X}^{(1)})^\top \widehat{\boldsymbol{V}}^\lambda \mathbf{1}_n & (\boldsymbol{X}_{\text{cs}}^{(1)})^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(1)} & \cdots & (\boldsymbol{X}_{\text{cs}}^{(1)})^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{X}_{\text{cs}}^{(K)})^\top \widehat{\boldsymbol{V}}^\lambda \mathbf{1}_n & (\boldsymbol{X}_{\text{cs}}^{(K)})^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(1)} & \cdots & (\boldsymbol{X}_{\text{cs}}^{(K)})^\top \widehat{\boldsymbol{V}}^\lambda \boldsymbol{X}_{\text{cs}}^{(K)} \end{bmatrix}.$$

Supplementary Methods 3 Selected box-constrained optimization algorithm and stopping criteria

3.1 Two-metric projected newton algorithm

The convexity of the dual problem to solve at the response-node ensures that a unique solution exists on the domain of the objective function. The algorithm used to solve the problem should allow sufficient descent to reach an adequate approximation of this unique solution. Since the components of $\boldsymbol{\alpha}$ are bounded by a compact set included in the open set $(0, 1)$ (see Supplementary Methods 1), an algorithm adequate for box-constrained convex optimization problem had to be selected. While many methods exist for box-constrained optimization [6], the chosen method should allow to reach convergence with sufficient precision given the potentially small magnitude of the dual parameter $\boldsymbol{\alpha}$ while still offering efficient computation when the dimension of the dual is high. We used the Two-Metric Projected Newton method suggested by Bertsekas

[7], applicable because Lemma S1 ensures that the dual parameter estimates lie in a compact parameter space $\Theta^{\boldsymbol{X}}_{\boldsymbol{\alpha},\lambda} \subset (0,1)$. We refer to [8] for an extensive description of the method and convergence details. Briefly, all components of the estimate $\widehat{\boldsymbol{\alpha}}^{\lambda}_{(t)}$ at step t at a boundary of the search domain and for which the gradient would pull the search toward the opposite side of the search domain are updated through gradient descent projected in the domain, while all other components are updated using Newton descent projected in the domain. The update is therefore $\widehat{\boldsymbol{\alpha}}^{\lambda}_{(t+1)} = \mathrm{Proj}[\widehat{\boldsymbol{\alpha}}^{\lambda}_{(t)} - \theta \boldsymbol{D}_{(t)} \nabla_{\boldsymbol{\alpha}} J^{\lambda}(\widehat{\boldsymbol{\alpha}}^{\lambda}_{(t)})]$, where $\boldsymbol{D}_{(t)}$ depends of the component as described before and $\mathrm{Proj}[\cdot]$ denotes the projection under the Euclidean norm. The step size θ is selected through backtracking line search (Armijo rule) along projection arc detailed in [6, 8]. An initial admissible estimate has to be provided, which was set at $\widehat{\boldsymbol{\alpha}}^{\lambda}_{(0)} = [0.1, \ldots, 0.1]^{\top}$.

3.2 Stopping criteria

The error entailed by the approximation of $\widehat{\boldsymbol{\alpha}}^{\lambda}$ in the chosen algorithm should ideally be low enough such that it preserve the asymptotic properties derived for the primal estimate. We notice that λ holds a scaling role over the dual parameter $\boldsymbol{\alpha}$ when it comes to retrieving the associated primal parameter $\boldsymbol{\beta}$. A restriction in function of λ consequently needs to be imposed in the estimation of the dual parameter to preserve the asymptotic properties of the primal parameters. The following proposition will allow to derive a stopping criteria for the dual estimation that ensures the asymptotic properties of the primal parameter hold.

Proposition S2. *For any $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \in \{-1, 1\} \times \mathbb{R}^p$ and any $\epsilon > 0$, consider $\tilde{\boldsymbol{\alpha}}^{\lambda} := \tilde{\boldsymbol{\alpha}}^{\lambda}_{\epsilon} \in (0,1)^n$ such that*

$$\|\nabla_{\boldsymbol{\alpha}} J^{\lambda}(\tilde{\boldsymbol{\alpha}}^{\lambda})\|_2 \leq \frac{2\lambda}{\sqrt{p+1}} \Big(\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2 + n \Big)^{-1/2} \epsilon \,.$$

Then, $\max_{j \in \{0,\ldots,p\}} |\tilde{\beta}^{\lambda}_j - \check{\beta}^{\lambda}_j| < \epsilon$, where $\tilde{\beta}^{\lambda}_0 = \sum_{i=1}^{n} y_i \tilde{\alpha}^{\lambda}_i$ and $\tilde{\beta}^{\lambda}_j = \sum_{i=1}^{n} y_i \tilde{\alpha}^{\lambda}_i x_{ij}$ for $j \in \{1, \ldots, p\}$, and with the $\check{\beta}^{\lambda}_j$'s, $j \in \{0, \ldots, p\}$, defined as in Lemma S1.

Proof of Proposition S2. Fix $\epsilon > 0$, and let $\tilde{\boldsymbol{\alpha}}^{\lambda}$ and $\tilde{\beta}^{\lambda}_j$, for $j \in \{0, \ldots, p\}$, be as defined in the statement of the lemma. (Although $\tilde{\boldsymbol{\alpha}}^{\lambda}$ and the $\tilde{\beta}^{\lambda}_j$'s implicitly depend on ϵ, this dependence is not explicitly reflected in the notation, for simplicity in the exposition.) For simplicity in the proof, for $i \in \{1,\ldots,n\}$ let $x_{0i} = 1$, so that $\tilde{\beta}_j^\lambda = \sum_{i=1}^n y_i \tilde{\alpha}_i^\lambda x_{ij}$ and $\widehat{\beta}_j^\lambda = \sum_{i=1}^n y_i \widehat{\alpha}_i^\lambda x_{ij}$ for $j \in \{0,\ldots,p\}$.

Let $\breve{\boldsymbol{\alpha}}^\lambda$ be defined as in Lemma S1, and recall from that lemma that $\breve{\beta}_0^\lambda = \sum_{i=1}^n y_i \breve{\alpha}_i^\lambda$ and $\breve{\beta}_j^\lambda = \sum_{i=1}^n y_i \breve{\alpha}_i^\lambda x_{ij}$ for $j \in \{1,\ldots,p\}$.

Using the fact that $y_i \in \{-1,1\}$ for all $i \in \{1,\ldots,n\}$, we derive

$$\begin{aligned}
\max_{j\in\{0,\ldots,p\}} |\tilde{\beta}_j^\lambda - \breve{\beta}_j^\lambda| &= \max_{j\in\{0,\ldots,p\}} \Big|(\lambda n)^{-1} \sum_{i=1}^n x_{ij} y_i (\tilde{\alpha}_i^\lambda - \breve{\alpha}_i^\lambda)\Big| \\
&\leq (\lambda n)^{-1} \sum_{j=0}^p \sum_{i=1}^n |x_{ij}||\tilde{\alpha}_i^\lambda - \breve{\alpha}_i^\lambda| \\
&\leq (\lambda n)^{-1} \|\tilde{\boldsymbol{\alpha}}^\lambda - \breve{\boldsymbol{\alpha}}^\lambda\|_2 \sum_{j=0}^p \Big(\sum_{i=1}^n x_{ij}^2\Big)^{1/2} \\
&\leq \sqrt{p+1}(\lambda n)^{-1} \|\tilde{\boldsymbol{\alpha}}^\lambda - \breve{\boldsymbol{\alpha}}^\lambda\|_2 \Big(\sum_{j=0}^p \sum_{i=1}^n x_{ij}^2\Big)^{1/2}. \qquad \text{(S15)}
\end{aligned}$$

To obtain the one-to-last line, we used Cauchy-Schwartz inequality, and to obtain the last line, we used the fact that for any positive $a_0,\ldots,a_p$ we have $\sum_{j=0}^p a_j \leq \sqrt{p+1}(\sum_{j=0}^p a_j^2)^{1/2}$

Now observe that, using standard vector calculus manipulations, the Hessian matrix of $J^\lambda(\boldsymbol{\alpha})$ can be expressed as

$$\begin{aligned}
\nabla^2_{\boldsymbol{\alpha}} J^\lambda(\boldsymbol{\alpha}) =& (\lambda n^2)^{-1} \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\mathcal{K}} \operatorname{diag}(\boldsymbol{y}) \\
&+ n^{-1} \operatorname{diag}\left\{ \left[(\alpha_1(1-\alpha_1))^{-1}, \cdots, (\alpha_n(1-\alpha_n))^{-1}\right]^\top \right\},
\end{aligned}$$

where we used the notation $\boldsymbol{\mathcal{K}} = \boldsymbol{X}\boldsymbol{X}^\top + \mathbf{1}_n \mathbf{1}_n^\top$ as defined in Supplementary Tables 1.

In the equation above, the matrix in the first term of the right-hand side of the equality is semi-positive definite, since for any vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\boldsymbol{\alpha}^\top \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\mathcal{K}} \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\alpha} = \|[\boldsymbol{X}\ \ \mathbf{1}_n]^\top \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\alpha}\|_2^2 \geq 0$. As the matrix $n^{-1} \operatorname{diag}\{[(\alpha_1(1-\alpha_1))^{-1}, \cdots, (\alpha_n(1-\alpha_n))^{-1}]^\top\}$ is positive definite for all $\boldsymbol{\alpha} \in (0,1)^n$, with $(\alpha_i(1-\alpha_i))^{-1} \geq 4$ for all $i \in \{1,\ldots,n\}$, it follows that $\nabla^2_{\boldsymbol{\alpha}} J^\lambda(\boldsymbol{\alpha})$ is strongly convex, with strong convexity parameter $m = 4n^{-1}$, since it follows from the last discussion that the matrix

$$\nabla^2_{\boldsymbol{\alpha}} J^\lambda(\boldsymbol{\alpha}) - \frac{m}{2} \boldsymbol{I},$$

is positive definite.

This allows us to conclude as in e.g. [9], Section 9.1.2, p.459, that it holds for all $\boldsymbol{\alpha} \in (0,1)^n$ that

$$||\boldsymbol{\alpha} - \breve{\boldsymbol{\alpha}}^\lambda||_2 \leq \frac{2}{m}\|\nabla_{\boldsymbol{\alpha}} J^\lambda(\boldsymbol{\alpha})\|_2 = \frac{n}{2}\|\nabla_{\boldsymbol{\alpha}} J^\lambda(\boldsymbol{\alpha})\|_2.$$

Combining this result with the inequality derived at (S15), we obtain

$$\max_{j\in\{0,\ldots,p\}} |\tilde{\beta}_j^\lambda - \breve{\beta}_j^\lambda| \leq \frac{\sqrt{p+1}}{2\lambda}\|\nabla_{\boldsymbol{\alpha}} J^\lambda(\tilde{\boldsymbol{\alpha}}^\lambda)\|_2 \Big(\sum_{j=0}^{p}\sum_{i=1}^{n} x_{ij}^2\Big)^{1/2}.$$

The proof of the lemma follows from the assumption over $\|\nabla_{\boldsymbol{\alpha}} J^\lambda(\tilde{\boldsymbol{\alpha}}^\lambda)\|_2$.

□

As shown in the proof of Proposition S1, the maximizer $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ of $l_n^\lambda(\beta_0, \boldsymbol{\beta})$, defined in (2) in the manuscript, satisfies

$$\begin{bmatrix} \widehat{\beta}_0^\lambda \\ \widehat{\boldsymbol{\beta}}^\lambda \end{bmatrix} = \begin{bmatrix} \breve{\beta}_0^\lambda - \widehat{\boldsymbol{\mu}}^\top \widehat{\boldsymbol{\beta}}^\lambda \\ \widehat{\boldsymbol{\Sigma}}^{-1}\breve{\boldsymbol{\beta}}^\lambda \end{bmatrix},$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are defined in (S7), and where $(\breve{\beta}_0^\lambda, \breve{\boldsymbol{\beta}}^\lambda)$, satisfies

$$\begin{bmatrix} \breve{\beta}_0^\lambda \\ \breve{\boldsymbol{\beta}}^\lambda \end{bmatrix} = (n\lambda)^{-1}\sum_{i=1}^{n} y_i \widehat{\alpha}_i^\lambda \begin{bmatrix} 1 \\ \boldsymbol{x}_{i,\text{cs}} \end{bmatrix}.$$

In the last equation, $\widehat{\boldsymbol{\alpha}}^\lambda = (\widehat{\alpha}_1^\lambda, \ldots, \widehat{\alpha}_n^\lambda)^\top \in (0,1)^n$ denotes the unique solution to the minimization problem $\min_{\boldsymbol{\alpha}\in(0,1)^n} J^\lambda(\boldsymbol{\alpha})$, with J^λ as in (3).

From Proposition S2, if $\tilde{\boldsymbol{\alpha}}^\lambda$ is a point such that

$$\|\nabla_{\boldsymbol{\alpha}} J^\lambda(\tilde{\boldsymbol{\alpha}}^\lambda)\|_2 \leq \frac{2\lambda}{\sqrt{p+1}}\Big(\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij,\text{cs}}^2 + n\Big)^{-1/2}\epsilon,$$

it follows that $\max_{j\in\{0,\ldots,p\}} |\breve{\beta}_j^\lambda - \breve{\beta}_j^\lambda| < \epsilon$, where $\breve{\beta}_0^\lambda = \sum_{i=1}^n y_i\tilde{\alpha}_i^\lambda$ and $\breve{\beta}_j^\lambda = \sum_{i=1}^n y_i \tilde{\alpha}_i^\lambda x_{ij,\text{cs}}$ for $j \in \{1,\ldots,p\}$. Therefore, if $(\tilde{\beta}_0^\lambda, \tilde{\boldsymbol{\beta}}^\lambda)$ denotes a version of $(\widehat{\beta}_0^\lambda, \widehat{\boldsymbol{\beta}}^\lambda)$ computed based on $\tilde{\boldsymbol{\alpha}}^\lambda$ instead of $\widehat{\boldsymbol{\alpha}}^\lambda$, i.e, if $(\tilde{\beta}_0^\lambda, \tilde{\boldsymbol{\beta}}^\lambda)$

satisfies

$$\begin{bmatrix}\tilde{\beta}_0^\lambda \\ \tilde{\boldsymbol{\beta}}^\lambda\end{bmatrix} = \begin{bmatrix}\breve{\beta}_0^\lambda - \widehat{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\beta}}^\lambda \\ \widehat{\boldsymbol{\Sigma}}^{-1}\breve{\boldsymbol{\beta}}^\lambda\end{bmatrix},$$

then, for $j \in \{1, \ldots, p\}$, we have

$$|\widehat{\beta}_j^\lambda - \tilde{\beta}_j^\lambda| \leq s_{n,j}^{-1}|\breve{\beta}_j^\lambda - \breve{\beta}_j^\lambda| \leq s_{n,j}^{-1}\epsilon\,.$$

Supplementary Methods 4 Privacy-preserving properties

4.1 Proof of Proposition 1

Before proving Proposition 1, we state and prove the following lemma, which provides the foundation for the proof of Proposition 1. Let $\boldsymbol{e}_{i,p^{(k)}} \in \mathbb{R}^{p^{(k)}}$ denote the standard basis vector with a 1 in the i^{th} position and 0 elsewhere.

Lemma S4. *Let $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$, and consider $\boldsymbol{P} \in \mathcal{M}_{p^{(k)},p^{(k)}}(\mathbb{R})$ such that $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_{p^{(k)}}$. Then, if $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \leq i \leq p^{(k)} - 1$, we have $\boldsymbol{A}\boldsymbol{P} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$.*

Proof of Lemma S4. First note that $\mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ is non-empty since $\boldsymbol{X}_{\text{cs}}^{(k)} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$. From this, to show that $\boldsymbol{A}\boldsymbol{P} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ whenever $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$, $\boldsymbol{P} \in \mathcal{M}_{p^{(k)},p^{(k)}}(\mathbb{R})$ with $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_{p^{(k)}}$ and $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \leq i \leq p^{(k)} - 1$, we need to verify that for such $\boldsymbol{A}$ and $\boldsymbol{P}$ we have (1) $(\boldsymbol{A}\boldsymbol{P})(\boldsymbol{A}\boldsymbol{P})^\top = \boldsymbol{\mathcal{K}}^{(k)}$; (2) $(\boldsymbol{A}\boldsymbol{P})^\top \boldsymbol{1}_n = 0$; and (3) $\text{diag}_{\text{vec}}\{(\boldsymbol{A}\boldsymbol{P})^\top(\boldsymbol{A}\boldsymbol{P})\} = (n-1)\boldsymbol{1}_{p^{(k)}}$.

To verify (1), it suffices to note that, since by definition $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\mathcal{K}}^{(k)}$ and $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_{p^{(k)}}$, we have $(\boldsymbol{A}\boldsymbol{P})(\boldsymbol{A}\boldsymbol{P})^\top = \boldsymbol{A}(\boldsymbol{P}\boldsymbol{P}^\top)\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\mathcal{K}}^{(k)}$.

To verify (2), since $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ implies $\boldsymbol{A}^\top \boldsymbol{1}_n = 0$, one directly computes that $(\boldsymbol{A}\boldsymbol{P})^\top \boldsymbol{1}_n = \boldsymbol{P}^\top(\boldsymbol{A}^\top \boldsymbol{1}_n) = 0$.

To verify (3), note that $\text{diag}_{\text{vec}}\{(\boldsymbol{A}\boldsymbol{P})^\top(\boldsymbol{A}\boldsymbol{P})\} = (n-1)\boldsymbol{1}_{p^{(k)}}$ if and only if $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \leq i \leq p^{(k)}$. Since we have only assumed that $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \leq i \leq p^{(k)} - 1$, we need to prove that, under our conditions, $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{p^{(k)},p^{(k)}}\|_2^2 = (n-1)$. To see why this is the case, note that since $\text{diag}_{\text{vec}}(\boldsymbol{A}^\top \boldsymbol{A}) = (n-1)\boldsymbol{1}_{p^{(k)}}$, we have

$$\sum_{i=1}^{p^{(k)}} \|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = \text{Tr}\{(\boldsymbol{A}\boldsymbol{P})^\top(\boldsymbol{A}\boldsymbol{P})\}$$

$$= \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A}) = p^{(k)}(n-1)\,.$$

This implies that $\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}_{p^{(k)},p^{(k)}}\|_2^2 = (n-1)$, thereby concluding the proof of the proposition.

□

Proof of Proposition 1. First note that since $\boldsymbol{X}_{\mathrm{cs}}^{(k)} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$, Lemma S4 implies that for any orthogonal matrix $\boldsymbol{P} \in \mathcal{M}_{p^{(k)},p^{(k)}}(\mathbb{R})$ such that $\|\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \le i \le p^{(k)} - 1$, we have $\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$. Since $\boldsymbol{P}\boldsymbol{e}_{j,p^{(k)}}$ corresponds to the j^{th} column of $\boldsymbol{P}$, say $\boldsymbol{p}_j$, which has unit norm, and each column of $\boldsymbol{X}_{\mathrm{cs}}^{(k)}$ has squared Euclidian norm equal to $n-1$, it follows that $\|\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P}\boldsymbol{e}_{j,p^{(k)}}\|_2^2 = n-1$ for all $1 \le j \le p^{(k)}$ if and only if

$$\boldsymbol{p}_j^\top \left\{ \frac{(\boldsymbol{X}_{\mathrm{cs}}^{(k)})^\top \boldsymbol{X}_{\mathrm{cs}}^{(k)}}{n-1} - \boldsymbol{I}_{p^{(k)}} \right\} \boldsymbol{p}_j = 0 \qquad \text{for all } j \in \{1, \ldots, p^{(k)}\}\,,$$

or equivalently, if and only if

$$\sum_{\ell=1}^{p^{(k)}-1} \sum_{\ell'=\ell+1}^{p^{(k)}} [\boldsymbol{P}]_{\ell j}[\boldsymbol{P}]_{\ell' j} \Big\{ \sum_{i=1}^{n} x_{i\ell,\mathrm{cs}}^{(k)} x_{i\ell',\mathrm{cs}}^{(k)} \Big\} = 0 \quad \text{for all } j \in \{1, \ldots, p^{(k)}\}.$$

One of these equations is redundant, since $\sum_{j=1}^{p^{(k)}} [\boldsymbol{P}]_{\ell j}[\boldsymbol{P}]_{\ell' j} = 0$ when $\ell \neq \ell'$, due to the orthogonality of the rows of $\boldsymbol{P}$. We conclude by dividing each side of the previous equation by n that if $\boldsymbol{P}$ is orthogonal and satisfies

$$\sum_{\ell=1}^{p^{(k)}-1} \sum_{\ell'=\ell+1}^{p^{(k)}} [\boldsymbol{P}]_{\ell j}[\boldsymbol{P}]_{\ell' j} \tau_{\ell\ell'}^{(k)} = 0 \quad \text{for all } j \in \{1, \ldots, p^{(k)} - 1\},$$

then $\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$. From this, to conclude the proof of the proposition, we need to show that any $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ expresses as $\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P}$, with $\boldsymbol{P}$ an orthogonal matrix such that $\sum_{\ell=1}^{p^{(k)}-1} \sum_{\ell'=\ell+1}^{p^{(k)}} [\boldsymbol{P}]_{\ell j}[\boldsymbol{P}]_{\ell' j} \tau_{\ell\ell'}^{(k)} = 0$ for all $j \in \{1, \ldots, p^{(k)} - 1\}$.

First, from Theorem 7.3.11 in [10] (p.452), if $\boldsymbol{A} \in \mathcal{M}_{n,p^{(k)}}(\mathbb{R})$ is such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{X}_{\mathrm{cs}}^{(k)}(\boldsymbol{X}_{\mathrm{cs}}^{(k)})^\top$, then there exists an orthogonal matrix $\boldsymbol{P} \in \mathcal{M}_{p^{(k)},p^{(k)}}(\mathbb{R})$ such that $\boldsymbol{A} = \boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P}$.

The proposition result then follows from the fact that, since $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ implies $\mathrm{diag}_{\mathrm{vec}}(\boldsymbol{A}^\top \boldsymbol{A}) = (n-1)\mathbf{1}_{p^{(k)}}$, we have

$$\mathrm{diag}_{\mathrm{vec}}\{(\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P})^\top(\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P})\} = (n-1)\mathbf{1}_{p^{(k)}} ,$$

which implies that $\|\boldsymbol{X}_{\mathrm{cs}}^{(k)}\boldsymbol{P}\boldsymbol{e}_{i,p^{(k)}}\|_2^2 = (n-1)$ for all $1 \le i \le p^{(k)} - 1$. □

4.2 Orthogonal matrices that preserve the binary nature of covariates

Proposition S3. *Let $p^{(k)} \ge 2$, and for each $j \in \{1, \dots, p^{(k)}\}$, define $\mathcal{D}_j = \{a_j, b_j\}$ with $a_j \ne b_j$. Assume that $\boldsymbol{A} \in \mathcal{M}_{n,p^{(k)}}(\mathbb{R})$ satisfies $\boldsymbol{A}_{ij} \in \mathcal{D}_j$ for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p^{(k)}\}$. If $n \ge 2^{p^{(k)}}$ and $\boldsymbol{A}$ contains exactly $2^{p^{(k)}}$ distinct rows, then any orthogonal matrix $\boldsymbol{P} \in \mathbb{R}^{p^{(k)} \times p^{(k)}}$ such that $[\boldsymbol{A}\boldsymbol{P}]_{ij} \in \{a'_j, b'_j\}$ for some $a'_j \ne b'_j$ for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p^{(k)}\}$ must be a sign-permutation matrix.*

Proof of Proposition S3. Let $\boldsymbol{A}' \in \mathcal{M}_{2^{p^{(k)}},p^{(k)}}(\mathbb{R})$ denote a submatrix of $\boldsymbol{A}$ consisting of $2^{p^{(k)}}$ distinct rows. Assume without loss of generality that the rows of $\boldsymbol{A}'$ are arranged in a way that

$$[\boldsymbol{A}']_{ij} = \begin{cases} a_j & \text{if } [\mathrm{bin}(i-1)]_j = 0 \\ b_j & \text{if } [\mathrm{bin}(i-1)]_j = 1 \,, \end{cases}$$

where, for any integer i, we use $\mathrm{bin}(i)$ to denote a vector containing its binary representation. In this notation, $\boldsymbol{A}'$ has the form

$$\boldsymbol{A}' = \begin{bmatrix} a_1 & a_2 & \cdots & a_{p^{(k)}-1} & a_{p^{(k)}} \\ a_1 & a_2 & \cdots & a_{p^{(k)}-1} & b_{p^{(k)}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_1 & b_2 & \cdots & a_{p^{(k)}-1} & a_{p^{(k)}} \\ b_1 & b_2 & \cdots & b_{p^{(k)}-1} & b_{p^{(k)}} \end{bmatrix} .$$

To prove the proposition, it suffices to show that, if $\boldsymbol{p} \in \mathbb{R}^{p^{(k)}}$ is a unit vector such that the entries of the vector $\boldsymbol{A}'\boldsymbol{p}$ satisfy $[\boldsymbol{A}'\boldsymbol{p}]_i \in \{r, s\}$ for some $r \ne s$, for all $i \in \{1, \dots, 2^{p^{(k)}}\}$, then $\boldsymbol{p}$ has exactly one non-zero entry. To do this, we proceed by induction on $p^{(k)}$: we first show that the result is true for $p^{(k)} = 2$, then, we prove that if it is true for $p^{(k)} - 1$, it implies that it is also true for $p^{(k)}$.

In the case $p^{(k)} = 2$, $\boldsymbol{A}'$ satisfies

$$\boldsymbol{A}' = \begin{bmatrix} a_1 & a_2 \\ a_1 & b_2 \\ b_1 & a_2 \\ b_1 & b_2 \end{bmatrix} .$$

We then need to show that if $\boldsymbol{p} \in \mathbb{R}^2$ is a unit vector such that the entries of the vector $\boldsymbol{A}'\boldsymbol{p}$ satisfy $[\boldsymbol{A}'\boldsymbol{p}]_i \in \{r, s\}$ for some $r \neq s$, for all $i \in \{1, \ldots, 4\}$, then $\boldsymbol{p}$ has exactly one non-zero entry. To this end, assume that it is not the case and that both entries of $\boldsymbol{p} = [p_1, p_2]^\top$ are non-zero. In this case, since $a_j \neq b_j$ for $j \in \{1, 2\}$, we have $[\boldsymbol{A}'\boldsymbol{p}]_1 - [\boldsymbol{A}'\boldsymbol{p}]_2 \neq 0$, $[\boldsymbol{A}'\boldsymbol{p}]_1 - [\boldsymbol{A}'\boldsymbol{p}]_3 \neq 0$ and $[\boldsymbol{A}'\boldsymbol{p}]_2 - [\boldsymbol{A}'\boldsymbol{p}]_4 \neq 0$. As $[\boldsymbol{A}'\boldsymbol{p}]_i \in \{r, s\}$ for some $r \neq s$, we therefore have $[\boldsymbol{A}'\boldsymbol{p}]_1 = [\boldsymbol{A}'\boldsymbol{p}]_4$ and $[\boldsymbol{A}'\boldsymbol{p}]_2 = [\boldsymbol{A}'\boldsymbol{p}]_3$. These equalities implies that

$$\begin{cases} (a_1 - b_1)p_1 + (a_2 - b_2)p_2 = 0 \\ (a_1 - b_1)p_1 - (a_2 - b_2)p_2 = 0 \end{cases} .$$

This system of equations shows a contradiction, since adding these equations implies $(a_1 - b_1)p_1 = 0$, which is not possible since $a_1 \neq b_1$ and we had assumed that $p_1 \neq 0$. This implies that $\boldsymbol{p} = [p_1, p_2]^\top$ has at least one non-zero entry. Since $\boldsymbol{p}$ has norm equal to 1, and therefore has exactly one non-zero entry (equal to ± 1), which concludes the proof for the case $p^{(k)} = 2$.

We next show that if it is true for $p^{(k)} - 1$, it implies that it is also true for $p^{(k)}$. To this end, first note that for any $i < j \in \{1, \ldots, 2^{p^{(k)}}\}$ such that $\text{bin}(i-1)$ and $\text{bin}(j-1)$ differ by exactly one bit, say the ℓ^{th} one, we have by construction that

$$[\boldsymbol{A}'\boldsymbol{p}]_i - [\boldsymbol{A}'\boldsymbol{p}]_j = (a_\ell - b_\ell)p_\ell \,.$$

Now assume that all components of $\boldsymbol{p}$ are different from 0. By the last equation, this implies that for any $i < j \in \{1, \ldots, 2^{p^{(k)}}\}$ such that $\text{bin}(i-1)$ and $\text{bin}(j-1)$ differ by exactly one bit, we have

$$[\boldsymbol{A}'\boldsymbol{p}]_i \neq [\boldsymbol{A}'\boldsymbol{p}]_j \,, \qquad \text{since } a_\ell \neq b_\ell \text{ for all } \ell \in \{1, \ldots, p^{(k)}\} \,.$$

Within the first four elements of $\boldsymbol{A}'\boldsymbol{P}$, this implies that $[\boldsymbol{A}'\boldsymbol{p}]_1 \neq [\boldsymbol{A}'\boldsymbol{p}]_2$, $[\boldsymbol{A}'\boldsymbol{p}]_1 \neq [\boldsymbol{A}'\boldsymbol{p}]_3$ and $[\boldsymbol{A}'\boldsymbol{p}]_2 \neq [\boldsymbol{A}'\boldsymbol{p}]_4$, or again that $[\boldsymbol{A}'\boldsymbol{p}]_1 = [\boldsymbol{A}'\boldsymbol{p}]_4$ and

$[\boldsymbol{A}'\boldsymbol{p}]_2 = [\boldsymbol{A}'\boldsymbol{p}]_3$ since $[\boldsymbol{A}'\boldsymbol{p}]_i$ can only take one of two possible values $\{r, s\}$. By extending this logic, we deduce that for all i such that $\text{bin}(i-1)$ has an even number of 1's, we have $[\boldsymbol{A}'\boldsymbol{p}]_1 = [\boldsymbol{A}'\boldsymbol{p}]_i$, and $[\boldsymbol{A}'\boldsymbol{p}]_1 \neq [\boldsymbol{A}'\boldsymbol{p}]_i$ otherwise. Specifically, with $\mathcal{A}_1 = \{j \in \{1, \ldots, 2^{p^{(k)}}\} : \sum_{\ell=1}^{p^{(k)}} [\text{bin}(j-1)]_\ell \text{ is even}\}$ and $\mathcal{A}_2 = \{1, \ldots, 2^{p^{(k)}}\} \setminus \mathcal{A}_1$, we have $[\boldsymbol{A}'\boldsymbol{p}]_i = [\boldsymbol{A}'\boldsymbol{p}]_j$ for all $i, j \in \mathcal{A}_1$, and $[\boldsymbol{A}'\boldsymbol{p}]_i = [\boldsymbol{A}'\boldsymbol{p}]_j$ for all $i, j \in \mathcal{A}_2$.

Now note that for any $\ell \neq \ell' \in \{1, \ldots, p^{(k)}\}$, there exists $i, j \in \mathcal{A}_1$ such that $\text{bin}(i-1)$ and $\text{bin}(j-1)$ are equal everywhere except in position ℓ and ℓ', where their bits are flipped. Since $i, j \in \mathcal{A}_1$ implies $[\boldsymbol{A}'\boldsymbol{p}]_i = [\boldsymbol{A}'\boldsymbol{p}]_j$, then

$$
\begin{aligned}
0 &= [\boldsymbol{A}'\boldsymbol{p}]_i - [\boldsymbol{A}'\boldsymbol{p}]_j \\
&= \begin{cases}
(a_\ell - b_\ell)p_\ell + (a_{\ell'} - b_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 0\,, [\text{bin}(i-1)]_{\ell'} = 0 \\
(b_\ell - a_\ell)p_\ell + (a_{\ell'} - b_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 1\,, [\text{bin}(i-1)]_{\ell'} = 0 \\
(a_\ell - b_\ell)p_\ell + (b_{\ell'} - a_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 0\,, [\text{bin}(i-1)]_{\ell'} = 1 \\
(b_\ell - a_\ell)p_\ell + (b_{\ell'} - a_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 1\,, [\text{bin}(i-1)]_{\ell'} = 1\,.
\end{cases}
\end{aligned}
$$

Now let i', j' be such that $\text{bin}(i-1)$ and $\text{bin}(i'-1)$ are identical except at position ℓ', where their bits are flipped, and $\text{bin}(j-1)$ and $\text{bin}(j'-1)$ are identical except at position ℓ', where their bits are flipped. Then, since $i, j \in \mathcal{A}_1$, we have $i', j' \in \mathcal{A}_2$, which therefore implies $[\boldsymbol{A}'\boldsymbol{p}]_{i'} = [\boldsymbol{A}'\boldsymbol{p}]_{j'}$ and that

$$
\begin{aligned}
0 &= [\boldsymbol{A}'\boldsymbol{p}]_{i'} - [\boldsymbol{A}'\boldsymbol{p}]_{j'} \\
&= \begin{cases}
(a_\ell - b_\ell)p_\ell - (a_{\ell'} - b_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 0\,, [\text{bin}(i-1)]_{\ell'} = 0 \\
(b_\ell - a_\ell)p_\ell - (a_{\ell'} - b_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 1\,, [\text{bin}(i-1)]_{\ell'} = 0 \\
(a_\ell - b_\ell)p_\ell - (b_{\ell'} - a_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 0\,, [\text{bin}(i-1)]_{\ell'} = 1 \\
(b_\ell - a_\ell)p_\ell - (b_{\ell'} - a_{\ell'})p_{\ell'} & \text{if } [\text{bin}(i-1)]_\ell = 1\,, [\text{bin}(i-1)]_{\ell'} = 1\,.
\end{cases}
\end{aligned}
$$

Adding the equations $0 = [\boldsymbol{A}'\boldsymbol{p}]_i - [\boldsymbol{A}'\boldsymbol{p}]_j$ and $0 = [\boldsymbol{A}'\boldsymbol{p}]_{i'} - [\boldsymbol{A}'\boldsymbol{p}]_{j'}$ leads to a contradiction, as their validity would require that $(a_\ell - b_\ell)p_\ell = 0$. However, this cannot hold under the assumption that all p_ℓ are nonzero and $a_\ell \neq b_\ell$. Therefore, there must exist at least one component of $\boldsymbol{p}$ that is equal to 0. Without loss of generality, assume that this component is the first one, i.e., p_1.

Consider the matrix $\bar{\boldsymbol{A}}'$, obtained by removing the first column of $\boldsymbol{A}'$ and discarding its last $2^{p^{(k)}-1}$ rows. Let $\bar{\boldsymbol{p}} \in \mathbb{R}^{p^{(k)}-1}$ denote the vector obtained by removing the first component of $\boldsymbol{p}$. Then $\bar{\boldsymbol{p}}$ is a unit vector such that the entries of the vector $\bar{\boldsymbol{A}}'\bar{\boldsymbol{p}}$ satisfy $[\bar{\boldsymbol{A}}'\bar{\boldsymbol{p}}]_i \in r, s$ for some $r \neq s$ and for all $i \in 1, \ldots, 2^{p^{(k)}-1}$. By the induction hypothesis, it follows that $\bar{\boldsymbol{p}}$ has exactly one nonzero component. Since $\boldsymbol{p} = [0, \bar{\boldsymbol{p}}^\top]^\top$, it follows $\boldsymbol{p}$ also has exactly one nonzero component. This concludes the proof that if $\boldsymbol{p} \in \mathbb{R}^{p^{(k)}}$ is a unit vector such that the entries of the vector $\boldsymbol{A}'\boldsymbol{p}$ satisfy $[\boldsymbol{A}'\boldsymbol{p}]_i \in \{r, s\}$ for some $r \neq s$, for all $i \in \{1, \ldots, 2^{p^{(k)}}\}$, then $\boldsymbol{p}$ has exactly one non-zero entry. The result follows. $\square$

4.3 Proof of Proposition 2

Before establishing the proof of Proposition 2, consider the case $p^{(k)} = 3$. In this case, any orthogonal matrix $\boldsymbol{P}$ can be obtained as $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{\theta}}\boldsymbol{P}_\pi^\pm$ where, for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in [0, 2\pi)$,

$$\boldsymbol{P}_{\boldsymbol{\theta}} = \begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) & 0 \\ \sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta_2) & 0 & -\sin(\theta_2) \\ 0 & 1 & 0 \\ \sin(\theta_2) & 0 & \cos(\theta_2) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_3) & -\sin(\theta_3) \\ 0 & \sin(\theta_3) & \cos(\theta_3) \end{bmatrix}$$

(see for example [11]). Matrices of the form $\boldsymbol{P}_{\boldsymbol{\theta}}$ generate all orthogonal matrices $\boldsymbol{P}$ with determinant equal to 1. Since $\boldsymbol{X}_{\text{cs}}^{(k)}\boldsymbol{P}_{\boldsymbol{\theta}}\boldsymbol{P}_\pi^\pm \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ if and only if $\boldsymbol{X}_{\text{cs}}^{(k)}\boldsymbol{P}_{\boldsymbol{\theta}} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$, to characterize the matrices belonging in the set $\mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ it suffices to search for the values of $\boldsymbol{\theta} \in (-\pi, \pi]^3$ such that $\boldsymbol{X}_{\text{cs}}^{(k)}\boldsymbol{P}_{\boldsymbol{\theta}} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$. It follows from Proposition 1 that $\boldsymbol{X}_{\text{cs}}^{(k)}\boldsymbol{P}_{\boldsymbol{\theta}} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ provided $\boldsymbol{\theta} \in (-\pi, \pi]^3$ satisfies $\boldsymbol{g}(\boldsymbol{\theta} \mid \tau_{12}^{(k)}, \tau_{13}^{(k)}, \tau_{23}^{(k)}) = [0, 0]^\top$, where

$$\begin{aligned} g_1(\boldsymbol{\theta} \mid \tau_{12}^{(k)}, \tau_{13}^{(k)}, \tau_{23}^{(k)}) &= \cos(\theta_1)\sin(\theta_1)\cos^2(\theta_2)\tau_{12}^{(k)} + \cos(\theta_1)\cos(\theta_2)\sin(\theta_2)\tau_{13}^{(k)} \\ &\quad + \sin(\theta_1)\cos(\theta_2)\sin(\theta_2)\tau_{23}^{(k)} \end{aligned}$$

and

$$\begin{aligned} &g_2(\boldsymbol{\theta} \mid \tau_{12}^{(k)}, \tau_{13}^{(k)}, \tau_{23}^{(k)}) \\ &= \{-\sin(\theta_1)\cos(\theta_3) - \cos(\theta_1)\sin(\theta_2)\sin(\theta_3)\} \\ &\qquad \times \{\cos(\theta_1)\cos(\theta_3) - \sin(\theta_1)\sin(\theta_2)\sin(\theta_3)\}\tau_{12}^{(k)} \\ &\quad - \{\sin(\theta_1)\cos(\theta_3) + \cos(\theta_1)\sin(\theta_2)\sin(\theta_3)\}\cos(\theta_2)\sin(\theta_3)\tau_{13}^{(k)} \\ &\quad + \{\cos(\theta_1)\cos(\theta_3) - \sin(\theta_1)\sin(\theta_2)\sin(\theta_3)\}\cos(\theta_2)\sin(\theta_3)\tau_{23}^{(k)}. \end{aligned}$$

By the Implicit Function Theorem [12], since the system consists of two continuously differentiable equations and three unknowns, the existence of a solution in the interior of $(-\pi, \pi]^3$ implies that the system admits infinitely many solutions. Such a solution always exists, since $\boldsymbol{\theta}_0 = \mathbf{0}$ corresponds to $\boldsymbol{P}_{\boldsymbol{\theta}_0} = \boldsymbol{I}_{p^{(k)}}$ always satisfies $\boldsymbol{X}_{\text{cs}}^{(k)} \boldsymbol{P}_{\boldsymbol{\theta}_0} = \boldsymbol{X}_{\text{cs}}^{(k)} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$. The rank r of the Jacobian matrix of $\boldsymbol{g}(\cdot \mid \tau_{12}^{(k)}, \tau_{13}^{(k)}, \tau_{23}^{(k)})$ at the solution determines the dimension of the solution set, but regardless of the rank, the solution set is infinite: it is locally a manifold of dimension $3 - r$.

The following proof formalizes this argument.

Proof of Proposition 2. By Proposition 1, $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ if and only if $\boldsymbol{A} = \boldsymbol{X}_{\text{cs}}^{(k)} \boldsymbol{P}$, with $\boldsymbol{P} \in \mathcal{M}_{p^{(k)}, p^{(k)}}(\mathbb{R})$ an orthogonal matrix satisfying (14). Moreover, any orthogonal $\boldsymbol{P} \in \mathcal{M}_{p^{(k)}, p^{(k)}}(\mathbb{R})$ can be expressed as $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{\theta}} \boldsymbol{P}_{\pi}^{\pm}$, where $\boldsymbol{P}_{\boldsymbol{\theta}}$ is a rotation matrix parametrized using the Givens rotation basis described in the statement of the proposition. Recalling the definition of $\boldsymbol{g}(\boldsymbol{\theta})$ in the statement of the proposition, and noting that $\boldsymbol{A} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$ if and only if $\boldsymbol{A}\boldsymbol{P}_{\pi}^{\pm} \in \mathbb{S}(\boldsymbol{\mathcal{K}}^{(k)})$, it suffices, to prove the proposition, to show that the set $\{\boldsymbol{\theta} : \boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}\}$ has infinite cardinality. This follows directly from the smoothness of $\boldsymbol{g}(\boldsymbol{\theta})$, the fact that when $p^{(k)} \geq 3$ the dimension of $\boldsymbol{\theta}$ exceeds the number of equations defined by $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$, the existence of a solution $\boldsymbol{\theta}_0 \in (-\pi, \pi)^{p^{(k)}(p^{(k)}-1)/2}$ satisfying $\boldsymbol{g}(\boldsymbol{\theta}_0) = \mathbf{0}$ (i.e., the solution $\boldsymbol{\theta} = \mathbf{0}$, which corresponds to $\boldsymbol{P}_{\boldsymbol{\theta}} = \boldsymbol{I}_{p^{(k)}}$), and an application of the Implicit Function Theorem. □

4.4 Privacy assessment - Empirical criterion

An empirical criterion was derived to verify if, using the quantities available at the covariate-nodes, every entry of the response-node's data can be flipped (recall $\boldsymbol{y} \in \{-1, 1\}^n$) while still leading to an admissible candidates for the response vector. This criterion, based on theoretical details from in Methods 4.3.2, is described in Algorithm 1 to support numerical implementation. We recall that the solution space that defines the solutions derived from the shared quantity $\widehat{\boldsymbol{c}}^{\lambda(k)}$ is given by

$$\mathbb{S}(\widehat{\boldsymbol{c}}^{\lambda(k)}) = \Big\{\boldsymbol{y}^{\dagger} \in \{-1,1\}^n : \boldsymbol{y}^{\dagger} = \text{sign}\{\text{diag}(\widehat{\boldsymbol{\alpha}}^{\lambda})\boldsymbol{y} + \boldsymbol{W}\boldsymbol{b}\}, \\ \text{with } \text{diag}(\widehat{\boldsymbol{\alpha}}^{\lambda})\boldsymbol{y} + \boldsymbol{W}\boldsymbol{b} \in (-1,1)^n\Big\},$$

This criterion can be verified at the response-node for any covariate-node k not co-located at the response-node.

Algorithm 1 Empirical criterion for the privacy assessment of the response vector $\boldsymbol{y}$ at covariate-node k

Input: Gram matrix $\boldsymbol{\mathcal{K}}^{(k)}$ from covariate-node k, response vector $\boldsymbol{y}$ and dual numerical estimate $\tilde{\boldsymbol{\alpha}}^{\lambda}$.

Output: Number of entries of the vector $\boldsymbol{y}$ that could be flipped.

Procedure:

1. Generate $\boldsymbol{W}$ in the null-space of $\boldsymbol{\mathcal{K}}^{(k)}$.
2. For every $i \in \{1, \dots, n\}$, verify if $\exists\ \boldsymbol{x}_0$ such that $\text{sign}(x_{0i}) \neq \text{sign}(y_i)$, where $\boldsymbol{x}_0 = \text{diag}(\tilde{\boldsymbol{\alpha}}^{\lambda})\boldsymbol{y} + \boldsymbol{W}\boldsymbol{b} \in (-1, 1)^n$.
3. Count the number of entries y_i that satisfied the condition.

References

[1] Sion, M.: On general minimax theorems. Pacific Journal of Mathematics **8**(1), 171–176 (1958)

[2] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ, USA (1970). https://doi.org/10.1515/9781400873173

[3] Wedderburn, R.W.M.: On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. Biometrika **63**(1), 27–32 (1976) https://doi.org/10.2307/2335080

[4] Albert, A., Anderson, J.A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrika **71**(1), 1–10 (1984)

[5] Vaart, A.W.v.d.: Asymptotic Statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK (2000)

[6] Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont, MA, USA (1999)

[7] Bertsekas, D.P.: Projected Newton Methods for Optimization Problems with Simple Constraints. SIAM Journal on Control and Optimization **20**(2), 221–246 (1982) https://doi.org/10.1137/0320018

[8] Schmidt, M., Kim, D., Sra, S.: Projected Newton-type Methods in Machine Learning. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning, pp. 305–330. The MIT Press, Cambridge, MA, USA (2011). https://doi.org/10.7551/mitpress/8996.003.0013

[9] Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge university press, Cambridge, United Kingdom (2004)

[10] Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge university press, New York, NY, USA (2012)

[11] Cornwell, J.F.: Group Theory in Physics: An Introduction vol. 1. Academic press, San Diego, CA (1997)

[12] Rudin, W.: Principles of Mathematical Analysis, Third edition edn. International series in pure and applied mathematics. McGraw-Hill, New York, NY, USA (1976)