Predicting Rare Events by Shrinking Towards Proportional Odds

Gregory Faletto[*] and Jacob Bien

Department of Data Sciences and Operations
University of Southern California Marshall School of Business

April 2, 2023

Abstract

Training classifiers is difficult with severe class imbalance, but many rare events are the culmination of a sequence with much more common intermediate outcomes. For example, in online marketing a user first sees an ad, then may click on it, and finally may make a purchase; estimating the probability of purchases is difficult because of their rarity. We show both theoretically and through data experiments that the more abundant data in earlier steps may be leveraged to improve estimation of probabilities of rare events. We present PRESTO, a relaxation of the proportional odds model for ordinal regression. Instead of estimating weights for one separating hyperplane that is shifted by separate intercepts for each of the estimated Bayes decision boundaries between adjacent pairs of categorical responses, we estimate separate weights for each of these transitions. We impose an L1 penalty on the differences between weights for the same feature in adjacent weight vectors in order to shrink towards the proportional odds model. We prove that PRESTO consistently estimates the decision boundary weights under a sparsity assumption. Synthetic and real data experiments show that our method can estimate rare probabilities in this setting better than both logistic regression on the rare category, which fails to borrow strength from more abundant categories, and the proportional odds model, which is too inflexible.

1 Introduction

Estimating probabilities of rare events is known to be difficult due to class imbalance. However, sometimes these events are the culmination of a sequential process with intermediate outcomes. For example, (1) in online marketing, a customer is first served an ad, then may

[*]Corresponding author: gregory.faletto@marshall.usc.edu

click on it, then may indicate interest in making a purchase (by "liking" the product, for example), and finally may make a purchase. (2) In health and medicine, many outcomes can be encoded as ordered categorical variables, like reported quality of life and disease progression [Norris et al., 2006]. (3) Sales of high-price durable goods typically follow a *sales funnel* [Duncan and Elkan, 2015]. For example, when buying a car often a potential buyer first comes in to see a car, may take a test drive, and finally may buy the car.

In many of these cases, the intermediate events are much more common than the rare events. Though these intermediate events may not be of direct interest, if the features that contribute to the probability of advancing through earlier classes also contribute to the probability of advancing through later classes, then the more abundant intermediate events can be leveraged to improve estimation of the rare event probabilities.

The *proportional odds model* [McCullagh, 1980], also called the *ordered logit model* [Cameron and Trivedi, 2005, Section 15.9.1], satisfies, for ordinal outcomes $k \in \{1, \ldots, K-1\}$

$$\log\left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)}\right) = \alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of weights and $\boldsymbol{x} \in \mathbb{R}^p$ is a vector of features. This implies that for all $k \in \{1, \ldots, K-1\}$

$$p_k(\boldsymbol{x}) := \mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right) = F\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}\right), \tag{2}$$

where $F(\cdot)$ is the logistic cumulative distribution function, $F(t) = \exp\{t\}/[1 + \exp\{t\}]$. Notice that $\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}$ is the Bayes decision boundary for the binary random variable $\mathbb{1}\left\{y \leq k\right\} \mid \boldsymbol{x}$. This problem could instead be cast as $K-1$ binary classification problems of the form (2) for adjacent classes:

$$\log\left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)}\right) = \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}, \qquad k \in \{1, \ldots, K-1\}. \tag{3}$$

The condition that the weight vectors $\boldsymbol{\beta}_k$ of the separating hyperplanes in (3) are all equal, as in (1), has been called the *proportional odds assumption* [McCullagh, 1980] or the *parallel regression assumption* [Greene, 2012, Section 18.3.2]. One way to motivate this model is by supposing that the response is driven by a latent (unobserved) variable U,

$$U = \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon, \tag{4}$$

where ϵ has a standard logistic distribution and is independent of $\boldsymbol{x}$. Response k is observed if and only if $-\alpha_k \leq U_i < -\alpha_{k-1}$ (where we define $\alpha_0 := -\infty$ and $\alpha_K := \infty$). This model leads to (2). (See Section 3.3.2 of Agresti 2010 for a more detailed explanation.)

Because the proportional odds model assumes that the decision boundaries between adjacent classes are all governed by the same hyperplane defined by $\boldsymbol{\beta}$ (separated only by different intercepts α_k), it assumes that the decision boundary between any two classes perfectly explains the decision boundary between any two other classes, other than an intercept term. If a rare event has much more common intermediate events before it, this model can therefore be very useful for better estimating the parameters of the model, and therefore better estimating the rare event probabilities. However, it could be that the proportional odds assumption is too rigid to be realistic, because observed features may have varying influence at different decision boundaries. For example: (1) in online marketing, users may click on an ad only to realize that the product is not what they were expecting, resulting in a particularly low probability of purchase. (2) For expensive goods like a home or car, potential buyers may express interest by going on a tour or taking a test drive purely out of curiosity; this may be distinct from their level of interest in actually making a purchase. (3) Students may place weights on different factors when deciding whether to apply to graduate school than they did when deciding whether to apply to an undergraduate program—they may have more appealing alternatives to additional schooling, they may face new financial or personal constraints because they are older, etc.

In each of these settings, if specific features vary in relevance for different decision boundaries while other features have about the same influence at every boundary, the proportional odds assumption may be too strong. Violations or relaxations of the proportional odds assumption along the lines of (3) have previously been considered by, for example, Brant [1990]. Peterson and Harrell Jr [1990] developed *partial proportional odds models*, which allow the proportional odds assumption to hold for some features but not others, an idea previously mentioned by Armstrong and Sloan [1989]. (See Section 3.6.1 of Agresti 2010 for a textbook-level discussion). These relaxations have not been widely adopted because fitting separate weights for each outcome is too flexible unless $p(K-1) \ll n$ and all classes are reasonably common (and we discuss additional difficulties of this kind of model in Sections 3.1 and 4.1).

1.1 Our Contributions

In this paper we propose relaxing the proportional odds assumption as in (3), but controlling the amount of relaxation by placing ℓ_1 penalties on the differences in weights corresponding to the same features in adjacent $\boldsymbol{\beta}_k$ vectors, in a way that is reminiscent of the fused lasso [Tibshirani et al., 2005]. This model allows us to borrow strength from outcomes where data is much more abundant to improve rare probability estimates when outcomes are much more rare without making the strong assumption that the weights in these models are exactly equal. In particular, it allows for the proportional odds model to hold for some specific features in some adjacent pairs of decision boundaries, but not others.

We formalize the intuitive argument we outline above—that the proportional odds model allows for precise estimation of the $\boldsymbol{\beta}$ vector as long as at least two adjacent classes are fairly common, and this allows for improved estimation of rare probabilities at the end of the sequence—through theoretical results in Section 2. Motivated by this argument but skeptical of the proportional odds assumption holding exactly, we propose PRESTO in

Section 3 and prove that it consistently estimates $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ under a sparsity assumption in Section 3.1. In Section 4 we demonstrate through synthetic and real data experiments that PRESTO can outperform both logistic regression and the proportional odds model, both in settings where the differences in adjacent $\boldsymbol{\beta}_k$ vectors are sparse, as PRESTO assumes, and in settings where these differences are not sparse. Before we move on from the introduction, we review related literature.

1.2 Related Work

The difficulty of classification with class imbalance has been well-known for decades. Johnson and Khoshgoftaar [2019] provide a recent review focusing on deep learning methods for handling class imbalance, and they also provide references for many other ways of dealing with class imbalance. One particularly closely related work is Owen [2007], which explores how logistic regression handles a vanishingly rare class. A particularly popular approach, SMOTE [Chawla et al., 2002], has its own recent review paper [Fernández et al., 2018].

Tutz and Gertheiss [2016] discuss the possibility of penalizing differences in weights between adjacent models, including briefly proposing an ℓ_1 penalty between weights in corresponding categories for proportional hazard models, though this is not the focus of their article and they only mention the idea very briefly without investigating it.

Wurm et al. [2021] propose a generalization of a proportional odds model (and implement it in the R package `ordinalNet`) that allows for the possibility that adjacent categories have equal (or very close) weights, but their method differs from ours. The most closely related model Wurm et al. propose is an over-parameterized *semi-parallel* model with both a matrix of separate parameters for each level, an approach reminiscent of Peterson and Harrell Jr [1990]. This results in more flexible, less structured models than our approach, which assumes similarity between adjacent $\boldsymbol{\beta}_k$ vectors. Further, Wurm et al. [2021] do not investigate the theoretical properties of their model, or the use of their model for improving estimates of rare event probabilities.

Ugba et al. [2021] and Pößnecker and Tutz [2016] implement an ℓ_2 rather than ℓ_1 penalty between weights in models for adjacent decision boundaries. However, these works also focus on ordinal regression more generally, while we focus both theoretically and in simulations on leveraging common classes to improve estimated probabilities of rare events. Further, the ℓ_1 penalty, which imposes sparse differences, allows the proportional odds assumption to hold for some features and decision boundaries and not others, while an ℓ_2 penalty relaxes the proportional odds assumption for all features but regularizes the relaxation.

2 Motivating Theory

We present the following theoretical results to motivate PRESTO. The thrust of our motivation is as follows: (1) Logistic regression does arbitrarily badly as class imbalance worsens (Theorem 4 in the supplement). (2) However, as one would expect, a logistic regression model's ability to estimate probabilities improves when the parameters $\boldsymbol{\beta}$ are known (Theorem 1). (3) The proportional odds model allows for precise estimation of $\boldsymbol{\beta}$ as long as two adjacent classes are reasonably common, even if the remaining classes are arbitrarily rare (Theorem 2). (4) Taking 2 and 3 together, our conclusion is that we can better estimate probabilities of rare events by using a method that leverages data from decision boundaries between abundant classes to better estimate decision boundaries near rare classes. (Both the proportional odds model and PRESTO leverage the data in this way.)

Before we present our results, we discuss the metrics we will use in our results and some of the assumptions we will make.

2.1 Preliminaries

Our goal is to characterize and compare the prediction error of estimated conditional probabilities from both logistic regression and the proportional odds model when one class is rare. There are many settings where estimating rare probabilities accurately (as opposed to, for example, predicting class labels accurately) is important. For example, in online advertising, advertisers bid on the price to display an ad to a given user. Advertisers could bid optimally if they knew the true probability each user would click a given ad, so they'd like to estimate these probabilities as precisely as possible [He et al., 2014, Zhang et al., 2014]. Another example is public policy, where scarce resources may be allocated based on estimated probability of bad outcomes [Von Wachter et al., 2019]. To prioritize optimally, precisely estimated probabilities are needed, not just accurate labels.

A natural metric in an estimation setting is mean squared error, $\mathbb{E}\left[(\pi(\boldsymbol{x}) - \hat{\pi}(\boldsymbol{x}))^2\right]$, where $\pi(\boldsymbol{x})$ is the actual probability of a rare event conditional on $\boldsymbol{x}$ and $\hat{\pi}(\boldsymbol{x})$ is an estimate. Further, we leverage asymptotic statistics and present results for *large-sample* estimators. We define the notions of asymptotic mean squared error we will use below:

Definition 1. *Let $\hat{\theta}_n$ be a maximum likelihood estimator for a parameter $\theta \in \mathbb{R}$ from a sample size of n. Under regularity conditions, the sequence of random variables $\{\sqrt{n} \cdot (\hat{\theta}_n - \theta)\}$ converges in distribution to a Gaussian random variable. Then we define the* ***asymptotic mean squared error*** *of $\hat{\theta}_n$ to be (suppressing n from the notation)*

$$\text{Asym.MSE}(\hat{\theta}) := \mathbb{E}\left[\left(\lim_{n\to\infty} \sqrt{n}\left[\hat{\theta}_n - \theta\right]\right)^2\right].$$

Asymptotic metrics are commonly used to compare the performance of estimators. The *asymptotic relative efficiency* of two estimators is the ratio of their asymptotic variances,

$$\text{Asym.Var}(\hat{\theta}) := \text{Var}\left(\lim_{n\to\infty} \sqrt{n}\left[\hat{\theta}_n - \theta\right]\right),$$

which is equal to Asym.MSE$(\hat{\theta}_n)$ for the (asymptotically unbiased) maximum likelihood estimators we consider. See Section 10.1.3 of Casella and Berger [2021], Section 8.2 of van der Vaart [2000], or Section 4.4.5 of Greene [2012] for textbook-level discussions. The asymptotic MSE could also be used as an estimator of the MSE for large (but finite) n, under the heuristic reasoning that for large n,

$$\text{MSE}(\hat{\theta}) = \frac{1}{n}\mathbb{E}\left[\left(\sqrt{n}\cdot\left[\hat{\theta}-\theta\right]\right)^2\right] \approx \frac{1}{n}\mathbb{E}\left[\left(\lim_{n\to\infty}\sqrt{n}\cdot\left[\hat{\theta}-\theta\right]\right)^2\right] = \frac{1}{n}\text{Asym.MSE}(\hat{\theta}).$$

See Section 4.4 of Greene [2012], Section 7.3 of Hansen [2022], or Section 3.5 of Wooldridge [2002] for more discussion of this kind of finite-sample estimation using asymptotic quantities.

We briefly present and discuss some of our assumptions.

- **Assumption** $X(\mathcal{A})$**:** The random vectors $\boldsymbol{x}_i \in \mathbb{R}^p$ are independent and identically distributed (iid) for $i \in \{1, \ldots, n\}$, each with probability measure $dF(\boldsymbol{x})$ with measurable, bounded support $\mathcal{S} \subset \mathcal{A} \subseteq \mathbb{R}^p$, with $\text{Cov}(\boldsymbol{X})$ positive definite.

- **Assumption** $Y(K)$**:** The response $y_i \in \{1, \ldots, K\}$ is distributed conditionally on $\boldsymbol{x}_i$ as in the proportional odds model (1). (Note that if $K = 2$, this is equivalent to the logistic regression model.) All classes have positive probability for all $\boldsymbol{x}$ on the support of $\boldsymbol{x}_i$ (equivalently, the intercepts strictly differ: $\alpha_1 < \ldots < \alpha_{K-1}$.)

Assumption $X(\mathcal{A})$ allows a very broad class of distributions, including both discrete and continuous random variables. Notice that the boundedness assumption within $X(\mathcal{A})$ implies that the matrix $\tilde{\boldsymbol{X}} := (\mathbf{1}, \boldsymbol{X})$ (where $\mathbf{1}$ is an n-vector of ones) has a finite maximum eigenvalue. When we will refer to it, we call it $\lambda_{\max}$ and write **Assumption** $X(\mathcal{A}, \boldsymbol{\lambda_{\max}})$.

From (2) we see that in the proportional odds model if the intercepts strictly differ ($\alpha_1 < \ldots < \alpha_{K-1}$) then for any $\boldsymbol{x}$ all of the classes have conditional probability strictly between 0 and 1. That said, if the support of $\boldsymbol{X}$ is unbounded then all of the probabilities for individual classes can become arbitrarily close to 0 or 1. Under Assumption $X(\mathcal{A})$, however, we can strictly bound quantities like $\sup_{\boldsymbol{x}\in\mathcal{S}}\{\pi_k(\boldsymbol{x})\}$ (where $\pi_k(\boldsymbol{x}) := p_k(\boldsymbol{x}) - p_{k-1}(\boldsymbol{x}) = \mathbb{P}(y = k \mid \boldsymbol{x})$) away from 1 or 0.

Theorem 1 holds under Assumption $X([0,\infty)^p)$, though for any bounded $\mathcal{S} \subseteq \mathbb{R}^p$, there is some finite a one could add to each coordinate to shift $\mathcal{S}$ to a subset of $[0,\infty)^p$; Theorem 1 would then apply to these translated features.

2.2 Theorem 1

Theorem 1 suggests a possible way to circumvent the problem of class imbalance. We compare the typical logistic regression intercept estimate $\hat{\alpha}$ to the *quasi-estimated* estimator $\hat{\alpha}_q$ obtained when one estimates only the intercept of the logistic regression model with a known $\boldsymbol{\beta}$. We also compare the resulting estimators of conditional probabilities for any $\boldsymbol{z} \in \mathbb{R}^p$: the usual logistic regression estimator $\hat{\pi}(\boldsymbol{z})$ and $\hat{\pi}_q(\boldsymbol{z})$, the estimator when $\boldsymbol{\beta}$ is known. Theorem 1 proves the reasonable intuition that $\hat{\alpha}_q$ must be a better estimator than $\hat{\alpha}$, and likewise for $\hat{\pi}_q(\boldsymbol{z})$ and $\hat{\pi}(\boldsymbol{z})$.

Theorem 1. *Assume $X([0,\infty)^p, \lambda_{max})$ and $Y(2)$ hold. Let $\pi(\boldsymbol{x}) := \mathbb{P}(y = 2 \mid \boldsymbol{x})$, and let $\pi_{min} := \inf_{\boldsymbol{x} \in \mathcal{S}} \{\pi(\boldsymbol{x}) \wedge 1 - \pi(\boldsymbol{x})\}$. Then*

1.
$$\frac{\text{Asym.MSE}(\hat{\alpha}) - \text{Asym.MSE}(\hat{\alpha}_q)}{[\text{Asym.MSE}(\hat{\alpha}_q)]^2} \geq \Delta$$

   where
$$\Delta := \frac{4\pi_{min}^2(1-\pi_{min})^2 \left\| \mathbb{E}\left[\boldsymbol{X}\right] \right\|_2^2}{\lambda_{max}},$$

   and

2. *For any $\boldsymbol{z} \in \mathbb{R}^p \setminus \{\boldsymbol{z}^*\}$, where*
$$\boldsymbol{z}^* := \frac{\mathbb{E}\left[\boldsymbol{X}\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]}{\mathbb{E}\left[\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]},$$

   it holds that
$$\text{Asym.MSE}(\hat{\pi}_q(\boldsymbol{z})) < \text{Asym.MSE}(\hat{\pi}(\boldsymbol{z})).$$

   (For $\boldsymbol{z}^*$, the above holds with $\leq$ rather than $<$.)

Examining the first result, it is sensible that the gap between the asymptotic variances of the two estimators vanishes as $\pi_{\min}$ vanishes because if $\min\{\pi_1(\boldsymbol{x}), 1-\pi_1(\boldsymbol{x})\}$ becomes very small on the bounded support, then the imbalance between the two classes potentially becomes very large, and estimating the intercept becomes difficult regardless of whether or not $\boldsymbol{\beta}$ is known. As the class balance improves ($\pi_{\min}$ becomes closer to its upper bound 1/2), the guaranteed gap between Asym.MSE($\hat{\alpha}$) and Asym.MSE($\hat{\alpha}_q$) becomes larger.

2.3 Theorem 2

Theorem 1 suggests that if only we could estimate $\boldsymbol{\beta}$ very well, we could improve our estimated probabilities even in the face of class imbalance. Theorem 2 suggests a way to leverage abundant data among other classes to do this.

In the proportional odds model (1), $\mathbb{R}^p$ is partitioned into regions with separating hyperplanes defined by $\boldsymbol{\beta}$, which we note are Bayes decision boundaries: for $\boldsymbol{x} \in \mathbb{R}^p$ such that $\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x} = 0$, we have $p_k(\boldsymbol{x}) = 1/2$.

Consider the setting of ordered categorical data generated by the proportional odds model with categories 1 and 2 similarly common over the support of a bounded distribution of $\boldsymbol{x}_i$ and categories $3, \ldots, K$ all rare. In this setting, for many of the observed values of $\boldsymbol{x}_i$, the probabilities of being in class 1 or 2 will both be close to 1/2. Intuitively it should be relatively easy to estimate $\boldsymbol{\beta}$ and α_1, the parameters that define the Bayes decision boundary between classes 1 and 2, and therefore $p_1(\boldsymbol{x}_i)$. Theorem 1 suggests this should help us in estimating the rare class probabilities. In Theorem 2, we prove that even if class K becomes arbitrarily rare, as long as the first two classes are reasonably well balanced, the proportional odds model still learns $\boldsymbol{\beta}$ quite well.

Theorem 2. *Assume $X(\mathbb{R}^p)$ and $Y(3)$ hold. Assume for all $\boldsymbol{x} \in \mathcal{S}$ it holds that $|\pi_k(\boldsymbol{x}) - 1/2| \leq \Delta$ for $k \in \{1, 2\}$ for some $\Delta \in (0, 1/2)$ and let $M := \sup_{\boldsymbol{x} \in \mathcal{S}} \|\boldsymbol{x}\|_2$ (notice that $X(\mathbb{R}^p)$ ensures that $M < \infty$). Suppose $\sup_{\boldsymbol{x} \in \mathcal{S}} \{\pi_3(\boldsymbol{x})\} = \pi_{rare}$, where π_{rare} is no greater than*

$$\min \left\{ \frac{1}{2} \left(\frac{1}{2} - \Delta \right) \left(\frac{1}{2} + \Delta \right), \frac{\lambda_{min} \left(I_{\beta\beta} - 2 \frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}} \right)}{3M^2 (M+2)} \right\}, \tag{5}$$

where $\lambda_{min}(\cdot)$ denotes the minimum eigenvalue of $\cdot$ and $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ is a symmetric matrix composed of terms from the Fisher information matrix for the proportional odds model (see the definitions of these terms in Equations 8, 9, and 10 in the appendix). Then there exists $C < \infty$ not depending on π_{rare} such that for any fixed $\boldsymbol{v} \in \mathbb{R}^p$,

$$\frac{1}{\boldsymbol{v}^\top \boldsymbol{v}} \text{Asym.MSE} \left(\boldsymbol{v}^\top \hat{\boldsymbol{\beta}}^{prop.\ odds} \right) \leq C.$$

Theorem 2 shows that in contrast to logistic regression, the proportional odds model still learns $\boldsymbol{\beta}$ within a fixed precision even as π_{rare} vanishes.

Remark 1. We briefly discuss the upper bound (5). For this bound to make sense, it must hold that the symmetric matrix $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ is positive definite so that its minimum eigenvalue is positive. The matrix $\boldsymbol{S} := I_{\beta\beta} - \frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ is the Schur complement of $I_{\alpha_1\alpha_1} = M_1$ in the submatrix

$$\begin{pmatrix} I_{\alpha_1\alpha_1} & I_{\beta\alpha_1}^\top \\ I_{\beta\alpha_1} & I_{\beta\beta} \end{pmatrix} \tag{6}$$

of the Fisher information matrix $I^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the proportional odds model (see Lemma 5 in the appendix). Note (6) is s a principal submatrix of the positive definite $I^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, so is positive definite by Observation 7.1.2 in Horn and Johnson [2012]. From (8) we also know that $I_{\alpha_1\alpha_1} > 0$, so $\boldsymbol{S}$ is positive definite by Theorem 1.12 in Zhang [2005]. It seems plausible that

$$I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}} = \boldsymbol{S} - \frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$$

is also positive definite because $I_{\beta\beta}$ is the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}^{\text{ideal}}$, the maximum likelihood estimator of $\boldsymbol{\beta}$ when α_1 and α_2 are known. We expect that $\text{Cov}(\hat{\boldsymbol{\beta}}^{\text{ideal}})$ would be small (and the eigenvalues of $I_{\beta\beta}$ would be large) in this setting because we can estimate $\boldsymbol{\beta}$ well due to the abundant observations in classes 1 and 2 (ensured if Δ is not too large), so we should be able to learn the decision boundary between these classes well. If the eigenvalues of $I_{\beta\beta}$ are indeed large, it might be reasonable to expect $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ to be positive definite. In Remark 3 in the appendix, we present more detailed analysis as well as the results of synthetic experiments that indicate that it is plausible both that $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ is positive definite and that the upper bound (5) is reasonable.

3 Predicting Rare Events by Shrinking Towards proportional Odds (PRESTO)

Theorems 1 and 2 suggest a path to improve estimated probabilities for a rare event that is at the end of an ordered sequence: use the more common events that come before it to improve the estimation of the decision boundary affecting the rare class. In practice, however, the proportional odds model assumption is strong and unlikely to hold in many settings. PRESTO allows for this assumption to be relaxed; instead of assuming the $\boldsymbol{\beta}$ vectors governing the decision boundaries are identical, we assume they are in general different, but with differences that are (approximately) sparse.

One concrete model to motivate this is a relaxation of (4) along the lines of (3). Suppose that $U_1 := U$ as defined in (4) (with $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$), and it still holds that an observation is in class 1 if $U_1 \geq -\alpha_1$. However, for $k \in \{2, \ldots, K-1\}$, outcome k is observed if and only if $-\alpha_k \leq U_k < -\alpha_{k-1} + \boldsymbol{\psi}_k^\top \boldsymbol{x}$ for sparse vectors $\boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{K-1} \in \mathbb{R}^p$ satisfying $\boldsymbol{\psi}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}$, so $U_k = U_{k-1} + \boldsymbol{\psi}_k^\top \boldsymbol{x}$ for $k \in \{2, \ldots, K-1\}$. Note that this is within the scope of (3), but we assume a structure on the differing $\boldsymbol{\beta}_k$ vectors rather than allowing for arbitrary differences.

Assuming sparse differences in adjacent $\boldsymbol{\beta}_k$ vectors in this way suggests the following optimization problem for data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$\begin{aligned} \underset{\boldsymbol{\beta}, \boldsymbol{\alpha}}{\arg\min} \Bigg\{ & -\frac{1}{n} \sum_{i=1}^n \log \left[F\left(\alpha_{y_i} + \boldsymbol{\beta}_{y_i}^\top \boldsymbol{x}_i\right) - F\left(\alpha_{y_i - 1} + \boldsymbol{\beta}_{y_i-1}^\top \boldsymbol{x}_i\right) \right] \\ & + \lambda_n \left(\sum_{j=1}^p |\beta_{j1}| + \sum_{j=1}^p \sum_{k=2}^{K-1} |\beta_{jk} - \beta_{j,k-1}| \right) \Bigg\} \end{aligned} \tag{7}$$

where we define $\alpha_K := \infty, \alpha_0 := -\infty$ and $\boldsymbol{\beta}_0 := \mathbf{0}$. The penalties on the $|\beta_{j1}|$ terms are sufficient to regularize all of the weights given the penalties on the difference terms starting from the $\boldsymbol{\beta}_1$ vector, improving parameter estimation. Like the proportional odds model and the generalized lasso [Tibshirani and Taylor, 2011] optimization problem, (7) is strictly convex if and only if $\alpha_{y_i} + \boldsymbol{\beta}_{y_i}^\top \boldsymbol{x}_i > \alpha_{y_i-1} + \boldsymbol{\beta}_{y_i-1}^\top \boldsymbol{x}_i$ for all i [Pratt, 1981]. This can be violated if the decision boundaries, which are not parallel, cross in the support of $\boldsymbol{X}$. In Section 4.1, we discuss the practical issues this presents when implementing relaxed proportional odds models like PRESTO, and in the next session, we prove PRESTO is consistent relying in part on an asumption that these decision boundaries do not cross in the support of $\boldsymbol{X}$. See Appendix G for details on how we estimate PRESTO in practice.

3.1 Theoretical Analysis

In this section, we present Theorem 3, which shows that PRESTO is a consistent estimator of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ under suitable assumptions. Before stating Theorem 3, we present and briefly discuss some of the new assumptions we will make.

- **Assumption** $S(s, c)$: The distribution of $y_i \mid \boldsymbol{x}_i$ is distributed according to the PRESTO likelihood (7), where the true coefficients $\boldsymbol{\theta}_* = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\psi}_2^\top, \ldots, \boldsymbol{\psi}_{K-1}^\top\right)^\top \in \mathbb{R}^{p(K-1)}$ are s-sparse (have s nonzero entries for a fixed s not increasing in n or p). Further, $\|\boldsymbol{\theta}_*\|_\infty \leq c$ for a fixed c.

- **Assumption** $T(c)$: For all small enough $\rho > 0$, for all $\boldsymbol{\theta} \in \mathbb{R}^{p(K-1)}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_1 \leq \rho$ it holds that none of the decision boundaries defined by $\boldsymbol{\theta}$ and the true $\alpha_1, \ldots, \alpha_{K-1}$ cross in $\mathcal{S}$. Also, $\max_{k \in \{1, \ldots, K-1\}} |\alpha_k| \leq c$.

The fixed sparsity assumption $S(s, c)$ is helpful theoretically and also because without it in higher dimensions it becomes increasingly difficult to have nonparallel decision boundaries that do not cross. The first part of Assumption $T(c)$ can be interpreted to mean that none of the decision boundaries cross "too closely" to $\mathcal{S}$. Other than these aspects, Assumptions $S(s, c)$ and $T(c)$ are mild.

Theorem 3. *In a setting with fixed $K \geq 3$ and $p = p_n \to \infty$ as $n \to \infty$ and satisfying $p_n \leq C_1 n^{C_2}$ for some $C_1 > 0$ and $C_2 \in (0, 1)$, consider estimating* PRESTO *with penalty $\lambda_n = C_3 \log(p_n[K-1])/n$ for some $C_3 > 0$. Suppose Assumption $X(\mathbb{R}^{p_n})$ holds and there is some $C_4 < \infty$ such that $\sup_{\boldsymbol{x} \in \mathcal{S}} \|\boldsymbol{x}\|_\infty \leq C_4$ and Assumptions $S(s, C_4)$ and $T(C_4)$ hold. Assume for some fixed $b > 0$ it holds that $\lambda_{min}^* := \min_{k \in \{1, \ldots, K\}} \lambda_{min}(\boldsymbol{\Sigma}_k) > b$, where $\boldsymbol{\Sigma}_k := \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^\top \mid y_i = k\right]$. Then* PRESTO *is a consistent estimator of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$.*

Theorem 3 shows that under fairly mild regularity conditions and a sparsity assumption in a high-dimensional setting, PRESTO consistently estimates all of the decision boundaries. That is, it is consistent both if the proportional odds assumption holds and in more flexible settings, where the proportional odds model is unrealistic, under sparsity. Theorem 1 suggests this should be helpful for estimating rare class probabilities.

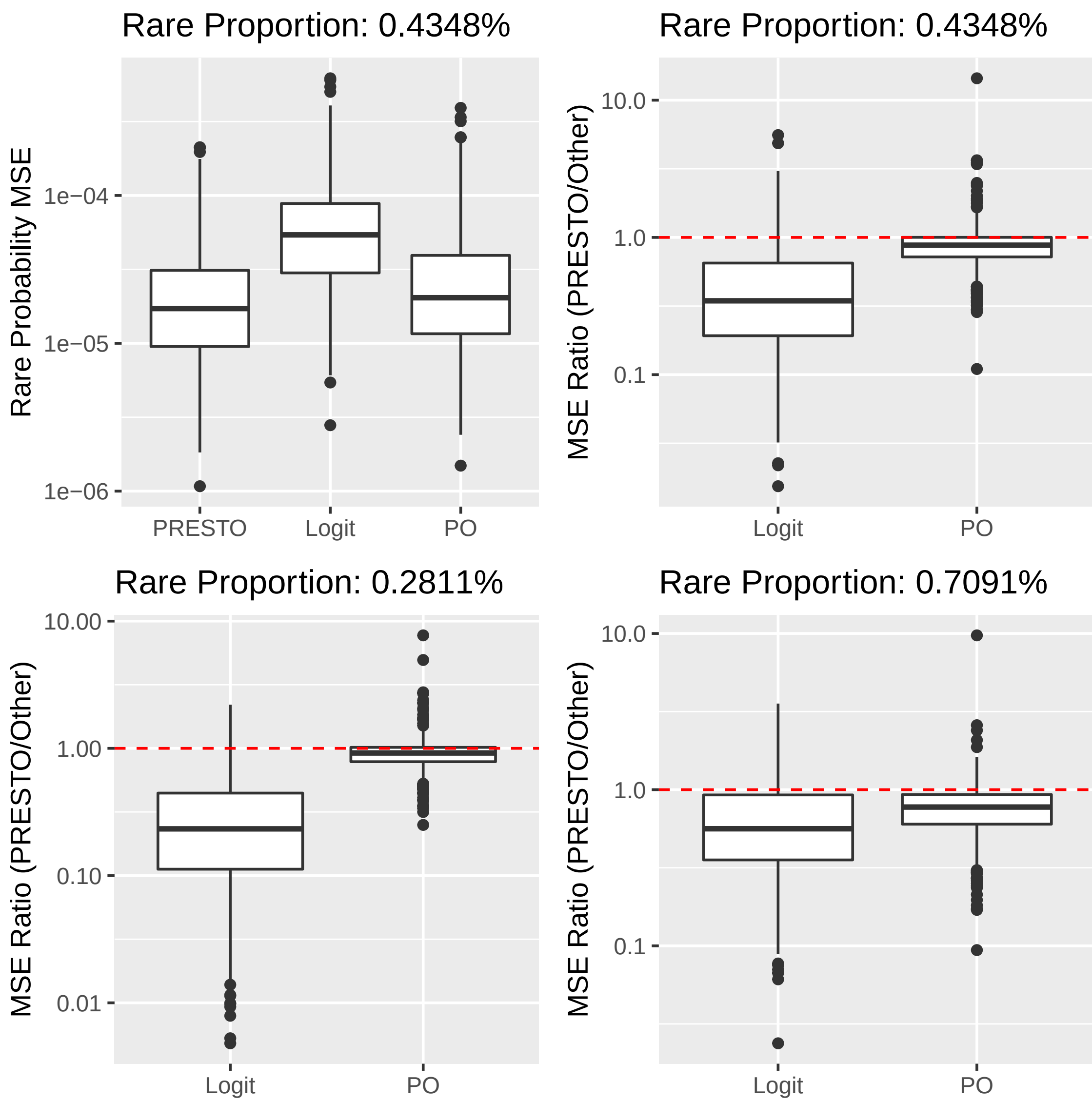

Figure 1: Top left: MSE of estimated rare class probabilities for each method across all $n = 2500$ observations, across 700 simulations, in sparse differences simulation setting of Section 4.1, for intercept setting yielding rare class probabilities of about 0.43% on average and sparsity $1/2$. Remaining plots: ratios of MSE for PRESTO *divided by MSE of each other method for each of three sets of intercepts with sparsity $1/2$ (*PRESTO *performs better if ratio is less than 1). All plots on log scale.*

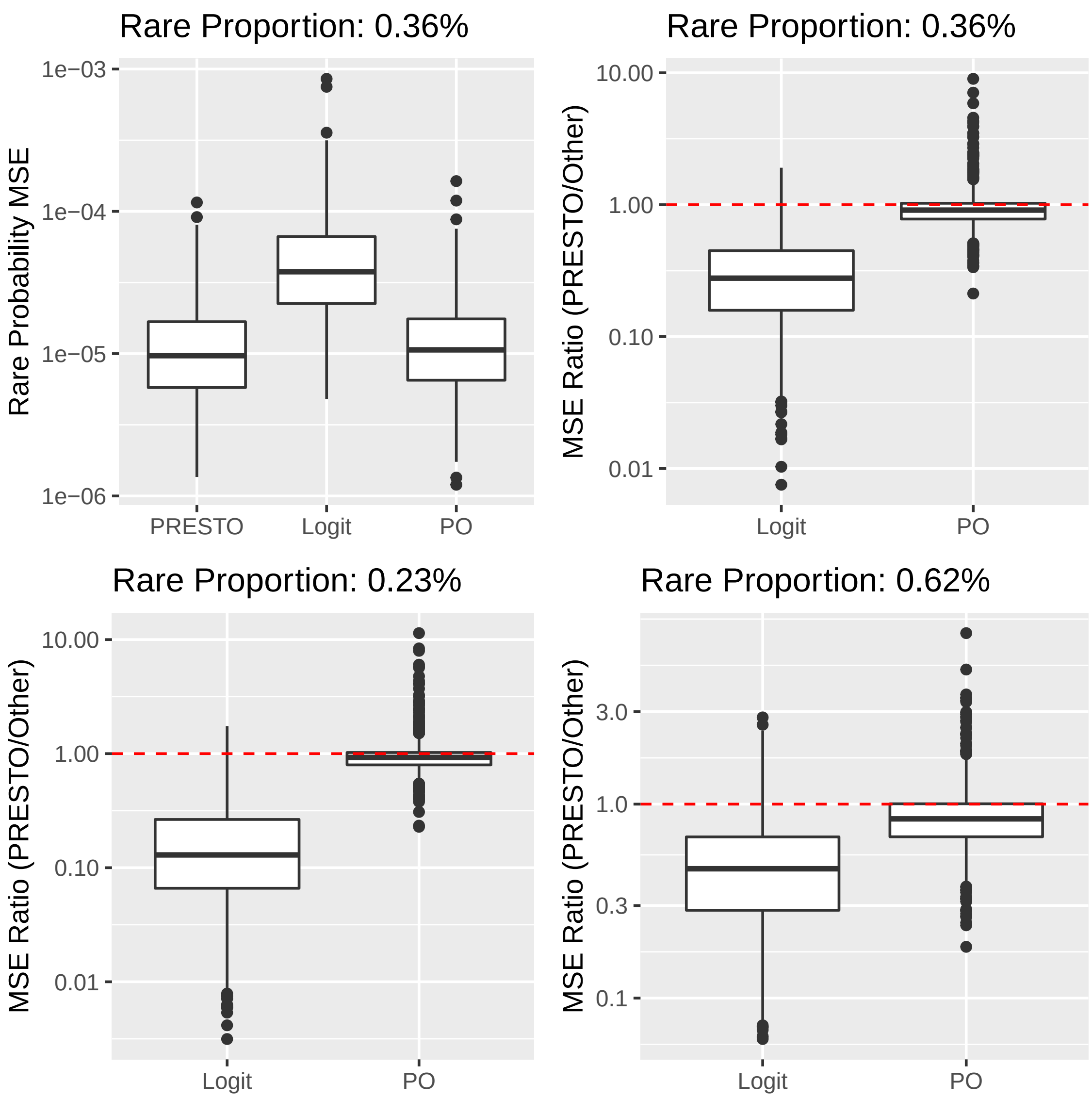

Figure 2: Same plots as in Figure 1, but for uniform differences synthetic experiment in Section 4.2.

4 Experiments

To illustrate the efficacy of PRESTO, we conduct two synthetic experiments and examine two real data sets[1]. In Section 4.1, we generate random $\boldsymbol{y}$ that have conditional probabilities based on a relaxation of the proportional odds model with sparse differences between adjacent decision boundary parameter vectors, rather than parameterizing all decision boundaries with the same $\boldsymbol{\beta}$. This setting is well-suited to PRESTO. In Section 4.2, we show that PRESTO also performs well in a less favorable setting, where the differences between adjacent decision boundaries are instead dense; nonetheless, PRESTO still outperforms logistic regression and proportional odds models. Lastly, in Sections 4.3 and 4.4 we compare the performance of PRESTO to logistic regression and proportional odds at estimating rare probabilities in real data experiments.

4.1 Simulated Data: Sparse Differences Setting

We repeat the following procedure for 700 simulations. First we generate data using $n = 2500$, $p = 10$, and $K = 4$. We draw a random $\boldsymbol{X} \in [-1, 1]^{n \times p}$, where $X_{ij} \sim \text{Uniform}(-1, 1)$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. Then $\boldsymbol{y} \in \{1, \ldots, K\}^n$ is generated according to a relaxation of the proportional odds model; instead of (1), we generate probabilities according to (3) where the $\boldsymbol{\beta}_k$ are generated in the following way for sparsity settings of $\eta = 1/3, 1/2$: first, we generate $\boldsymbol{\beta}_1$ by taking the vector $(0.5, \ldots, 0.5)^\top$, but setting all of the entries equal to 0 randomly with probability $1 - \eta$ for each entry independently. Then we set $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k, \qquad k \in \{2, \ldots, K-1\}$, where $\boldsymbol{\psi}_k \in \mathbb{R}^p$ are iid random vectors for each $k \in \{2, \ldots, K-1\}$ generated according to the following distribution:

$$\psi_{kj} = \begin{cases} 0, & \text{with probability } 1 - \eta, \\ 0.5, & \text{with probability } \eta/2, \\ -0.5, & \text{with probability } \eta/2, \end{cases} \qquad j \in \{1, \ldots, p\}.$$

We consider three possible sets of intercepts: $\boldsymbol{\alpha} = (0, 3.5, 5.5)$, $(0, 4, 6)$, and $(0, 4, 6.5)$, so that the first two categories are common and the remaining categories are rare. The final rare class is the one of interest; in the three settings, the average proportions of observations falling in the rare class are 0.61%, 0.37%, and 0.23%, respectively, for the $\eta = 1/3$ setting and 0.71%, 0.43%, and 0.28% for the $\eta = 1/2$ setting.

The fact that the decision boundaries may cross in the support of $\boldsymbol{X}$, which would mean that for such $\boldsymbol{x}$ some class probabilities are defined to be negative, puts practical limits on the magnitude of $\boldsymbol{\psi}_k$ in simulations. (See Section 3.6.1 of Agresti 2010 for a discussion.)

[1]Simulation studies were conducted in R Version 4.2.1 running on macOS 10.15.7. We also used the R packages `MASS` (version 7.3-58.1) and `ordinalNet` (version 2.12), available for download on CRAN, and the R `simulator` package, available for download at `https://github.com/jacobbien/simulator`. We implemented PRESTO by slightly modifying the code for `ordinalNet`; see Section G the appendix for details. We will release all code by the camera-ready deadline.

Also, for this reason, in each simulation we check whether or not the conditional probabilities are positive for each class for every sampled $\boldsymbol{x}$; if not, we generate new $\boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{K-1}$ for a limited number of iterations, ending the simulation study in failure if no suitable $\boldsymbol{\psi}_k$ can be found in a reasonable number of attempts. The parameters we used generated positive probabilities for all observations across all simulations.

We then estimate a model for each method; for logistic regression, we estimate the binary classification problem of whether or not each observation is in class K, and for proportional odds and PRESTO, we fit a full model on all K responses. For PRESTO, we use 5-fold cross-validation to choose a value of λ among 20 choices, selecting the λ with the best out-of-fold Brier score (other metrics, like negative log likelihood, failed because some values of λ in some folds resulted in models yielding negative probabilities, so these other metrics were undefined). The 20 candidate values of λ are generated in the following way: the largest λ value, λ_{20}, is the smallest λ for which all of the estimated sparse differences equal 0; the smallest λ value is set to $\lambda_1 = 0.01 \cdot \lambda_{20}$, and the remaining λ values are generated at equal intervals on a logarithmic scale between these two values.

Each of these models yields estimated probabilities that each observation lies in class K. In the final step of each simulation run, we compute the mean squared error of these estimated probabilities for each method.

In Figure 1, we show boxplots of the empirical mean squared errors for each method in the setting where the rare class is observed in 0.43% of observations when $\eta = 1/2$. In order to see how the methods compare pairwise on each simulation, we also show boxplots of the ratio between the mean squared error of PRESTO and the other two methods in each of the three simulation settings. We also conduct one-tailed paired t-tests of the alternative hypothesis that the mean MSE for PRESTO was lower than each of the competitor methods in each setting; all 12 of the p-values (provided in Table 3 of Appendix B) were below 10^{-5}. Finally, in Appendix B we also provide a table with the mean and standard errors for the MSE of each method in each simulation setting, as well as boxplots like the one in the top left corner of Figure 1 for the other two intercept settings and all boxplots for the $\eta = 1/3$ setting.

We see that PRESTO typically estimates these rare probabilities better than logistic regression, which despite being correctly specified struggles with class imbalance and does not draw strength from estimating the easier decision boundary between classes 1 and 2, and the proportional odds model, whose assumptions are not satisfied in this setting.

4.2 Simulated Data: Dense Differences Setting

In real data sets the differences between adjacent decision boundary parameter vectors may not always be sparse, so we conduct another synthetic experiment in the same way as in Section 4.1, except $\boldsymbol{\beta}_{1j} \sim \text{Uniform}(-.5, .5)$ and each $\psi_{kj} \sim \text{Uniform}(-.5, .5)$, iid across $j \in \{1, \ldots, p\}$ and $k \in \{2, \ldots, K-1\}$. This yields average rare class proportions of 0.62%, 0.36%, and 0.23% using the same intercepts as the experiments in Section 4.1. This setting can be considered “approximately" sparse in the sense that while no deviations will exactly equal 0, some will be large and important to estimate, and some will be essentially negligible.

Figure 2 and Table 1 summarize the results, along with additional figures and tables in Appendix B. We again see that PRESTO outperforms both competitor methods by statistically significant margins (except that PRESTO statistically ties the proportional odds model in one setting).

Table 1: Calculated p-values for one-tailed paired t-tests for uniform differences simulation setting of Section 4.2 (statistically significant p-values indicate better performance for PRESTO).

Rare Class Proportion	Logit p-value	PO p-value
0.62%	< 1e-04	< 1e-04
0.36%	< 1e-04	2.42e-04
0.23%	< 1e-04	0.21

4.3 Real Data Experiment 1: Soup Tasting

We conduct a real data experiment using the `soup` data set from the R `ordinal` package [R. H. B. Christensen, 2019]. The data come from a study [Christensen et al., 2011] of participants who tasted soups and responded whether they thought each soup was a reference product they had previously been familiarized with or a new test product. The respondents also stated how sure they were in their response on a three-level scale, yielding a total of $K = 6$ possible ordered outcomes for $n = 1847$ observations. The outcome of interest corresponds to the respondent being sure the tasted soup was the reference and is observed in 228 observations (about 12% of the total). All of the features are categorical, and after one-hot encoding we have $p = 22$ binary features related to the soup, the respondent, and the testing environment[2]. This may be a promising setting for PRESTO because though the responses have a well-defined ordering, it seems plausible that different features might have different effects at different levels of respondent sureness.

We complete the following procedure 350 times: first, we randomly split the data into training (90% of the data) and test (10%) sets. We estimate models using PRESTO, logistic regression, and the proportional odds model on the training data and evaluate on the test set.

We are interested in the accuracy of the rare class probabilities, but we can’t evaluate rare probability MSE since we don’t observe the true probabilities. Brier score could be a reasonable proxy, but it is known to be a poor metric in the presence of class imbalance

[2]The categorical predictors `PRODID` and `RESP` are omitted because in some splits not all levels of these features are observed in the training set, making it impossible to estimate parameters for these features.

[Benedetti, 2010]. Instead we estimate rare probability MSE using the following procedure. For each method, we sort the estimated test set rare class probabilities and assign the observations into 10 bins: the first 1/10 observations go in the first bin, and so on. Then we estimate the mean squared error of the estimated probabilities by $\frac{1}{n}\sum_{i=1}^{n}(o_{b(i)} - \hat{\pi}_1^{(i)})^2$, where $o_{b(i)}$ is the observed rare class proportion in the bin containing observation i and $\hat{\pi}_1^{(i)}$ is the estimated rare class probability for observation i. This is similar to *expected calibration error* [Naeini et al., 2015], though we use squared error rather than absolute error. 10 equal frequency bins follows the default of the R `CalibratR` package that implements expected calibration error [Schwarz and Heider, 2018].

By this metric, the mean error for PRESTO is 0.0096, 0.0157 for logistic regression and 0.0135 for proportional odds. Figure 3 displays boxplots of the results as in the synthetic experiments which indicate that PRESTO typically outperforms the other methods. We do not report p-values or standard errors since the observed samples are dependent (random splits of the same data set).

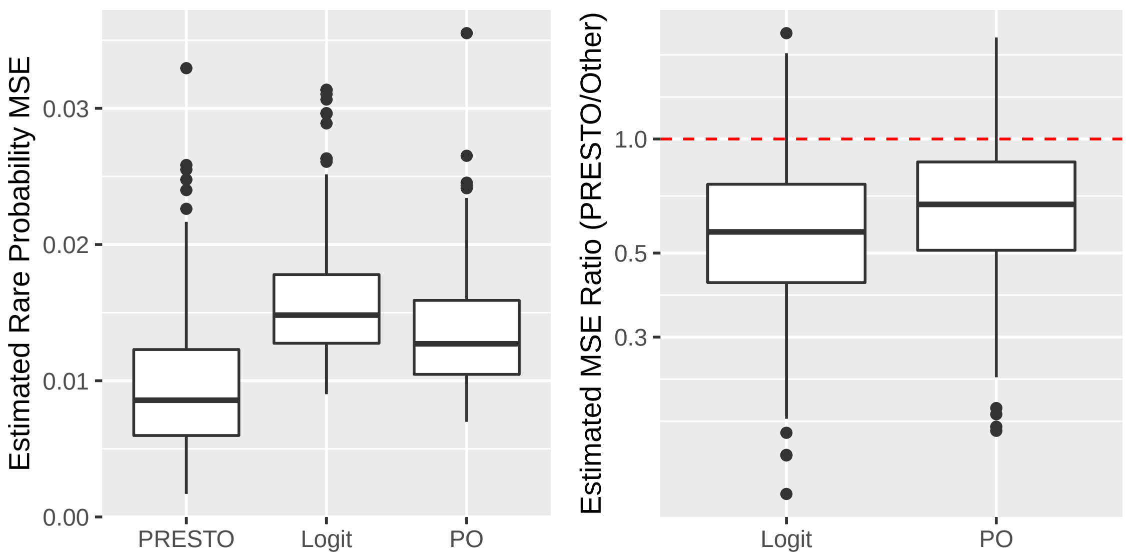

Figure 3: Left: Estimated MSEs of estimated rare class probabilities for each method across 280 random draws of training and test sets in real data experiment from Section 4.3. Right: ratios of estimated MSE for PRESTO *divided by MSE of each other method (*PRESTO *performs better if ratio is less than 1).*

4.4 Real Data Experiment 2: Diabetes

We present another real data experiment using the data set `PreDiabetes` from the R `MLDataR` package [Hutson et al., 2022]. This data set contains $n = 3059$ observations of

Table 2: Estimated rare class MSE for each method at each age cutoff in prediabetes real data experiment.

Age cutoff	PRESTO	Logit	PO
30	0.001095549	0.010443706	0.009903144
35	0.005306542	0.024368822	0.022397283
40	0.018049663	0.048436103	0.049834974
45	0.062173756	0.116159268	0.118422832
50	0.125441153	0.211305468	0.213431380
55	0.235600869	0.334510770	0.338530124
60	0.353343723	0.412752902	0.417347807
65	0.378428563	0.445375361	0.444832225

patients who were eventually diagnosed with diabetes. Each observation consists of the age at which the patient was diagnosed with prediabetes and diabetes as well as $p = 5$ covariates. Given an age a, we form an ordinal variable based on the patient's status of non-diabetic, prediabetic, or diabetic at age $a - 1$. We do this for ages $a \in \{30, 35, 40, \ldots, 65\}$. The number of patients diagnosed with diabetes increases with a, so varying a allows us to change the rarity of the rarest class in a natural way. The proportion of patients in the data diagnosed with diabetes before age $a = 30$ is 0.92, and 50.93% of the patients were diagnosed with diabetes before age $a = 65$.

We use PRESTO, logistic regression, and the proportional odds model to estimate the probability that each patient was diagnosed with diabetes before age a for each a. Much like our soup tasting data application, in each setting we take repeated random splits of the data, using 90% of the data selected at random for training and 10% for testing. In each iteration we again evaluate each method on the test data using the same estimator for mean squared error of the estimated rare class probabilities. We repeat this procedure 35 times in each of the 8 settings.

We display the results in a plot in Figure 4. We also provide the mean MSEs for each method at each age cutoff in Table 2. We see that PRESTO outperforms both logistic regression and the proportional odds model in all of these settings. (For age cutoffs $a = 29$ and below we were unable to estimate the proportional odds model on all subsamples because of the difficulty of having at least one observation from each class in both the training and test sets). PRESTO seems to outperform the other methods at all class rarities, though the performance gap increases as the rare class becomes less rare.

5 Conclusion

By leveraging data from earlier decision boundaries, but relaxing the rigid proportional odds assumption, PRESTO can substantially improve estimation of the probability of rare events, even when the assumption of sparse differences between adjacent decision boundary

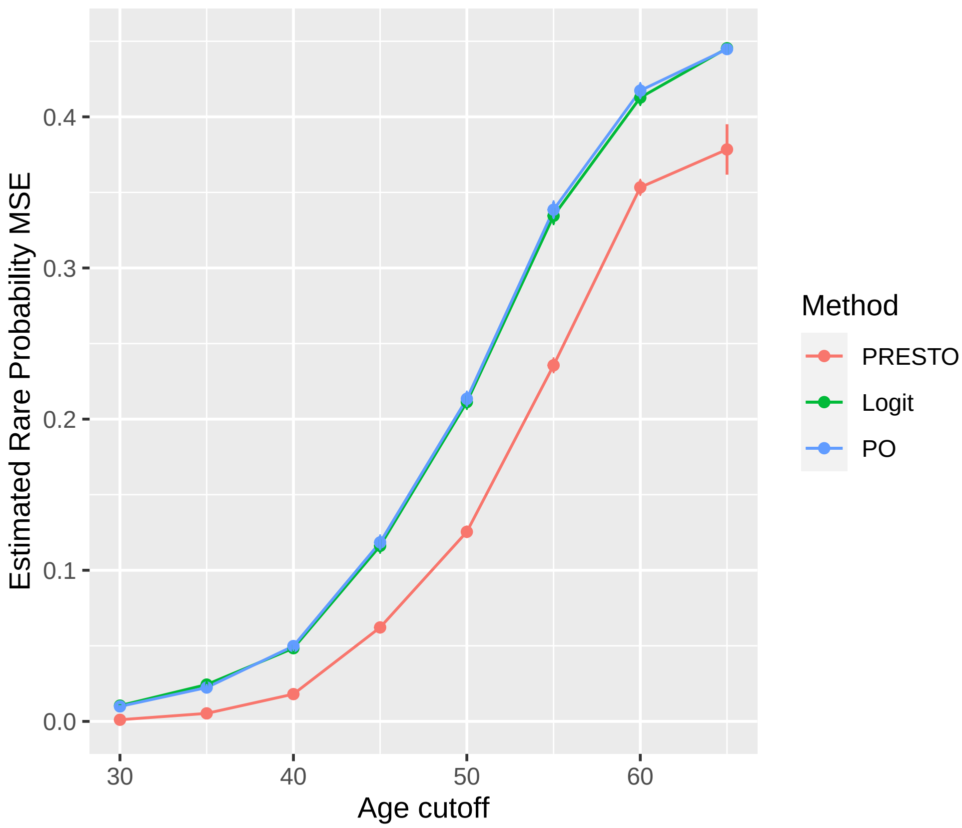

Figure 4: Left: Estimated MSEs of estimated rare class probabilities for each method across 280 random draws of training and test sets in real data experiment from Section 4.3. Right: ratios of estimated MSE for PRESTO *divided by MSE of each other method (*PRESTO *performs better if ratio is less than 1).*

weight vectors does not exactly hold. Future work could explore ℓ_1 penalties for the coefficients themselves, not just the differences between the coefficients, to allow for simultaneous feature selection and model estimation. Inference for PRESTO could also be possible by extending the method for exact post-selection inference for the generalized lasso path by Hyun et al. [2018] to our generalized linear model setting.

References

A. Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

B. G. Armstrong and M. Sloan. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1):191–204, 1989.

R. Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138(1): 203–211, 2010.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous Analysis of LASSO and Dantzig Selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. doi: 10.1214/08-AOS620. URL https://projecteuclid-org.libproxy1.usc.edu/download/pdfview_1/euclid.aos/1245332830.

R. Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, pages 1171–1178, 1990.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

R. H. B. Christensen, G. Cleaver, and P. B. Brockhoff. Statistical and Thurstonian models for the A-not A protocol with and without sureness. *Food Quality and Preference*, 22 (6):542–549, 2011. ISSN 09503293. doi: 10.1016/j.foodqual.2011.03.003. URL http://dx.doi.org/10.1016/j.foodqual.2011.03.003.

B. A. Duncan and C. P. Elkan. Probabilistic modeling of a sales funnel to prioritize leads. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1751–1758, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788578. URL https://doi.org/10.1145/2783258.2788578.

K. Ekvall and M. Bottai. Concave likelihood-based regression with finite-support response variables. *Biometrics*, (March):1–12, 2022. ISSN 0006-341X. doi: 10.1111/biom.13760.

A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

W. H. Greene. *Econometric Analysis*. Pearson Education, 7th edition, 2012.

B. Hansen. *Econometrics*. Princeton University Press, 2022. ISBN 9780691235899. URL `https://books.google.com/books?id=Pte7zgEACAAJ`.

T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143, 2015.

X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the eighth international workshop on data mining for online advertising*, pages 1–9, 2014.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/9781139020411.

G. Hutson, A. Laldin, and I. Velásquez. *MLDataR: Collection of Machine Learning Datasets for Supervised Machine Learning*, 2022. URL `https://CRAN.R-project.org/package=MLDataR`. R package version 0.1.3.

S. Hyun, M. G'Sell, and R. J. Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. doi: 10.1186/s40537-019-0192-5. URL `https://doi.org/10.1186/s40537-019-0192-5`.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

C. M. Norris, W. A. Ghali, L. D. Saunders, R. Brant, D. Galbraith, P. Faris, M. L. Knudtson, A. Investigators, et al. Ordinal regression model and the linear regression model were superior to the logistic regression models. *Journal of clinical epidemiology*, 59(5): 448–456, 2006.

A. B. Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(4), 2007.

B. Peterson and F. E. Harrell Jr. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(2): 205–217, 1990.

W. Pößnecker and G. Tutz. A general framework for the selection of effect type in ordinal regression, 2016. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-26912-0.

J. W. Pratt. Concavity of the log likelihood. *Journal of the American Statistical Association*, 76(373):103–106, 1981. ISSN 1537274X. doi: 10.1080/01621459.1981.10477613.

R. H. B. Christensen. ordinal—Regression Models for Ordinal Data, 2019. URL https://CRAN.R-project.org/package=ordinal.

J. Schwarz and D. Heider. GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making. *Bioinformatics*, 35(14):2458–2465, 11 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty984. URL https://doi.org/10.1093/bioinformatics/bty984.

R. Serfling. *Approximation theorems of mathematical statistics.* Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980. ISBN 0471024031. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

G. Tutz and J. Gertheiss. Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200, 2016.

E. R. Ugba, D. Mörlein, and J. Gertheiss. Smoothing in ordinal regression: An application to sensory data. *Stats*, 4(3):616–633, 2021.

A. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504. URL https://books.google.com/books?id=UEuQEM5RjWgC.

R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

T. Von Wachter, M. Bertrand, H. Pollack, J. Rountree, and B. Blackwell. Predicting and preventing homelessness in los angeles. *California Policy Lab and University of Chicago Poverty Lab*, 2019.

J. M. Wooldridge. *Econometric analysis of cross section and panel data.* MIT Press, Cambridge and London, 2002. ISBN 0262232197.

M. J. Wurm, P. J. Rathouz, and B. M. Hanlon. Regularized ordinal regression and the ordinalnet r package. *arXiv preprint arXiv:1706.05003*, 2017.

M. J. Wurm, P. J. Rathouz, and B. M. Hanlon. Regularized ordinal regression and the ordinalNet R package. *Journal of Statistical Software*, 99(6):1–42, 2021.

F. Zhang. *The Schur Complement and Its Applications.* Numerical Methods and Algorithms. Springer, 2005. ISBN 9780387242712. URL `https://books.google.com/books?id=Wjd8_AwjiIIC`.

W. Zhang, S. Yuan, and J. Wang. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1077–1086, 2014.

We state Theorem 4, which shows that logistic regression does arbitrarily badly at estimating rare class probabilities as the class imbalance increases, in Section A. In Section B, we display summary statistics and additional figures for the observed mean squared errors (MSEs) for each method from the synthetic data experiments from Sections 4.1 and 4.2. We provide the proofs of Theorems 1 and 4 in Section C. In Section D, we present synthetic data experiments and analysis justifying the validity of one of the assumptions of Theorem 2 in Remark 3, and we then prove Theorem 2. Theorems 1, 4, and 2 depend on Lemma 5, which is stated at the beginning of Section C and proven in Section E. We prove Theorem 3 in Section F. Finally, in Section G we provide implementation details for estimating PRESTO.

A Theorem 4

It is well-known that class imbalance poses a major challenge for classifiers. Theorem 4 exhibits this concretely for logistic regression.

Theorem 4. *Assume* $X(\mathbb{R}^p, \lambda_{max})$ *and* $Y(2)$ *hold. Let* $\pi(\boldsymbol{x}) := \mathbb{P}(y = 2 \mid \boldsymbol{x})$*, and assume that* $\sup_{\boldsymbol{x} \in \mathcal{S}} \pi(\boldsymbol{x}) = \pi_{rare}$ *for some* $\pi_{rare} \leq 1/2$*. Then*

1. *for any fixed* $\boldsymbol{v} \in \mathbb{R}^{p+1}$*,*

$$\frac{1}{\boldsymbol{v}^\top \boldsymbol{v}} \text{Asym.MSE} \left((\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top) \boldsymbol{v} \right) \geq \frac{1}{\lambda_{max} \pi_{rare}},$$

   and

2. *for any fixed* $\boldsymbol{z} \in \mathcal{S}$*,*

$$\text{Asym.MSE} \left(\frac{\hat{\pi}(\boldsymbol{z})}{\pi(\boldsymbol{z})} \right) \geq \frac{1 - \pi_{rare}}{\pi_{rare}} \frac{1}{\lambda_{max}}.$$

Proof. Provided in Section C.2. □

To give an example of applying part 1 of this result, consider the choice $\boldsymbol{v} = (0, 1, 0, \ldots, 0)$. Then we have that $\text{Asym.MSE}\left(\hat{\beta}_1\right) \geq 1/(\lambda_{\text{max}} \pi_{\text{rare}})$, so $\hat{\beta}_1$ (or any other estimated coefficient) has arbitrarily large asymptotic mean squared error as π_{rare} vanishes. Part 2 shows that the same thing happens to the asymptotic mean squared error for the estimated probabilities of the logistic regression estimator, when scaled by $\pi(\boldsymbol{z})$.

B More Simulation Results

For more results from the synthetic experiments, see Tables 3, 4, and 5, and Figures 7, 6, 5, and 8.

Table 3: Calculated p-values for one-tailed paired t-tests for sparse differences simulation setting of Section 4.1 (statistically significant p-values indicate better performance for PRESTO).

Rare Prop.	Sparsity	Logit p-value	PO p-value
0.61%	1/3	5.19e-74	4.21e-19
0.71%	1/2	8.68e-48	3.38e-35
0.37%	1/3	3.08e-61	1.65e-03
0.43%	1/2	3.75e-64	2.57e-11
0.23%	1/3	3.34e-38	2.85e-06
0.28%	1/2	4.01e-52	4.71e-06

Table 4: Means and standard errors of empirical MSEs for each method in each of three intercept settings in the sparse differences synthetic experiment setting of Section 4.1.

Rare Class Proportion	Sparsity	PRESTO	Logistic Regression	Proportional Odds
0.61%	1/3	3.03e-05 (1.1e-06)	6.90e-05 (2.1e-06)	3.64e-05 (1.4e-06)
0.71%	1/2	5.22e-05 (1.9e-06)	8.89e-05 (2.5e-06)	7.50e-05 (3e-06)
0.37%	1/3	1.40e-05 (6e-07)	5.66e-05 (2.4e-06)	1.49e-05 (6.1e-07)
0.43%	1/2	2.63e-05 (1e-06)	7.21e-05 (2.7e-06)	3.17e-05 (1.4e-06)
0.23%	1/3	6.15e-06 (2.6e-07)	5.74e-05 (3.8e-06)	6.56e-06 (2.8e-07)
0.28%	1/2	1.39e-05 (6.3e-07)	6.77e-05 (3.4e-06)	1.58e-05 (7.1e-07)

Table 5: Means and standard errors of empirical MSEs for each method in each of three intercept settings in the uniform differences synthetic experiment setting of Section 4.2.

Rare Class Proportion	PRESTO	Logistic Regression	Proportional Odds
0.62%	2.86e-05 (8.3e-07)	6.51e-05 (1.7e-06)	3.43e-05 (9.8e-07)
0.36%	1.33e-05 (4.4e-07)	5.36e-05 (2.2e-06)	1.43e-05 (5e-07)
0.23%	5.96e-06 (2.3e-07)	6.24e-05 (3.8e-06)	6.07e-06 (2.1e-07)

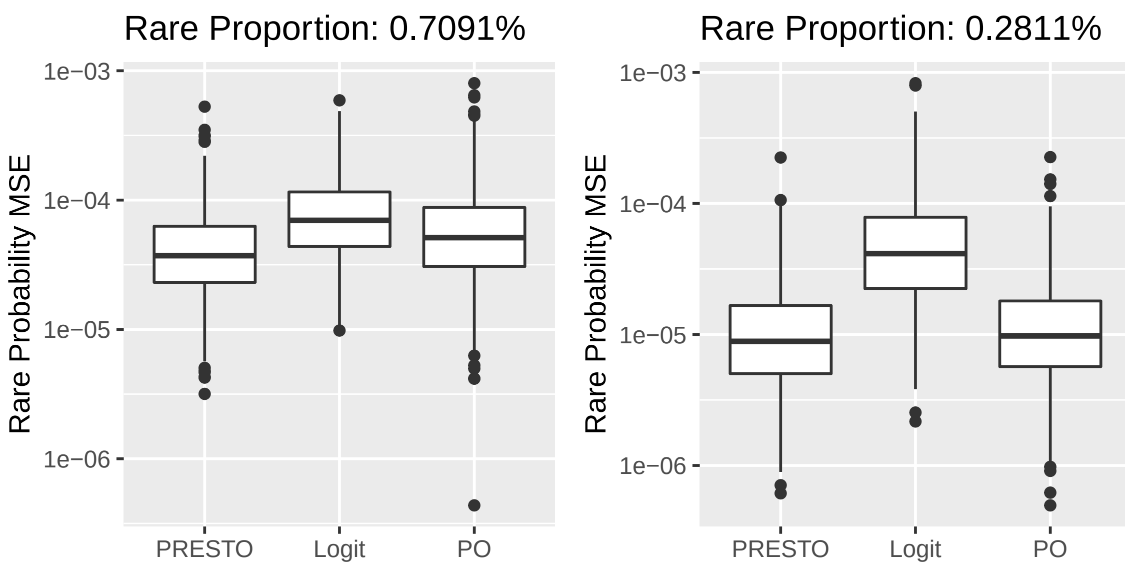

Figure 5: MSE of predicted rare class probabilities for each method across all $n = 2500$ *observations, across 700 simulations, in sparse differences synthetic experiment setting of Section 4.1 with sparsity* $1/2$. *(These plots are for the two intercept settings that weren't shown in the main text for the sparsity setting of* $1/2$. *All plots on a log scale.)*

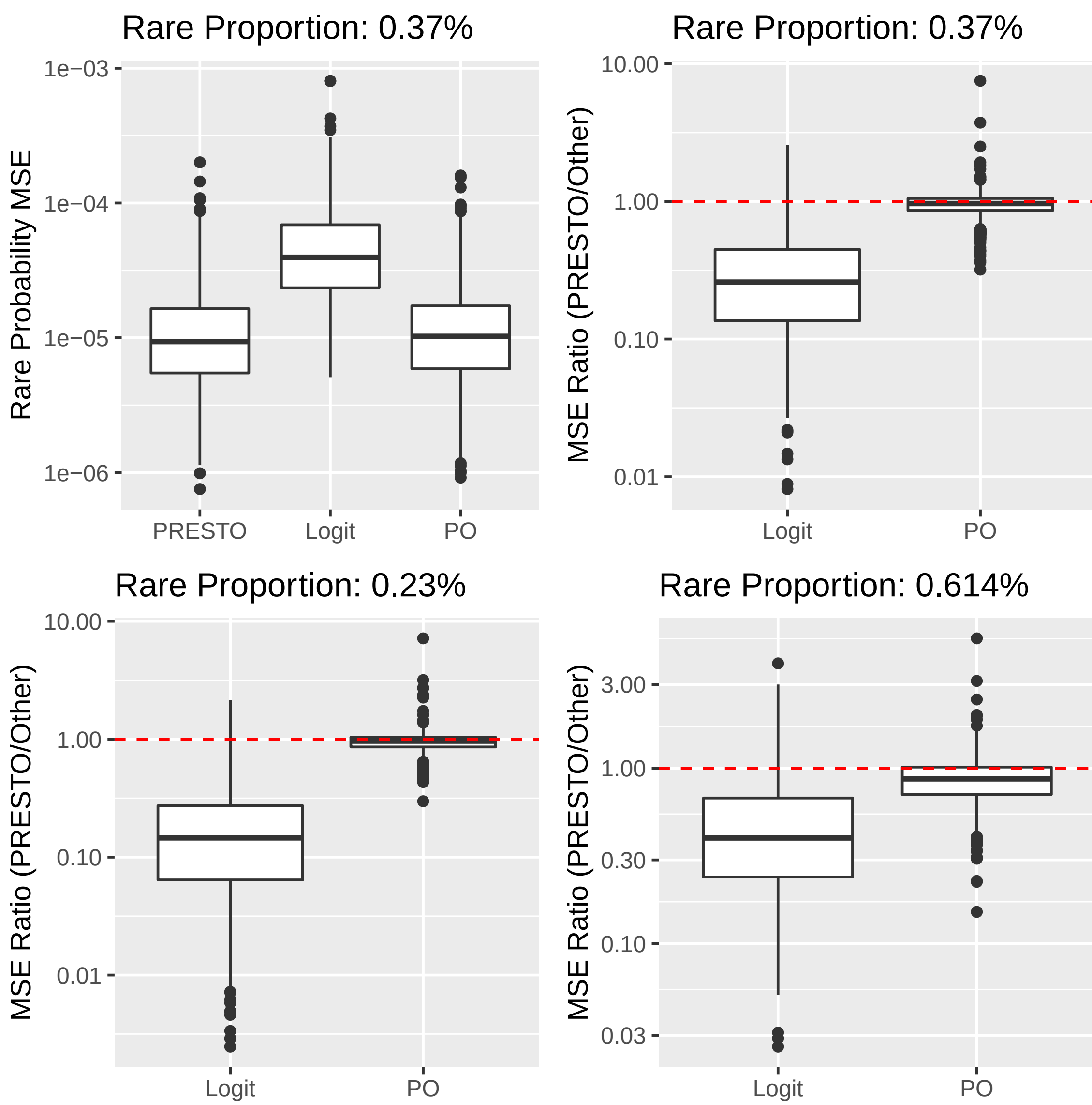

Figure 6: Same as Figure 1, but for the simulations with sparsity 1/3.

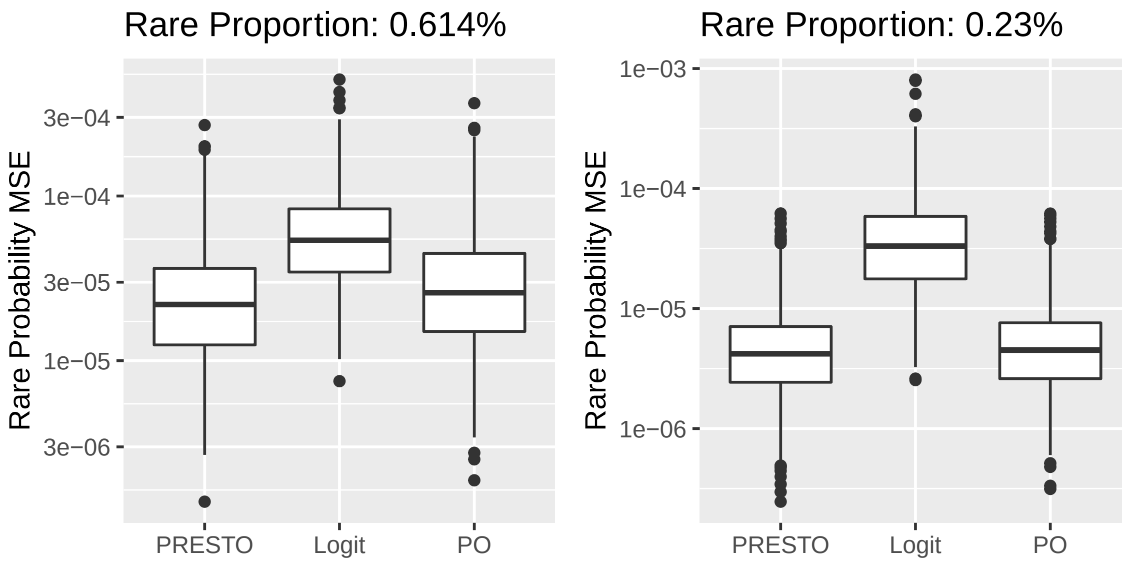

Figure 7: Same as Figure 5, but for the simulations with sparsity 1/2.

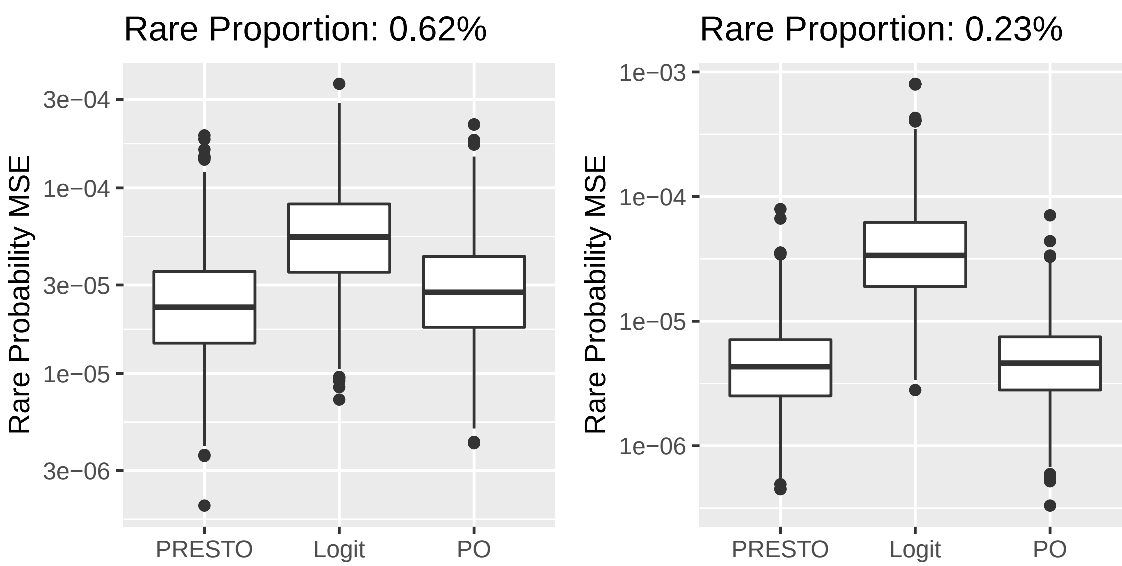

Figure 8: MSE of predicted rare class probabilities for each method across all $n = 2500$ *observations, across 700 simulations, in uniform differences synthetic experiment setting of Section 4.2. (These plots are for the two intercept settings that weren't shown in the main text.)*

C Statement of Lemma 5 and Proofs of Theorems 4 and 1

In Section C.1 we state Lemma 5, and we prove Theorems 4 and 1 in Section C.2.

C.1 Statement of Lemma 5

Theorems 4, 1, and 2 relate to the asymptotic covariance matrices of the maximum likelihood estimators of the parameters of the proportional odds and logistic regression models. Under mild regularity conditions, the asymptotic covariance matrix of any maximum likelihood estimator (when scaled by $\sqrt{n}$) is known to be the inverse of the Fisher information matrix

$$-\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathcal{L}(\boldsymbol{\theta})\right],$$

where $\boldsymbol{\theta}$ are the parameters estimated by the model and $\mathcal{L}(\boldsymbol{\theta})$ is the log likelihood [Serfling, 1980, Section 4.2.2]. In the proof of Lemma 5, we calculate these Fisher information matrices for the proportional odds and logistic regression models and verify the needed regularity conditions.

Lemma 5. Assume that no class has probability 0 for any $\boldsymbol{x} \in \mathcal{S}$ (equivalently, assume that all of the intercepts in the proportional odds model (1) are not equal, so $\alpha_1 < \ldots < \alpha_{K-1}$). Assume that $dF(\boldsymbol{x})$ has bounded support.

1. The Fisher information matrix for the maximum likelihood estimator of the proportional odds model (1) is

$$I^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} I_{\alpha\alpha}^{\text{prop. odds}} & \left(I_{\beta\alpha}^{\text{prop. odds}}\right)^\top \\ I_{\beta\alpha}^{\text{prop. odds}} & I_{\beta\beta}^{\text{prop. odds}} \end{pmatrix} \in \mathbb{R}^{(K-1+p)\times(K-1+p)}$$

where

$$I_{\alpha\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{pmatrix} M_1 & -\tilde{M}_2 & 0 & \cdots & 0 & 0 \\ -\tilde{M}_2 & M_2 & -\tilde{M}_3 & \cdots & 0 & 0 \\ 0 & -\tilde{M}_3 & M_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & M_{K-2} & -\tilde{M}_{K-1} \\ 0 & 0 & 0 & \cdots & -\tilde{M}_{K-1} & M_{K-1} \end{pmatrix}, \tag{8}$$

$$I_{\beta\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{pmatrix} J_1^{\boldsymbol{x}} + \tilde{J}_2^{\boldsymbol{x}} \\ J_2^{\boldsymbol{x}} + \tilde{J}_3^{\boldsymbol{x}} \\ \vdots \\ J_{K-1}^{\boldsymbol{x}} + \tilde{J}_K^{\boldsymbol{x}} \end{pmatrix}, \quad \text{and} \tag{9}$$

$$I_{\beta\beta}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_k^{\boldsymbol{x}\boldsymbol{x}^\top} \right), \tag{10}$$

where

$$M_k := \int [p_k(\boldsymbol{x})(1-p_k(\boldsymbol{x}))]^2 \left(\frac{1}{\pi_k(\boldsymbol{x})} + \frac{1}{\pi_{k+1}(\boldsymbol{x})} \right) dF(\boldsymbol{x}), \qquad k \in \{1,\ldots,K-1\}, \tag{11}$$

$$\tilde{M}_k := \int p_k(\boldsymbol{x})(1-p_k(\boldsymbol{x}))p_{k-1}(\boldsymbol{x})(1-p_{k-1}(\boldsymbol{x})) \cdot \frac{1}{\pi_k(\boldsymbol{x})} \, dF(\boldsymbol{x}), \qquad k \in \{2,\ldots,K-1\} \tag{12}$$

and

$$\begin{aligned} J_k &:= \int \pi_k(\boldsymbol{x})p_k(\boldsymbol{x})[1-p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}, \\ J_k^{\boldsymbol{x}} &:= \int \boldsymbol{x}\pi_k(\boldsymbol{x})p_k(\boldsymbol{x})[1-p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}^p, \\ J_k^{\boldsymbol{x}\boldsymbol{x}^\top} &:= \int \boldsymbol{x}\boldsymbol{x}^\top\pi_k(\boldsymbol{x})p_k(\boldsymbol{x})[1-p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}^{p\times p}, \\ \tilde{J}_k &:= \int \pi_k(\boldsymbol{x})p_{k-1}(\boldsymbol{x})[1-p_{k-1}(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}, \\ \tilde{J}_k^{\boldsymbol{x}} &:= \int \boldsymbol{x}\pi_k(\boldsymbol{x})p_{k-1}(\boldsymbol{x})[1-p_{k-1}(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}^p, \qquad \text{and} \\ \tilde{J}_k^{\boldsymbol{x}\boldsymbol{x}^\top} &:= \int \boldsymbol{x}\boldsymbol{x}^\top\pi_k(\boldsymbol{x})p_{k-1}(\boldsymbol{x})[1-p_{k-1}(\boldsymbol{x})] \, dF(\boldsymbol{x}) \in \mathbb{R}^{p\times p} \end{aligned}$$

for all $k \in [K]$.

2. The Fisher information matrix for the maximum likelihood estimator of the logistic regression model predicting whether or not each observation is in class 1 is

$$I^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = \begin{pmatrix} I_{\alpha\alpha}^{\text{logistic}} & \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \\ I_{\beta\alpha}^{\text{logistic}} & I_{\beta\beta}^{\text{logistic}} \end{pmatrix} = \mathbb{E}\left[\pi_1(\boldsymbol{X})[1-\pi_1(\boldsymbol{X})]\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right] \in \mathbb{R}^{(p+1)\times(p+1)} \tag{13}$$

where $\tilde{\boldsymbol{X}} := \begin{pmatrix}\mathbf{1} & \boldsymbol{X}\end{pmatrix}$ (an n-vector of all ones followed by $\boldsymbol{X}$) and

$$I_{\alpha\alpha}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = M_1^{\text{logistic}}, \tag{14}$$

$$I_{\beta\alpha}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = J_1^{\boldsymbol{x};\text{ logistic}} + \tilde{J}_2^{\boldsymbol{x};\text{ logistic}}, \qquad \text{and} \tag{15}$$

$$I_{\beta\beta}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = J_1^{\boldsymbol{x}\boldsymbol{x}^\top;\text{ logistic}} + \tilde{J}_2^{\boldsymbol{x}\boldsymbol{x}^\top;\text{ logistic}}, \tag{16}$$

where we define

$$M_1^{\text{logistic}} := \int \pi_1(\boldsymbol{x})(1-\pi_1(\boldsymbol{x}))\ dF(\boldsymbol{x}) \tag{17}$$

and

$$\begin{aligned}
J_1^{\boldsymbol{x};\text{ logistic}} &:= \int \boldsymbol{x}\pi_1(\boldsymbol{x})^2[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) \in \mathbb{R}^p = J_1^{\boldsymbol{x}}, \\
J_1^{\boldsymbol{x}\boldsymbol{x}^\top;\text{ logistic}} &:= \int \boldsymbol{x}\boldsymbol{x}^\top\pi_1(\boldsymbol{x})^2[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) \in \mathbb{R}^{p\times p} = J_1^{\boldsymbol{x}\boldsymbol{x}^\top}, \\
\tilde{J}_2^{\boldsymbol{x};\text{ logistic}} &:= \int \boldsymbol{x}\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]^2\ dF(\boldsymbol{x}) \in \mathbb{R}^p, \qquad \text{and} \\
\tilde{J}_2^{\boldsymbol{x}\boldsymbol{x}^\top;\text{ logistic}} &:= \int \boldsymbol{x}\boldsymbol{x}^\top\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]^2\ dF(\boldsymbol{x}) \in \mathbb{R}^{p\times p}.
\end{aligned}$$

3. The information matrices $I^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $I^{\text{logistic}}(\alpha_1, \boldsymbol{\beta})$ are finite and positive definite, and the following convergences hold:

$$\sqrt{n}\cdot\left(\hat{\boldsymbol{\theta}}_1^{\text{prop. odds}} - \boldsymbol{\theta}_1\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(I^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)^{-1}\right)$$

and

$$\sqrt{n}\cdot\left(\hat{\boldsymbol{\theta}}_1^{\text{logistic}} - \boldsymbol{\theta}_1\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(I^{\text{logistic}}(\alpha_1, \boldsymbol{\beta})\right)^{-1}\right).$$

Further, because these information matrices are symmetric and positive definite, by Observation 7.1.2 in Horn and Johnson [2012] the principal submatrices $I_{\alpha\alpha}^{\text{prop. odds}} := I_{\alpha\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $I_{\beta\beta}^{\text{prop. odds}} := I_{\beta\beta}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and $I_{\beta\beta}^{\text{logistic}} := I_{\beta\beta}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta})$ are all positive definite.

Proof. Provided in Section E. □

Remark 2. Note that $J_K = 0$ because $1 - p_K(\boldsymbol{x}) = 1 - \mathbb{P}(y(\boldsymbol{x}) \leq K \mid \boldsymbol{x}) = 0$ for all $\boldsymbol{x}$, and similarly for $J_K^{\boldsymbol{x}}$ and $J_K^{\boldsymbol{x}\boldsymbol{x}^\top}$. Likewise, $\tilde{J}_1 = 0$ because $p_0(\boldsymbol{x}) = \mathbb{P}(y(\boldsymbol{x}) \leq 0 \mid \boldsymbol{x}) = 0$ for all $\boldsymbol{x}$, and similarly for $\tilde{J}_1^{\boldsymbol{x}}$ and $\tilde{J}_1^{\boldsymbol{x}\boldsymbol{x}^\top}$.

We also take a moment to briefly establish some identities we will use later. For any $k \in \{1, \ldots, K-1\}$,

$$\begin{aligned} J_k^{\boldsymbol{x}} + \tilde{J}_{k+1}^{\boldsymbol{x}} = & \int \boldsymbol{x} \pi_k(\boldsymbol{x}) p_k(\boldsymbol{x})[1 - p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}) + \int \boldsymbol{x} \pi_{k+1}(\boldsymbol{x}) p_k(\boldsymbol{x})[1 - p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}) \\ = & \int \boldsymbol{x} \left[\pi_k(\boldsymbol{x}) + \pi_{k+1}(\boldsymbol{x})\right] p_k(\boldsymbol{x})[1 - p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}). \end{aligned} \tag{18}$$

Similarly,

$$J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_{k+1}^{\boldsymbol{x}\boldsymbol{x}^\top} = \int \boldsymbol{x}\boldsymbol{x}^\top \left[\pi_k(\boldsymbol{x}) + \pi_{k+1}(\boldsymbol{x})\right] p_k(\boldsymbol{x})[1 - p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}),$$

so from (10) we have

$$\begin{aligned} I_{\beta\beta}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = & \sum_{k=1}^{K} \left(J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_k^{\boldsymbol{x}\boldsymbol{x}^\top} \right) \\ = & \tilde{J}_1^{\boldsymbol{x}\boldsymbol{x}^\top} + \sum_{k=1}^{K-1} \left(J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_{k+1}^{\boldsymbol{x}\boldsymbol{x}^\top} \right) + J_K^{\boldsymbol{x}\boldsymbol{x}^\top} \\ = & \int \boldsymbol{x}\boldsymbol{x}^\top \pi_1(\boldsymbol{x}) \underbrace{p_0(\boldsymbol{x})}_{=\mathbb{P}(y \leq 0 | \boldsymbol{x}) = 0} [1 - p_0(\boldsymbol{x})] \, dF(\boldsymbol{x}) + \sum_{k=1}^{K-1} \left(J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_{k+1}^{\boldsymbol{x}\boldsymbol{x}^\top} \right) \\ & + \int \boldsymbol{x}\boldsymbol{x}^\top \pi_K(\boldsymbol{x}) p_K(\boldsymbol{x}) \left[1 - \underbrace{p_K(\boldsymbol{x})}_{=\mathbb{P}(y \leq K | \boldsymbol{x}) = 1} \right] \, dF(\boldsymbol{x}) \\ = & \sum_{k=1}^{K-1} \int \boldsymbol{x}\boldsymbol{x}^\top \left[\pi_k(\boldsymbol{x}) + \pi_{k+1}(\boldsymbol{x})\right] p_k(\boldsymbol{x})[1 - p_k(\boldsymbol{x})] \, dF(\boldsymbol{x}). \end{aligned} \tag{19}$$

C.2 Proofs of Theorems 4 and 1

Equipped with the results of Lemma 5, we proceed to prove Theorems 4 and 1. (Recall that for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of equal dimension p, we say $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.)

Proof of Theorem 4. 1. Note that the assumptions of Lemma 5 are satisfied. Since the inverse of the asymptotic covariance matrix is

$$\mathbb{E}\left[\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right] \preceq \pi_{\text{rare}}(1-\pi_{\text{rare}})\mathbb{E}\left[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right]$$

(where the second step is valid because $t \mapsto t(1-t)$ is monotone increasing in t for $t \in [0, 1/2]$), by Corollary 7.7.4 in Horn and Johnson [2012] the largest eigenvalue of the inverse of the asymptotic covariance matrix is no larger than $\pi_{\text{rare}}(1-\pi_{\text{rare}})\lambda_{\max}$. Therefore

$$\text{Asym.Cov}\left((\sqrt{n}\cdot\hat{\alpha}, \sqrt{n}\cdot\hat{\boldsymbol{\beta}})\right) = \left(\mathbb{E}\left[\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right]\right)^{-1} \succeq \left(\pi_{\text{rare}}(1-\pi_{\text{rare}})\mathbb{E}\left[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right]\right)^{-1}$$

has smallest eigenvalue at least $1/[\lambda_{\max}\pi_{\text{rare}}(1-\pi_{\text{rare}})]$, which is larger than $1/[\lambda_{\max}\pi_{\text{rare}}]$, again using Corollary 7.7.4 in Horn and Johnson [2012]. So for any $\boldsymbol{v} \in \mathbb{R}^{p+1}$,

$$\text{Asym.Var}\left((\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v}\right) = \boldsymbol{v}^\top \text{Asym.Cov}\left((\sqrt{n}\cdot\hat{\alpha}, \sqrt{n}\cdot\hat{\boldsymbol{\beta}})\right)\boldsymbol{v} \geq \frac{\boldsymbol{v}^\top\boldsymbol{v}}{\lambda_{\max}\pi_{\text{rare}}}.$$

Finally, since we have already shown that $(\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)$ is asymptotically unbiased, the asymptotic MSE is equal to this asymptotic variance:

$$\begin{aligned}
\text{Asym.MSE}((\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v})) = \; & \mathbb{E}\left[\lim_{n\to\infty}\left(\sqrt{n}\cdot\left[(\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v} - (\alpha, \boldsymbol{\beta}^\top)\boldsymbol{v}\right]\right)^2\right] \\
= \; & \mathbb{E}\Bigg[\lim_{n\to\infty}\left(\sqrt{n}\cdot\left[\mathbb{E}\left[(\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v}\right] - (\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v}\right]\right)^2 \\
& + \lim_{n\to\infty}\left(\sqrt{n}\cdot\left[(\alpha, \boldsymbol{\beta}^\top)\boldsymbol{v} - \mathbb{E}[(\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v}\right]\right)^2\Bigg] \\
= \; & \text{Asym.Var}\left(\sqrt{n}\left[(\alpha, \boldsymbol{\beta}^\top)\boldsymbol{v} - (\hat{\alpha}, \hat{\boldsymbol{\beta}}^\top)\boldsymbol{v}\right]\right) + 0 \\
\geq \; & \frac{\boldsymbol{v}^\top\boldsymbol{v}}{\lambda_{\max}\pi_{\text{rare}}}.
\end{aligned}$$

2. Because $(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \mapsto \hat{\pi}(\boldsymbol{z})$ is differentiable for all $\boldsymbol{z} \in \mathbb{R}^p$, by the delta method (Theorem 3.1 in van der Vaart 2000)

$$\sqrt{n}\cdot[\hat{\pi}(\boldsymbol{z}) - \pi(\boldsymbol{z})] \xrightarrow{d} \mathcal{N}\left(0, \pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\left(1, \boldsymbol{z}^\top\right)\left(I^{\text{logistic}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)^{-1}\left(1, \boldsymbol{z}^\top\right)^\top\right)$$

for any $\boldsymbol{z} \in \mathbb{R}^p$. Therefore

$$\begin{aligned}
\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\pi}(\boldsymbol{z})\right) &= \pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\left(1,\boldsymbol{z}^\top\right)\left(I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right)^{-1}\left(1,\boldsymbol{z}^\top\right)^\top \\
&\geq \pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\left\|\left(1,\boldsymbol{z}^\top\right)\right\|_2^2\lambda_{\min}\left(\left(I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right)^{-1}\right) \\
&= \frac{\pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\left\|\left(1,\boldsymbol{z}^\top\right)\right\|_2^2}{\|I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta})\|_{\text{op}}} \\
&\geq \frac{\pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\left\|\left(1,\boldsymbol{z}^\top\right)\right\|_2^2}{\pi_{\text{rare}}(1-\pi_{\text{rare}})\left\|\mathbb{E}\left[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right]\right\|_{\text{op}}} \\
&\overset{(*)}{\geq} \frac{\pi(\boldsymbol{z})^2\left[1-\pi_{\text{rare}}\right]^2}{\pi_{\text{rare}}(1-\pi_{\text{rare}})\lambda_{\max}} \\
&= \frac{\pi(\boldsymbol{z})^2\left[1-\pi_{\text{rare}}\right]}{\pi_{\text{rare}}\lambda_{\max}},
\end{aligned}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of $\cdot$, $(*)$ uses $\left\|\left(1,\boldsymbol{z}^\top\right)\right\|_2^2 \geq 1$ and $\pi(\boldsymbol{z}) \leq \pi_{\text{rare}}$ for all $\boldsymbol{z} \in \mathcal{S}$. This yields

$$\text{Asym.Var}\left(\sqrt{n}\frac{\pi(\boldsymbol{x})-\hat{\pi}_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right) = \frac{1}{\pi(\boldsymbol{x})^2}\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\pi}(\boldsymbol{z})\right) \geq \frac{1-\pi_{\text{rare}}}{\pi_{\text{rare}}\lambda_{\max}},$$

where $\|\cdot\|_2$ denotes the operator norm. Similarly to the previous result, $\hat{\pi}(\boldsymbol{z})$ is asymptotically unbiased and its asymptotic MSE is equal to its asymptotic variance:

$$\begin{aligned}
\text{Asym.MSE}(\hat{\pi}(\boldsymbol{x})) &= \mathbb{E}\left[\lim_{n\to\infty}\left(\sqrt{n}\cdot\frac{\pi(\boldsymbol{x})-\hat{\pi}_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right)^2\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty}\left(\sqrt{n}\cdot\frac{\mathbb{E}[\hat{\pi}_n(\boldsymbol{x})]-\hat{\pi}_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right)^2+\lim_{n\to\infty}\left(\sqrt{n}\cdot\frac{\pi(\boldsymbol{x})-\mathbb{E}[\hat{\pi}_n(\boldsymbol{x})]}{\pi(\boldsymbol{x})}\right)^2\right] \\
&= \text{Asym.Var}\left(\sqrt{n}\frac{\pi(\boldsymbol{x})-\hat{\pi}_n(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right)+0 \\
&\geq \frac{1-\pi_{\text{rare}}}{\pi_{\text{rare}}\lambda_{\max}}.
\end{aligned}$$

□

Proof of Theorem 1. 1. Again, the assumptions of Lemma 5 are satisfied. Lemma 5 shows that the asymptotic covariance matrix of the scaled maximum likelihood estimates of the parameters of logistic regression (a special case of the proportional odds model with $K = 2$ categories) is

$$\text{Asym.Cov}\left(\sqrt{n}\cdot\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right)^\top\right) = \left(I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right)^{-1} = \begin{pmatrix} I_{\alpha\alpha}^{\text{logistic}} & \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \\ I_{\beta\alpha}^{\text{logistic}} & I_{\beta\beta}^{\text{logistic}} \end{pmatrix}^{-1},$$

so in the case that $\boldsymbol{\beta}$ is known, we have

$$\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\alpha}_q\right) = \left(I_{\alpha\alpha}^{\text{logistic}}\right)^{-1} = \frac{1}{I_{\alpha\alpha}^{\text{logistic}}}.$$

If $\boldsymbol{\beta}$ is not known, then if $I_{\beta\beta}^{\text{logistic}}$ is positive definite (and therefore invertible) the formula for block matrix inversion yields

$$\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\alpha}\right) = \frac{1}{I_{\alpha\alpha}^{\text{logistic}} - \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\left(I_{\beta\beta}^{\text{logistic}}\right)^{-1} I_{\beta\alpha}^{\text{logistic}}}. \tag{20}$$

We know that $I_{\beta\beta}^{\text{logistic}}$ is positive definite because $I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is finite and positive definite from Lemma 5, so the principal submatrix $I_{\beta\beta}^{\text{logistic}}$ is positive definite by Observation 7.1.2 in Horn and Johnson [2012]. Further, since we know from Lemma 5 that the covariance matrix of $(\hat{\alpha},\hat{\boldsymbol{\beta}})$ is finite and positive definite under our conditions, this also implies that

$$0 < I_{\alpha\alpha}^{\text{logistic}} - \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\left(I_{\beta\beta}^{\text{logistic}}\right)^{-1} I_{\beta\alpha}^{\text{logistic}} < \infty. \tag{21}$$

Now we seek a lower bound for $\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\alpha}\right)$. We see from (20) that we can get such a bound by lower-bounding $\left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\left(I_{\beta\beta}^{\text{logistic}}\right)^{-1} I_{\beta\alpha}^{\text{logistic}}$. Because $t \mapsto t(1-t)$ is upper-bounded by $1/4$ for all $t \in [0,1]$,

$$I_{\beta\beta}^{\text{logistic}} = \int \boldsymbol{x}\boldsymbol{x}^\top \pi_2(\boldsymbol{x})[1-\pi_2(\boldsymbol{x})]\, dF(\boldsymbol{x}) \preceq \int \boldsymbol{x}\boldsymbol{x}^\top \cdot \frac{1}{4}\, dF(\boldsymbol{x}) = \frac{1}{4}\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right].$$

Then

$$\begin{aligned}
\left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\left(I_{\beta\beta}^{\text{logistic}}\right)^{-1} I_{\beta\alpha}^{\text{logistic}} &\geq \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\left(\frac{1}{4}\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]\right)^{-1} I_{\beta\alpha}^{\text{logistic}} \\
&\geq 4\lambda_{\min}\left(\left(\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]\right)^{-1}\right)\left\|I_{\beta\alpha}^{\text{logistic}}\right\|_2^2 \\
&= \frac{4\left\|I_{\beta\alpha}^{\text{logistic}}\right\|_2^2}{\left\|\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]\right\|_{\text{op}}} \\
&\geq \frac{4\pi_{\min}^2(1-\pi_{\min})^2\left\|\mathbb{E}\left[\boldsymbol{X}\right]\right\|_2^2}{\lambda_{\max}} =: \Delta,
\end{aligned} \tag{22}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of $\cdot$ and the last step follows because

$$\begin{aligned}\left\|I_{\beta\alpha}^{\text{logistic}}\right\|_2 &= \left\|\int \boldsymbol{x}\pi_2(\boldsymbol{x})[1-\pi_2(\boldsymbol{x})]\ dF(\boldsymbol{x})\right\|_2 \\ &\geq \left\|\int \boldsymbol{x}\pi_{\min}(1-\pi_{\min})\ dF(\boldsymbol{x})\right\|_2 \\ &= \pi_{\min}(1-\pi_{\min})\left\|\mathbb{E}\left[\boldsymbol{X}\right]\right\|_2\end{aligned}$$

(where we used the fact that $\boldsymbol{X}$ has support only over nonnegative numbers). Therefore (20) and (22) yield

$$\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\alpha}\right) \geq \left(\frac{1}{\text{Asym.Var}\left(\sqrt{n}\cdot\hat{\alpha}_q\right)} - \Delta\right)^{-1}. \tag{23}$$

The remainder of the argument is similar to the end of the proof of Theorem 4: the asymptotic unbiasedness of these estimators yields

$$\text{Asym.MSE}(\hat{\alpha}) = \text{Asym.Var}\left(\sqrt{n}\cdot[\alpha-\hat{\alpha}]\right)$$

and

$$\text{Asym.MSE}(\hat{\alpha}_q) = \text{Asym.Var}\left(\sqrt{n}\cdot[\alpha-\hat{\alpha}_q]\right).$$

Then from (21) we know that

$$I_{\alpha\alpha}^{\text{logistic}} > \left(I_{\beta\alpha}^{\text{logistic}}\right)^{\top}\left(I_{\beta\beta}^{\text{logistic}}\right)^{-1} I_{\beta\alpha}^{\text{logistic}} \geq \frac{4\pi_{\min}^2(1-\pi_{\min})^2\left\|\mathbb{E}\left[\boldsymbol{X}\right]\right\|_2^2}{\lambda_{\max}}. \tag{24}$$

Making the appropriate substitutions into (24) yields

$$\frac{1}{\text{Asym.MSE}(\hat{\alpha}_q)} - \Delta > 0,$$

and then substituting into (23) yields

$$\begin{aligned}\text{Asym.MSE}(\hat{\alpha}) &\geq \left(\frac{1}{\text{Asym.MSE}(\hat{\alpha}_q)} - \Delta\right)^{-1} \\ &= \frac{\text{Asym.MSE}(\hat{\alpha}_q)}{1-\Delta\cdot\text{Asym.MSE}(\hat{\alpha}_q)} \\ &\overset{(*)}{\geq} \text{Asym.MSE}(\hat{\alpha}_q)\cdot(1+\Delta\cdot\text{Asym.MSE}(\hat{\alpha}_q)) \\ \iff \qquad \frac{\text{Asym.MSE}(\hat{\alpha})-\text{Asym.MSE}(\hat{\alpha}_q)}{[\text{Asym.MSE}(\hat{\alpha}_q)]^2} &\geq \Delta,\end{aligned}$$

where in $(*)$ we used the inequality $c/(1-ct) \leq c(1+ct)$ for any $c>0$, $t<\frac{1}{c}$.

2. Because $(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \mapsto \hat{\pi}(\boldsymbol{z})$ is differentiable for all $\boldsymbol{z} \in \mathbb{R}^p$, by the delta method (Theorem 3.1 in van der Vaart 2000)

$$\sqrt{n} \cdot [\hat{\pi}(\boldsymbol{z}) - \pi(\boldsymbol{z})] \xrightarrow{d} \mathcal{N}\left(0, \pi(\boldsymbol{z})^2 [1-\pi(\boldsymbol{z})]^2 \left(1, \boldsymbol{z}^\top\right) \left(I^{\text{logistic}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)^{-1} \left(1, \boldsymbol{z}^\top\right)^\top\right)$$

for any $\boldsymbol{z} \in \mathbb{R}^p$, and similarly

$$\sqrt{n} \cdot [\hat{\pi}_q(\boldsymbol{z}) - \pi(\boldsymbol{z})] \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi(\boldsymbol{z})^2 [1-\pi(\boldsymbol{z})]^2}{I_{\alpha\alpha}^{\text{logistic}}}\right).$$

We can find $\left(I^{\text{logistic}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)^{-1}$ using the formula for block matrix inversion if

$$\boldsymbol{D} := I_{\beta\beta}^{\text{logistic}} - \frac{I_{\beta\alpha}^{\text{logistic}} \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top}{I_{\alpha\alpha}^{\text{logistic}}}$$

is positive definite (and therefore invertible; note that $\boldsymbol{D}$ is symmetric) and $I_{\alpha\alpha}^{\text{logistic}} > 0$. We have

$$I_{\alpha\alpha}^{\text{logistic}} = \int \pi_2(\boldsymbol{x})[1-\pi_2(\boldsymbol{x})]\, dF(\boldsymbol{x}) \geq \int \pi_{\min}[1-\pi_{\min}]\, dF(\boldsymbol{x}) = \pi_{\min}[1-\pi_{\min}] > 0,$$

and by Theorem 1.12 in Zhang [2005], we then know $\boldsymbol{D}$ is positive definite (and invertible) since $I^{\text{logistic}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is by Lemma 5 and $I_{\alpha\alpha}^{\text{logistic}} > 0$. Let $\lambda_{\max}^{\boldsymbol{D}} := \|\boldsymbol{D}\|_{\text{op}}$ be the largest eigenvalue of $\boldsymbol{D}$; note that $1/\lambda_{\max}^{\boldsymbol{D}}$ is then the smallest eigenvalue of $\boldsymbol{D}^{-1}$. Then for any $\boldsymbol{z} \in \mathbb{R}^p$, we have

$$\begin{aligned}
&\begin{pmatrix} 1 & \boldsymbol{z}^\top \end{pmatrix} \left(I^{\text{logistic}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{z} \end{pmatrix} \\
= \ &\begin{pmatrix} 1 & \boldsymbol{z}^\top \end{pmatrix} \begin{pmatrix} \frac{1}{I_{\alpha\alpha}^{\text{logistic}}} + \frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2} \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \boldsymbol{D}^{-1} I_{\beta\alpha}^{\text{logistic}} & -\frac{1}{I_{\alpha\alpha}^{\text{logistic}}} \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \boldsymbol{D}^{-1} \\ -\frac{1}{I_{\alpha\alpha}^{\text{logistic}}} \boldsymbol{D}^{-1} I_{\beta\alpha}^{\text{logistic}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{z} \end{pmatrix} \\
= \ &\frac{1}{I_{\alpha\alpha}^{\text{logistic}}} + \frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2} \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \boldsymbol{D}^{-1} I_{\beta\alpha}^{\text{logistic}} + \boldsymbol{z}^\top \boldsymbol{D}^{-1} \boldsymbol{z} - 2 \frac{1}{I_{\alpha\alpha}^{\text{logistic}}} I_{\beta\alpha}^{\text{logistic}} \boldsymbol{D}^{-1} \boldsymbol{z} \\
\stackrel{(a)}{=} \ &\frac{1}{I_{\alpha\alpha}^{\text{logistic}}} + \frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2} \left\| \boldsymbol{D}^{-1/2} \left(I_{\beta\alpha}^{\text{logistic}} - I_{\alpha\alpha}^{\text{logistic}} \boldsymbol{z}\right) \right\|_2^2 \\
\stackrel{(b)}{\geq} \ &\frac{1}{I_{\alpha\alpha}^{\text{logistic}}} + \frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2 \lambda_{\max}^{\boldsymbol{D}}} \left\| \left(I_{\beta\alpha}^{\text{logistic}} - I_{\alpha\alpha}^{\text{logistic}} \boldsymbol{z}\right) \right\|_2^2 \qquad (25) \\
\geq \ &\frac{1}{I_{\alpha\alpha}^{\text{logistic}}},
\end{aligned}$$

where (a) follows from

$$\begin{aligned}
&\frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2}\left\|\boldsymbol{D}^{-1/2}\left(I_{\beta\alpha}^{\text{logistic}}-I_{\alpha\alpha}^{\text{logistic}}\boldsymbol{z}\right)\right\|_2^2\\
=\;&\frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2}\left(\left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \boldsymbol{D}^{-1}I_{\beta\alpha}^{\text{logistic}}+\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2\boldsymbol{z}^\top\boldsymbol{D}^{-1}\boldsymbol{z}-2I_{\alpha\alpha}^{\text{logistic}}I_{\beta\alpha}^{\text{logistic}}\boldsymbol{D}^{-1}\boldsymbol{z}\right)
\end{aligned}$$

and (b) uses the fact that $1/\sqrt{\lambda_{\max}^{\boldsymbol{D}}}$ is the smallest eigenvalue of $\boldsymbol{D}^{-1/2}$. If we can show that $\left\|\boldsymbol{D}^{-1/2}\left(I_{\beta\alpha}^{\text{logistic}}-I_{\alpha\alpha}^{\text{logistic}}\boldsymbol{z}\right)\right\|_2\neq 0$, then (25) is enough to establish the strict inequality in the result. Using (13), note that

$$\begin{aligned}
0 =\;& I_{\beta\alpha}^{\text{logistic}}-I_{\alpha\alpha}^{\text{logistic}}\boldsymbol{z}\\
=\;& \int \boldsymbol{x}\pi(\boldsymbol{x})[1-\pi(\boldsymbol{x})]\;dF(\boldsymbol{x})-\boldsymbol{z}\int \pi(\boldsymbol{x})[1-\pi(\boldsymbol{x})]\;dF(\boldsymbol{x})\\
\iff\qquad \boldsymbol{z} =\;& \frac{\int \boldsymbol{x}\pi(\boldsymbol{x})[1-\pi(\boldsymbol{x})]\;dF(\boldsymbol{x})}{\int \pi(\boldsymbol{x})[1-\pi(\boldsymbol{x})]\;dF(\boldsymbol{x})},
\end{aligned}$$

so for all $\boldsymbol{z}\neq\mathbb{E}\left[\boldsymbol{X}\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]/\mathbb{E}\left[\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]$, we have

$$\frac{1}{\left(I_{\alpha\alpha}^{\text{logistic}}\right)^2\lambda_{\max}^{\boldsymbol{D}}}\left\|\left(I_{\beta\alpha}^{\text{logistic}}-I_{\alpha\alpha}^{\text{logistic}}\boldsymbol{z}\right)\right\|_2^2>0.$$

So for any $\boldsymbol{z}\in\mathbb{R}^p\setminus\mathbb{E}\left[\boldsymbol{X}\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]/\mathbb{E}\left[\pi(\boldsymbol{X})[1-\pi(\boldsymbol{X})]\right]$, we have

$$\begin{aligned}
\pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2\begin{pmatrix}1 & \boldsymbol{z}^\top\end{pmatrix}\begin{pmatrix}I_{\alpha\alpha}^{\text{logistic}} & \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top\\ I_{\beta\alpha}^{\text{logistic}} & I_{\beta\beta}^{\text{logistic}}\end{pmatrix}^{-1}\begin{pmatrix}1\\ \boldsymbol{z}\end{pmatrix} &> \frac{\pi(\boldsymbol{z})^2\left[1-\pi(\boldsymbol{z})\right]^2}{I_{\alpha\alpha}^{\text{logistic}}}\\
\iff\qquad \text{Asym.Var}\left(\sqrt{n}\cdot\hat{\pi}_q(\boldsymbol{z})\right) &< \text{Asym.Var}\left(\sqrt{n}\cdot(\hat{\pi}(\boldsymbol{z})\right).
\end{aligned}$$

□

D Remark 3 and Proof of Theorem 2

Before we prove Theorem 2, we begin with a remark investigating the plausibility of the upper bound (5) required as an assumption for Theorem 2.

Remark 3. To investigate the plausibility of the assumption that $I_{\beta\beta}-2\frac{I_{\beta\alpha_1}I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ is positive definite (and that the upper bound for π_{rare} in Equation 5 can hold) empirically, we performed the following simulation study (with a setup similar to that of our simulation studies in Section 4 of the paper). In each of 25 simulations, using $n = 10^6$, $p = 10$, and $K = 3$, we generated $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with $\boldsymbol{X}_{ij} \sim \text{Uniform}(-1, 1)$ iid for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. We then generated $y_i \in \{1, 2, 3\}$ from $\boldsymbol{x}_i$ for each $i \in \{1, \ldots, n\}$ according to the proportional odds model (1), using $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ and $\boldsymbol{\alpha} = (0, 20)$ (so that class 3 would be very rare, with $\pi_{\text{rare}} \approx 4.54 \cdot 10^{-5}$). We estimated $I_{\beta\beta}$, $I_{\beta\alpha_1}$, and $I_{\alpha_1\alpha_1}$ using empirical estimates of the expressions in (8), (9), and (10); for example, using (19) we estimated $I_{\beta\beta}$ by

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-1} \boldsymbol{x}_i \boldsymbol{x}_i^\top \left[\pi_k(\boldsymbol{x}_i) + \pi_{k+1}(\boldsymbol{x}_i) \right] p_k(\boldsymbol{x}_i)[1 - p_k(\boldsymbol{x}_i)].$$

Finally, we used these estimated quantities to estimate $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$, and we calculated the minimum eigenvalue of this estimated matrix. Across all 25 simulations, the sample mean of this minimum eigenvalue was 0.02361, and the minimum was 0.02359. The standard error was $2.94 \cdot 10^{-6}$, and the 95% confidence interval for the mean of the minimum eigenvalue was $(0.02360, 0.02362)$. See Figure 9 for a boxplot of the 100 estimated minimum eigenvalues. These results seem to suggest that the assumption that $\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}} \right) > 0$ is reasonable under the assumptions of Theorem 2.

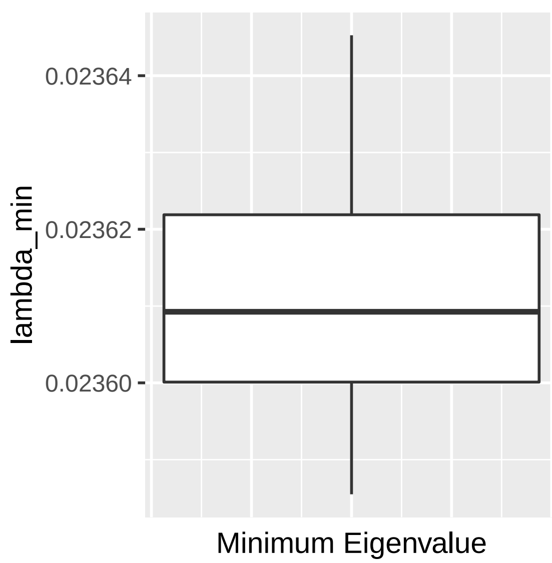

Figure 9: Boxplot of the estimated minimum eigenvalues of $I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}$ *in the simulation study described in Remark 3.*

Further, we also empirically investigated the plausibility of the assumption

$$\pi_{\text{rare}} \leq \frac{\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1}I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right)}{3M^2\left(2+M\right)} \tag{26}$$

directly. Note that in this simulation study, $M = \|(1,\dots,1)\|_2 = \sqrt{p}$, $\mathcal{S} = [-1,1]^p$ and

$$\sup_{\boldsymbol{x}\in\mathcal{S}}\{\pi_3(\boldsymbol{x})\} = \pi_3(\boldsymbol{x}^*) = 1 - \frac{1}{1+\exp\{-(20+\sum_{j=1}^p x_j^*)\}},$$

for $\boldsymbol{x}^* = (-1,\dots,-1)$. So all of the quantities in (26) are known except the minimum eigenvalue, which we were able to estimate with seemingly high precision. It turns out that (26) was satisfied even when we used the minimum value of $\lambda_{\min}$ across all 25 simulations as our estimate; in this case, the left side of (26) is $4.54\cdot 10^{-5}$ and the right side is $1.52\cdot 10^{-4}$.

Lastly, for Theorem 2 to make sense it should also hold that $\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1}I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right)$ does not vanish as π_{rare} becomes arbitrarily small; we investigate this analytically. To see that this seems reasonable, note that from (19) we have

$$\begin{aligned}
I_{\beta\beta} = & \sum_{k=1}^{2}\int \boldsymbol{x}\boldsymbol{x}^\top\left[\pi_k(\boldsymbol{x})+\pi_{k+1}(\boldsymbol{x})\right]p_k(\boldsymbol{x})[1-p_k(\boldsymbol{x})]\ dF(\boldsymbol{x}) \\
= & \int \boldsymbol{x}\boldsymbol{x}^\top\left(\left[\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})\right]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})] + \left[\pi_2(\boldsymbol{x})+\pi_3(\boldsymbol{x})\right]\pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]\right)\ dF(\boldsymbol{x}) \\
= & \int \boldsymbol{x}\boldsymbol{x}^\top\left(\left[1-\pi_3(\boldsymbol{x})\right]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})] + \left[1-\pi_1(\boldsymbol{x})\right]\pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]\right)\ dF(\boldsymbol{x}) \\
= & \int \boldsymbol{x}\boldsymbol{x}^\top[1-\pi_1(\boldsymbol{x})][1-\pi_3(\boldsymbol{x})]\left(\pi_1(\boldsymbol{x})+\pi_3(\boldsymbol{x})\right)\ dF(\boldsymbol{x}) \\
= & \int \boldsymbol{x}\boldsymbol{x}^\top[\pi_2(\boldsymbol{x})+\pi_3(\boldsymbol{x})][1-\pi_3(\boldsymbol{x})][\pi_1(\boldsymbol{x})+\pi_3(\boldsymbol{x})]\ dF(\boldsymbol{x}),
\end{aligned}$$

which is non-vanishing in π_{rare}. (In particular, there seems to be no reason to suspect that the eigenvalues of $I_{\beta\beta}$ change drastically for, say $\pi_3(\boldsymbol{x}) \leq 4.54\cdot 10^{-5}$ for all $\boldsymbol{x}\in\mathcal{S}$, as in our simulation study above, versus $\pi_3(\boldsymbol{x}) \leq 10^{-20}$ for all $\boldsymbol{x}\in\mathcal{S}$.) Further, (32) below shows that

$$\left\|\frac{I_{\beta\alpha_1}I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right\|_{\text{op}} = \frac{\|I_{\beta\alpha_1}\|_2^2}{I_{\alpha_1\alpha_1}}$$

(where $\|\cdot\|_{\text{op}}$ is the operator norm) is bounded from above by a constant not depending on π_{rare}. Taken together, this suggests that Assumption (26) does not become more implausible as π_{rare} becomes arbitrarily small.

Before we proceed with the proof of Theorem 2, we state a lemma with inequalities we will use.

Lemma 6. The following inequalities hold under the assumptions of Theorem 2:

$$I_{\alpha_2\alpha_2} \le \pi_{\text{rare}} I_{\alpha_1\alpha_1} \cdot \frac{1}{1/4 - \Delta^2}, \tag{27}$$

$$I_{\alpha_2\alpha_2} \le \pi_{\text{rare}} \left(1 + \frac{\pi_{\text{rare}}}{1/2 - \Delta}\right), \tag{28}$$

$$|I_{\alpha_1\alpha_2}| \le I_{\alpha_2\alpha_2}, \tag{29}$$

$$\|I_{\beta\alpha_2}\|_2 \le M \cdot I_{\alpha_2\alpha_2}, \tag{30}$$

$$\frac{\|I_{\beta\alpha_2}\|_2^2}{I_{\alpha_2\alpha_2}} \le M^2 \pi_{\text{rare}} \left(1 + \frac{\pi_{\text{rare}}}{1/2 - \Delta}\right), \qquad \text{and} \tag{31}$$

$$\frac{\|I_{\beta\alpha_1}\|_2^2}{I_{\alpha_1\alpha_1}} \le \frac{M^2}{4}. \tag{32}$$

Proof. Provided immediately after the proof of Theorem 2. □

Proof of Theorem 2. By an argument analogous to the one used at the end of the proof of Theorem 4, it is enough to upper-bound

$$\left\|\text{Cov}\left(\lim_{n\to\infty} \sqrt{n}\left[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{prop. odds}}\right]\right)\right\|_{\text{op}} = \left\|\text{Asym.Cov}\left(\sqrt{n} \cdot \hat{\boldsymbol{\beta}}^{\text{prop. odds}}\right)\right\|_{\text{op}}.$$

Using Lemma 5 and the block matrix inversion formula,

$$\begin{aligned} \text{Asym.Cov}\left(\sqrt{n} \cdot \hat{\boldsymbol{\beta}}^{\text{prop. odds}}\right) &= \left(I_{\beta\beta} - I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta}\right)^{-1} \\ \implies \quad \left\|\text{Asym.Cov}\left(\sqrt{n} \cdot \hat{\boldsymbol{\beta}}^{\text{prop. odds}}\right)\right\|_{\text{op}} &= \frac{1}{\lambda_{\min}\left(I_{\beta\beta} - I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta}\right)}, \end{aligned}$$

where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of $\cdot$ and $\|\cdot\|_{\text{op}}$ is the operator norm. $I_{\alpha\alpha}$ is a 2×2 matrix, so

$$I_{\alpha\alpha}^{-1} = \frac{1}{\det(I_{\alpha\alpha})} \begin{pmatrix} I_{\alpha_2\alpha_2} & -I_{\alpha_1\alpha_2} \\ -I_{\alpha_2\alpha_1} & I_{\alpha_1\alpha_1} \end{pmatrix},$$

and

$$\begin{aligned} I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta} &= \frac{1}{\det(I_{\alpha\alpha})} \begin{pmatrix} I_{\beta\alpha_1} & I_{\beta\alpha_2} \end{pmatrix} \begin{pmatrix} I_{\alpha_2\alpha_2} & -I_{\alpha_1\alpha_2} \\ -I_{\alpha_2\alpha_1} & I_{\alpha_1\alpha_1} \end{pmatrix} \begin{pmatrix} I_{\alpha_1\beta} \\ I_{\alpha_2\beta} \end{pmatrix} \\ &= \frac{1}{\det(I_{\alpha\alpha})} \left(I_{\alpha_2\alpha_2} I_{\beta\alpha_1} I_{\beta\alpha_1}^\top + I_{\alpha_1\alpha_1} I_{\beta\alpha_2} I_{\beta\alpha_2}^\top - I_{\alpha_1\alpha_2} I_{\beta\alpha_1} I_{\beta\alpha_2}^\top - I_{\alpha_1\alpha_2} I_{\beta\alpha_2} I_{\beta\alpha_1}^\top\right) \\ &= \frac{1}{\det(I_{\alpha\alpha})} \left(I_{\alpha_2\alpha_2} I_{\beta\alpha_1} I_{\beta\alpha_1}^\top + I_{\alpha_1\alpha_1} I_{\beta\alpha_2} I_{\beta\alpha_2}^\top + |I_{\alpha_1\alpha_2}| I_{\beta\alpha_1} I_{\beta\alpha_2}^\top + |I_{\alpha_1\alpha_2}| I_{\beta\alpha_2} I_{\beta\alpha_1}^\top\right), \end{aligned} \tag{33}$$

where in the last step we used that $I_{\alpha_1\alpha_2} = -\tilde{M}_2 < 0$, which is clear from (9) and (12). Because we know from Lemma 5 that $I_{\alpha\alpha}$ (and therefore also $I_{\alpha\alpha}^{-1}$) is positive definite, by Observation 7.1.6 in Horn and Johnson [2012] $I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta}$ is positive definite as well. Therefore we can use an upper bound on $1/\det(I_{\alpha\alpha})$ to upper bound $I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta}$. We have

$$\begin{aligned}
\det(I_{\alpha\alpha}) &= I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2} - I_{\alpha_1\alpha_2}^2 \\
&\overset{(a)}{\geq} I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2} - I_{\alpha_2\alpha_2}^2 \\
&= I_{\alpha_2\alpha_2}\left(I_{\alpha_1\alpha_1} - I_{\alpha_2\alpha_2}\right) \\
&\overset{(b)}{\geq} I_{\alpha_2\alpha_2}\left(I_{\alpha_1\alpha_1} - \pi_{\text{rare}} I_{\alpha_1\alpha_1} \cdot \frac{1}{1/4 - \Delta^2}\right) \\
&= \left(1 - \frac{\pi_{\text{rare}}}{1/4 - \Delta^2}\right) I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2} \\
&\overset{(c)}{\geq} \frac{1}{2} I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2},
\end{aligned}$$

where in (a) we used (29), in (b) we used (27), and (c) uses that from the upper bound (5) for π_{rare} we have

$$\begin{aligned}
\pi_{\text{rare}} &\leq \frac{1}{2}\left(\frac{1}{2} - \Delta\right)\left(\frac{1}{2} + \Delta\right) = \frac{1}{2}\left(\frac{1}{4} - \Delta^2\right) \\
\iff \quad \frac{\pi_{\text{rare}}}{1/4 - \Delta^2} &\leq \frac{1}{2} \\
\iff \quad 1 - \frac{\pi_{\text{rare}}}{1/4 - \Delta^2} &\geq \frac{1}{2}.
\end{aligned}$$

Now we can bound $I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta}$ using (33):

$$\begin{aligned}
I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta} &\preceq \frac{2}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(I_{\alpha_2\alpha_2} I_{\beta\alpha_1} I_{\beta\alpha_1}^\top + I_{\alpha_1\alpha_1} I_{\beta\alpha_2} I_{\beta\alpha_2}^\top + |I_{\alpha_1\alpha_2}| I_{\beta\alpha_1} I_{\beta\alpha_2}^\top + |I_{\alpha_1\alpha_2}| I_{\beta\alpha_2} I_{\beta\alpha_1}^\top\right) \\
&= 2\left(\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}} + \frac{I_{\beta\alpha_2} I_{\beta\alpha_2}^\top}{I_{\alpha_2\alpha_2}} + \frac{|I_{\alpha_1\alpha_2}|}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(I_{\beta\alpha_1} I_{\beta\alpha_2}^\top + I_{\beta\alpha_2} I_{\beta\alpha_1}^\top\right)\right),
\end{aligned}$$

and

$$I_{\beta\beta} - I_{\alpha\beta}^\top I_{\alpha\alpha}^{-1} I_{\alpha\beta} \succeq I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}} - 2\left(\frac{I_{\beta\alpha_2} I_{\beta\alpha_2}^\top}{I_{\alpha_2\alpha_2}} + \frac{|I_{\alpha_1\alpha_2}|}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(I_{\beta\alpha_1} I_{\beta\alpha_2}^\top + I_{\beta\alpha_2} I_{\beta\alpha_1}^\top\right)\right)$$

Note that $I_{\beta\beta}$, $I_{\beta\alpha_1} I_{\beta\alpha_1}^\top$, $I_{\beta\alpha_2} I_{\beta\alpha_2}^\top$, and $I_{\beta\alpha_1} I_{\beta\alpha_2}^\top + I_{\beta\alpha_2} I_{\beta\alpha_1}^\top$ are all symmetric. By Weyl's theorem, it holds that for symmetric matrices with matching dimensions $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\lambda_{\min}(\boldsymbol{A} - \boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) - \|\boldsymbol{B}\|_{\text{op}},$$

because for any $\boldsymbol{v}$

$$(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} - \boldsymbol{B}\boldsymbol{v} \geq \lambda_{\min}(\boldsymbol{A})\boldsymbol{v} - \|\boldsymbol{B}\|_{\text{op}}\boldsymbol{v} = \left(\lambda_{\min}(\boldsymbol{A}) - \|\boldsymbol{B}\|_{\text{op}}\right)\boldsymbol{v}.$$

So

$$\begin{aligned}
&\lambda_{\min}\left(I_{\beta\beta} - I_{\alpha\beta}^{\top} I_{\alpha\alpha}^{-1} I_{\alpha\beta}\right) \\
\geq\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - \frac{2}{I_{\alpha_2\alpha_2}}\left\|I_{\beta\alpha_2} I_{\beta\alpha_2}^{\top}\right\|_{\text{op}} - 2\frac{|I_{\alpha_1\alpha_2}|}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left\|I_{\beta\alpha_1} I_{\beta\alpha_2}^{\top} + I_{\beta\alpha_2} I_{\beta\alpha_1}^{\top}\right\|_{\text{op}} \\
\overset{(a)}{\geq}\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - \frac{2}{I_{\alpha_2\alpha_2}}\left\|I_{\beta\alpha_2} I_{\beta\alpha_2}^{\top}\right\|_{\text{op}} - 2\frac{|I_{\alpha_1\alpha_2}|}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(\left\|I_{\beta\alpha_1} I_{\beta\alpha_2}^{\top}\right\|_{\text{op}} + \left\|I_{\beta\alpha_2} I_{\beta\alpha_1}^{\top}\right\|_{\text{op}}\right) \\
\overset{(b)}{\geq}\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2\frac{\|I_{\beta\alpha_2}\|_2^2}{I_{\alpha_2\alpha_2}} - 2\frac{|I_{\alpha_1\alpha_2}|}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(2\|I_{\beta\alpha_1}\|_2\|I_{\beta\alpha_2}\|_2\right) \\
\overset{(c)}{\geq}\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2M^2\pi_{\text{rare}}\left(1 + \frac{\pi_{\text{rare}}}{1/2-\Delta}\right) - 2\frac{I_{\alpha_2\alpha_2}}{I_{\alpha_1\alpha_1} I_{\alpha_2\alpha_2}}\left(2\|I_{\beta\alpha_1}\|_2 \cdot M \cdot I_{\alpha_2\alpha_2}\right) \\
=\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2M\left(M\pi_{\text{rare}}\left(1 + \frac{\pi_{\text{rare}}}{1/2-\Delta}\right) + 2\frac{I_{\alpha_2\alpha_2}}{I_{\alpha_1\alpha_1}}\|I_{\beta\alpha_1}\|_2\right) \\
\overset{(d)}{\geq}\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2M\left(M\pi_{\text{rare}}\left(1 + \frac{\pi_{\text{rare}}}{1/2-\Delta}\right) + 2\frac{1}{I_{\alpha_1\alpha_1}}\|I_{\beta\alpha_1}\|_2 \cdot \pi_{\text{rare}}\left(1 + \frac{\pi_{\text{rare}}}{1/2-\Delta}\right)\right) \\
=\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2M\pi_{\text{rare}}\left(1 + \frac{\pi_{\text{rare}}}{1/2-\Delta}\right)\left(M + 2\frac{\|I_{\beta\alpha_1}\|_2}{I_{\alpha_1\alpha_1}}\right) \\
\overset{(e)}{\geq}\ & \lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) - 2M\pi_{\text{rare}}\left(1 + \frac{1}{2}\right)\left(M + \frac{M^2}{2}\right) \\
\overset{(f)}{\geq}\ & \frac{1}{2}\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right),
\end{aligned}$$

where in (a) we used the triangle inequality, in (b) we used that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ it holds that $\|\boldsymbol{a}\boldsymbol{b}^{\top}\|_{\text{op}} = |\boldsymbol{b}^{\top}\boldsymbol{a}| \leq \|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$ (note that $\boldsymbol{a}\boldsymbol{b}^{\top}$ is rank one with eigenvector $\boldsymbol{a}$) as well as the triangle inequality, in (c) we used (29), (30), and (31), in (d) we used (28), (e) follows from (32) and

$$\frac{\pi_{\text{rare}}}{1/2-\Delta} \leq \frac{1/2(1/2-\Delta)(1/2+\Delta)}{1/2-\Delta} = \frac{1}{2}\left(\frac{1}{2}+\Delta\right) \leq \frac{1}{2}\left(\frac{1}{2}+\frac{1}{2}\right) \leq \frac{1}{2}$$

(since $\Delta \leq 1/2$), and in (f) we used our assumptions that

$$\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^{\top}}{I_{\alpha_1\alpha_1}}\right) > 0$$

and

$$\pi_{\text{rare}} \leq \frac{\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right)}{3M^2(2+M)}$$

$$\iff \qquad 3M\pi_{\text{rare}}\left(M + \frac{M^2}{2}\right) \leq \frac{1}{2}\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right).$$

We have therefore shown that

$$\left\|\text{Cov}\left(\lim_{n\to\infty}\sqrt{n}\left[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{prop. odds}}\right]\right)\right\|_{\text{op}} \leq \frac{2}{\lambda_{\min}\left(I_{\beta\beta} - 2\frac{I_{\beta\alpha_1} I_{\beta\alpha_1}^\top}{I_{\alpha_1\alpha_1}}\right)}.$$

□

Proof of Lemma 6. We omit $\mathbf{x}$ in integrals, when it is clear (e.g., π_1 stands for $\pi_1(\mathbf{x})$).
Proof of (27)**:**

From (9) and (11), using $\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x}) + \pi_3(\boldsymbol{x}) = 1$ we have

$$\begin{aligned}
I_{\alpha_2\alpha_2} &= M_2 \\
&= \int [\pi_3(\boldsymbol{x})(1-\pi_3(\boldsymbol{x}))]^2 \left(\frac{1}{\pi_2(\boldsymbol{x})} + \frac{1}{\pi_3(\boldsymbol{x})}\right)\, dF(\boldsymbol{x}) \\
&= \int [\pi_3(\boldsymbol{x})(1-\pi_3(\boldsymbol{x}))]^2 \left(\frac{\pi_2(\boldsymbol{x}) + \pi_3(\boldsymbol{x})}{\pi_2(\boldsymbol{x})\pi_3(\boldsymbol{x})}\right)\, dF(\boldsymbol{x}) \\
&= \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]^2 \left(\frac{1-\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\, dF(\boldsymbol{x}) \qquad (34) \\
&\leq \pi_{\text{rare}} \int (1-\pi_3)^2(1-\pi_1)/\pi_2 dF \\
&= \pi_{\text{rare}} \int \pi_1(1-\pi_1)^2 \frac{(1-\pi_3)^2}{\pi_2}\left[\frac{1}{\pi_1(1-\pi_1)}\right] dF \\
&\overset{(a)}{\leq} \pi_{\text{rare}} \int \pi_1(1-\pi_1)^2 \frac{1-\pi_3}{\pi_2}\left[\frac{1}{(1/2-\Delta)(1/2+\Delta)}\right] dF \\
&= \pi_{\text{rare}} \int \pi_1^2(1-\pi_1)^2 \frac{\pi_1+\pi_2}{\pi_1\pi_2}\frac{1}{1/4-\Delta^2} \\
&= \pi_{\text{rare}} \int \pi_1^2(1-\pi_1)^2 \left(\frac{1}{\pi_1} + \frac{1}{\pi_2}\right)\frac{1}{1/4-\Delta^2} \\
&= \pi_{\text{rare}} M_1 \frac{1}{1/4-\Delta^2} \\
&= \pi_{\text{rare}} I_{\alpha_1\alpha_1} \frac{1}{1/4-\Delta^2},
\end{aligned}$$

where in (a) we used the fact that $t \mapsto 1/[t(1-t)]$ is nonincreasing in t for $t \in (0, 1/2]$, so it is largest when t is small as possible, and by assumption $\inf_{\boldsymbol{x} \in \mathcal{S}}\{\pi_1 \wedge 1 - \pi_1\} \geq 1/2 - \Delta$.

Proof of (28)**:** Using (34) we have

$$\begin{aligned} I_{\alpha_2\alpha_2} &= \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]^2 \left(\frac{\pi_2(\boldsymbol{x})+\pi_3(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right) dF(\boldsymbol{x}) \\ &= \int \pi_3(1-\pi_3)^2(1+\pi_3/\pi_2)dF \\ &\leq \pi_{\text{rare}} \int (1-0)^2 \left(1+\frac{\pi_{\text{rare}}}{1/2-\Delta}\right) dF. \end{aligned}$$

Proof of (29)**:** Using (34) along with (8) and (12) we have

$$\begin{aligned} I_{\alpha_2\alpha_2} &= \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]^2 \left(\frac{1-\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right) dF(\boldsymbol{x}) \\ &= \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] \left(\frac{\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right) dF(\boldsymbol{x}) \\ &= \int \pi_3(1-\pi_3)(1-\pi_1)\frac{\pi_1}{\pi_2} dF + \int \pi_3(1-\pi_3)(1-\pi_1)dF \\ &= \tilde{M}_2 + \int \pi_3(1-\pi_3)(1-\pi_1)dF \\ &= |I_{\alpha_1\alpha_2}| + \int \pi_3(1-\pi_3)(1-\pi_1)dF \\ &\geq |I_{\alpha_1\alpha_2}|. \end{aligned}$$

Proof of (30)**:** From (9) and (18) we have

$$
\begin{aligned}
\|I_{\beta\alpha_2}\|_2 &= \left\| \int \boldsymbol{x}\,[\pi_2(\boldsymbol{x}) + \pi_3(\boldsymbol{x})]\,\pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]\ dF(\boldsymbol{x}) \right\|_2 \\
&= \left\| \int \boldsymbol{x}\,[1-\pi_1(\boldsymbol{x})]\,\pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]\ dF(\boldsymbol{x}) \right\|_2 \\
&\le \int \|\boldsymbol{x}\|_2 \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] dF \\
&\le M \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] dF \\
&\le M \cdot \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] \left(1 + \frac{\pi_3(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\ dF(\boldsymbol{x}) \\
&= M \cdot \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] \left(\frac{\pi_2(\boldsymbol{x}) + \pi_3(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\ dF(\boldsymbol{x}) \\
&= M \cdot \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})][1-\pi_1(\boldsymbol{x})] \left(\frac{1-\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\ dF(\boldsymbol{x}) \\
&\le M \cdot \int \pi_3(\boldsymbol{x})[1-\pi_3(\boldsymbol{x})]^2 \left(\frac{1-\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\ dF(\boldsymbol{x}) \\
&= M \cdot I_{\alpha_2\alpha_2},
\end{aligned}
$$

where in the last inequality we used $\pi_3(\boldsymbol{x}) \le \pi_{\text{rare}} < 1/2 - \Delta \le \pi_1(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{S}$ and in the last step we used (34).

Proof of (31)**:** This follows from (28) and (30).

Proof of (32)**:** Using from (9)

$$
\begin{aligned}
\|I_{\beta\alpha_1}\|_2 = &\ \left\| \int \boldsymbol{x}\pi_1^2(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) + \int \boldsymbol{x}\pi_2(\boldsymbol{x})\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) \right\|_2 \\
= &\ \left\| \int \boldsymbol{x}[\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x})]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) \right\|_2 \\
\le &\ \int \|\boldsymbol{x}\|_2\,[\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x})]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x}) \\
\le &\ M \int [\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x})]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})]\ dF(\boldsymbol{x})
\end{aligned}
$$

and from (8)

$$\begin{aligned}
I_{\alpha_1\alpha_1} = M_1 = & \int (\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})])^2 \left(\frac{1}{\pi_1(\boldsymbol{x})} + \frac{1}{\pi_2(\boldsymbol{x})}\right) \, dF(\boldsymbol{x}) \\
= & \int (\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})])^2 \frac{\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})}{\pi_1(\boldsymbol{x})\pi_2(\boldsymbol{x})} \, dF(\boldsymbol{x}) \\
= & \int \pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})][\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})] \frac{1-\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})} \, dF(\boldsymbol{x}) \\
\geq & \int \pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})][\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})] \, dF(\boldsymbol{x}),
\end{aligned}$$

we have

$$\frac{\|I_{\beta\alpha_1}\|_2^2}{I_{\alpha_1\alpha_1}} \leq M^2 \int [\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})]\pi_1(\boldsymbol{x})[1-\pi_1(\boldsymbol{x})] \, dF(\boldsymbol{x}) \leq \frac{M^2}{4},$$

where in the last step we used that $t \mapsto t(1-t) \leq 1/4$ for all $t \in [0,1]$. □

E Proof of Lemma 5

First we will calculate the Fisher information matrices, then we will show the convergence results.

Since the logistic regression model can be considered a special case of the proportional odds model with $K = 2$ categories, we will mostly focus our calculations on the proportional odds model.

E.1 Calculating the Log Likelihood and Gradients

In the proportional odds model, the likelihood can be expressed as

$$\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{1}{1+\exp\left\{-(\alpha_k+\boldsymbol{\beta}^\top\boldsymbol{x}_i)\right\}} - \frac{1}{1+\exp\left\{-(\alpha_{k-1}+\boldsymbol{\beta}^\top\boldsymbol{x}_i)\right\}}\right)^{\mathbb{1}\{y_i=k\}}, \tag{35}$$

so the log likelihood can be expressed as

$$\begin{aligned}
\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i=k\}\log\left(\frac{1}{1+\exp\left\{-(\alpha_k+\boldsymbol{\beta}^\top\boldsymbol{x}_i)\right\}} - \frac{1}{1+\exp\left\{-(\alpha_{k-1}+\boldsymbol{\beta}^\top\boldsymbol{x}_i)\right\}}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i=k\}\log\left(p_{ik}-p_{i,k-1}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i=k\}\log\left(\pi_{ik}\right),
\end{aligned}$$

where

$$\begin{aligned}
p_{ik} &:= p_k\left(\boldsymbol{x}_i\right), \\
\pi_{ik} &:= \mathbb{P}(y_i = k) = p_{ik} - p_{i,k-1} = p_k\left(\boldsymbol{x}_i\right) - p_{k-1}\left(\boldsymbol{x}_i\right),
\end{aligned}$$

$\alpha_0 := -\infty$ and $\alpha_K := \infty$ (while $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{K-1})^\top$ are parameters to be estimated). Using

$$\frac{\partial}{\partial t}\frac{1}{1+\exp\{-t\}} = \frac{1}{1+\exp\{-t\}}\left(1 - \frac{1}{1+\exp\{-t\}}\right)$$

the gradient has entries corresponding to $\boldsymbol{\beta}$ equal to

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i = k\}\boldsymbol{x}_i\left(\frac{p_{ik}(1-p_{ik}) - p_{i,k-1}(1-p_{i,k-1})}{p_{ik}-p_{i,k-1}}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i = k\}\boldsymbol{x}_i\left(\frac{p_{ik} - p_{i,k-1} - \left(p_{ik}^2 - p_{i,k-1}^2\right)}{p_{ik}-p_{i,k-1}}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i = k\}\boldsymbol{x}_i\left(1 - p_{ik} - p_{i,k-1}\right),
\end{aligned}$$

and using

$$\begin{aligned}
\frac{\partial}{\partial\alpha_k}\pi_{ik} &= \frac{\partial}{\partial\alpha_k}\left(p_{ik} - p_{i,k-1}\right) = p_{ik}(1-p_{ik}) \qquad \text{and} \\
\frac{\partial}{\partial\alpha_k}\pi_{i,k+1} &= \frac{\partial}{\partial\alpha_k}\left(p_{i,k+1} - p_{ik}\right) = -p_{ik}(1-p_{ik}),
\end{aligned}$$

the entries corresponding to $\boldsymbol{\alpha}$ equal

$$\begin{aligned}
\frac{\partial}{\partial\alpha_k}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\frac{\partial}{\partial\alpha_k}\left(\mathbb{1}\{y_i = k\}\log\left(\pi_{ik}\right) + \mathbb{1}\{y_i = k+1\}\log\left(\pi_{i,k+1}\right)\right) \\
&= \sum_{i=1}^{n}\left(\mathbb{1}\{y_i = k\}\frac{p_{ik}(1-p_{ik})}{\pi_{ik}} - \mathbb{1}\{y_i = k+1\}\frac{p_{ik}(1-p_{ik})}{\pi_{i,k+1}}\right) \\
\implies \qquad \nabla_{\boldsymbol{\alpha}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\boldsymbol{e}_k p_{ik}(1-p_{ik})\left(\frac{\mathbb{1}\{y_i = k\}}{\pi_{ik}} - \frac{\mathbb{1}\{y_i = k+1\}}{\pi_{i,k+1}}\right),
\end{aligned}$$

where $\boldsymbol{e}_k \in \{0,1\}^{K-1}$ has the k^{th} entry equal to 1 and the rest equal to 0. (Note that since $\alpha_0 = -\infty$, $p_{i0} = 0$, and similarly $p_{iK} = 1$ as expected.)

E.2 Calculating the Hessian Matrices

The entries of the Hessian corresponding to $\boldsymbol{\beta}$, $H^{\text{prop. odds}}_{\beta\beta}$ are

$$\begin{aligned}\nabla^2_{\boldsymbol{\beta}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= -\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}\{y_i=k\}\boldsymbol{x}_i\boldsymbol{x}_i^\top\left[p_{ik}(1-p_{ik})+p_{i,k-1}\left(1-p_{i,k-1}\right)\right] \\ &= -\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\sum_{k=1}^{K}\left(\mathbb{1}\{y_i=k+1\}+\mathbb{1}\{y_i=k\}\right)\mathrm{Var}\left(\mathbb{1}\{y_i\le k\}\right).\end{aligned}$$

Using

$$\frac{\partial}{\partial\alpha_k}p_{ik}(1-p_{ik}) = p_{ik}(1-p_{ik}) - 2p_{ik}^2(1-p_{ik}) = p_{ik}(1-p_{ik})\left(1-2p_{ik}\right),$$

the entries corresponding to the $\boldsymbol{\alpha}$ block of the Hessian $H^{\text{prop. odds}}_{\alpha\alpha}$ are as follows:

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_k^2}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\frac{\partial}{\partial\alpha_k}\left(p_{ik}(1-p_{ik})\left(\frac{\mathbb{1}\{y_i=k\}}{\pi_{ik}}-\frac{\mathbb{1}\{y_i=k+1\}}{\pi_{i,k+1}}\right)\right) \\ &= \sum_{i=1}^{n}\left(\mathbb{1}\{y_i=k\}\frac{\pi_{ik}p_{ik}(1-p_{ik})\left(1-2p_{ik}\right)-p_{ik}^2(1-p_{ik})^2}{\pi_{ik}^2}\right. \\ &\qquad \left.-\,\mathbb{1}\{y_i=k+1\}\frac{\pi_{i,k+1}p_{ik}(1-p_{ik})\left(1-2p_{ik}\right)+p_{ik}^2(1-p_{ik})^2}{\pi_{i,k+1}^2}\right) \\ &= \sum_{i=1}^{n}p_{ik}(1-p_{ik})\left(\mathbb{1}\{y_i=k\}\frac{\pi_{ik}\left(1-2p_{ik}\right)-p_{ik}(1-p_{ik})}{\pi_{ik}^2}\right. \\ &\qquad \left.-\,\mathbb{1}\{y_i=k+1\}\frac{\pi_{i,k+1}\left(1-2p_{ik}\right)+p_{ik}(1-p_{ik})}{\pi_{i,k+1}^2}\right), \\ \frac{\partial^2}{\partial\alpha_k\partial\alpha_{k-1}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}p_{ik}(1-p_{ik})\frac{\partial}{\partial\alpha_{k-1}}\left(\frac{\mathbb{1}\{y_i=k\}}{\pi_{ik}}-\frac{\mathbb{1}\{y_i=k+1\}}{\pi_{i,k+1}}\right) \\ &= -\sum_{i=1}^{n}\mathbb{1}\{y_i=k\}p_{ik}(1-p_{ik})\left(\frac{-p_{i,k-1}(1-p_{i,k-1})}{\pi_{ik}^2}\right) \\ &= \sum_{i=1}^{n}\mathbb{1}\{y_i=k\}\cdot\frac{p_{ik}(1-p_{ik})p_{i,k-1}(1-p_{i,k-1})}{\pi_{ik}^2}, \qquad \text{and} \\ \frac{\partial^2}{\partial\alpha_k\partial\alpha_{k'}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= 0, \qquad \text{all other } k\neq k',\end{aligned}$$

where $\pi_{i,K+1}=0$. (Note that $\frac{\partial^2}{\partial\alpha_k\partial\alpha_{k+1}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is nonzero as well, but it matches the expression for $\frac{\partial^2}{\partial\alpha_{\tilde{k}}\partial\alpha_{\tilde{k}-1}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ with $\tilde{k}:=k+1$.)

Finally, the entries corresponding to the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ mixed blocks $H_{\alpha\beta}^{\text{prop. odds}}$ of the Hessian are

$$\begin{aligned}\frac{\partial}{\partial \alpha_k}\nabla_{\boldsymbol{\beta}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = & \sum_{i=1}^{n}\Bigg(\mathbb{1}\{y_i=k\}\boldsymbol{x}_i\frac{\partial}{\partial \alpha_k}\left(1-p_{ik}-p_{i,k-1}\right) \\ & + \mathbb{1}\{y_i=k+1\}\boldsymbol{x}_i\frac{\partial}{\partial \alpha_k}\left(1-p_{i,k+1}-p_{i,k}\right)\Bigg) \\ = & -\sum_{i=1}^{n}\boldsymbol{x}_i p_{ik}\left(1-p_{ik}\right)\left(\mathbb{1}\{y_i=k\}+\mathbb{1}\{y_i=k+1\}\right),\ k\in[K-1].\end{aligned}$$

E.3 Calculation of the Fisher Information Matrices

Now we find the Fisher information matrices

$$I^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{pmatrix} I_{\alpha\alpha}^{\text{prop. odds}} & I_{\beta\alpha}^{\text{prop. odds}} \\ \left(I_{\beta\alpha}^{\text{prop. odds}}\right)^\top & I_{\beta\beta}^{\text{prop. odds}}\end{pmatrix}$$

and

$$I^{\text{logistic}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{pmatrix} I_{\alpha\alpha}^{\text{logistic}} & \left(I_{\beta\alpha}^{\text{logistic}}\right)^\top \\ I_{\beta\alpha}^{\text{logistic}} & I_{\beta\beta}^{\text{logistic}}\end{pmatrix}$$

by taking the negative expectation of each block (using a single observation). For the α block, we have

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2}{\partial \alpha_k^2}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right] &= -\mathbb{E}\Bigg[p_{ik}(1-p_{ik})\Bigg(\mathbb{1}\{y_i=k\}\frac{\pi_{ik}\,(1-2p_{ik})-p_{ik}(1-p_{ik})}{\pi_{ik}^2} \\
&\quad -\mathbb{1}\{y_i=k+1\}\frac{\pi_{i,k+1}\,(1-2p_{ik})+p_{ik}(1-p_{ik})}{\pi_{i,k+1}^2}\Bigg)\Bigg] \\
&= -\mathbb{E}\Bigg[p_{ik}(1-p_{ik})\Bigg(\frac{\pi_{ik}\,(1-2p_{ik})-p_{ik}(1-p_{ik})}{\pi_{ik}} \\
&\quad -\frac{\pi_{i,k+1}\,(1-2p_{ik})+p_{ik}(1-p_{ik})}{\pi_{i,k+1}}\Bigg)\Bigg] \\
&= \mathbb{E}\left[p_{ik}^2(1-p_{ik})^2\left(\frac{1}{\pi_{ik}}+\frac{1}{\pi_{i,k+1}}\right)\right] \\
&= M_k, \qquad k\in[K-1], \\
-\mathbb{E}\left[\frac{\partial^2}{\partial \alpha_k\partial\alpha_{k-1}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right] &= -\mathbb{E}\left[\mathbb{1}\{y_i=k\}\cdot\frac{p_{ik}(1-p_{ik})p_{i,k-1}(1-p_{i,k-1})}{\pi_{ik}^2}\right] \\
&= -\mathbb{E}\left[\frac{p_{ik}(1-p_{ik})p_{i,k-1}(1-p_{i,k-1})}{\pi_{ik}}\right] \\
&= -\tilde{M}_k, \qquad k\in\{2,\ldots,K-1\}, \qquad \text{and} \\
\mathbb{E}\left[\frac{\partial^2}{\partial \alpha_k\partial\alpha_{k'}}\mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right] &= 0, \qquad \text{all other } k\neq k',
\end{aligned}
$$

where we used the definitions of M_k and $\tilde{M}_k$ in (11) and (12). Therefore $I_{\alpha\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathbb{R}^{(K-1)\times(K-1)}$ has the form

$$
I_{\alpha\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{pmatrix}
M_1 & -\tilde{M}_2 & 0 & \cdots & 0 & 0 \\
-\tilde{M}_2 & M_2 & -\tilde{M}_3 & \cdots & 0 & 0 \\
0 & -\tilde{M}_3 & M_3 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & M_{K-2} & -\tilde{M}_{K-1} \\
0 & 0 & 0 & \cdots & -\tilde{M}_{K-1} & M_{K-1}
\end{pmatrix},
$$

verifying (8). Observe that in the case of logistic regression ($K = 2$), we have

$$
\begin{aligned}
I_{\alpha\alpha}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = M_1 &= \int [p_1(\boldsymbol{x})(1 - p_1(\boldsymbol{x}))]^2 \left(\frac{1}{\pi_1(\boldsymbol{x})} + \frac{1}{\pi_2(\boldsymbol{x})} \right) \, dF(\boldsymbol{x}) \\
&= \int [\pi_1(\boldsymbol{x})(1 - \pi_1(\boldsymbol{x}))]^2 \left(\frac{1}{\pi_1(\boldsymbol{x})} + \frac{1}{1 - \pi_1(\boldsymbol{x})} \right) \, dF(\boldsymbol{x}) \\
&= \int [\pi_1(\boldsymbol{x})(1 - \pi_1(\boldsymbol{x}))]^2 \cdot \frac{1}{\pi_1(\boldsymbol{x}) \, (1 - \pi_1(\boldsymbol{x}))} \, dF(\boldsymbol{x}) \\
&= \int \pi_1(\boldsymbol{x})(1 - \pi_1(\boldsymbol{x})) \, dF(\boldsymbol{x}) \\
&= M_1^{\text{logistic}},
\end{aligned}
$$

which is (14). (This is equivalent to a logistic regression predicting whether $y_i = 1$ regardless of K.)

E.4 Mixed block

For the α-β mixed block, we have for all $k \in \{1, \ldots, K - 1\}$

$$
-\mathbb{E}\left[\frac{\partial^2}{\partial \alpha_k \partial \boldsymbol{\beta}} \mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] = \mathbb{E}\left[\boldsymbol{X}_1 p_k(\boldsymbol{X}_1)(1 - p_k(\boldsymbol{X}_1))(\pi_k(\boldsymbol{X}_1) + \pi_{k+1}(\boldsymbol{X}_1)) \right].
$$

Then

$$
-\mathbb{E}\left[\frac{\partial^2}{\partial \alpha_k \partial \boldsymbol{\beta}} \mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] = J_k^{\boldsymbol{x}} + \tilde{J}_{k+1}^{\boldsymbol{x}}, \qquad k \in \{1, \ldots, K - 1\},
$$

so $I_{\beta\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{(K-1) \times p}$ has the form

$$
I_{\beta\alpha}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} J_1^{\boldsymbol{x}} + \tilde{J}_2^{\boldsymbol{x}} \\ J_2^{\boldsymbol{x}} + \tilde{J}_3^{\boldsymbol{x}} \\ \vdots \\ J_{K-1}^{\boldsymbol{x}} + \tilde{J}_K^{\boldsymbol{x}} \end{pmatrix},
$$

as in (9). In the case of logistic regression predicting whether $y_i = 1$,

$$
I_{\beta\alpha}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = J_1^{\boldsymbol{x};\,\text{logistic}} + \tilde{J}_2^{\boldsymbol{x};\,\text{logistic}} = \int \boldsymbol{x} \pi_1(\boldsymbol{x})[1 - \pi_1(\boldsymbol{x})] \, dF(\boldsymbol{x}),
$$

verifying (15).

E.5 Beta block

Finally, for the β block, we have

$$
\begin{aligned}
I_{\beta\beta}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= -\mathbb{E}\left[\nabla_{\boldsymbol{\beta}}^2 \mathcal{L}^{\text{prop. odds}}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^\top \sum_{k=1}^{K} \left(\mathbb{1}\{y_i = k+1\} + \mathbb{1}\{y_i = k\}\right) p_k(\boldsymbol{X}_1)(1 - p_k(\boldsymbol{X}_1)) \mid \boldsymbol{X}\right]\right] \\
&= \mathbb{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^\top \sum_{k=1}^{K} \left(\pi_{k+1}\left(\boldsymbol{X}_{1\cdot}\right) + \pi_k\left(\boldsymbol{X}_{1\cdot}\right)\right) p_k(\boldsymbol{X}_1)(1 - p_k(\boldsymbol{X}_1))\right] \\
&= \sum_{k=1}^{K} \left(J_k^{\boldsymbol{x}\boldsymbol{x}^\top} + \tilde{J}_k^{\boldsymbol{x}\boldsymbol{x}^\top}\right),
\end{aligned}
$$

as in (10) (recall that $J_K^{\boldsymbol{x}\boldsymbol{x}^\top} = 0$ and $\tilde{J}_1^{\boldsymbol{x}\boldsymbol{x}^\top} = 0$ for all $\boldsymbol{x}$). In the case of logistic regression predicting whether $y_i = 1$ (for any K),

$$
I_{\beta\beta}^{\text{logistic}}(\alpha_1, \boldsymbol{\beta}) = J_1^{\boldsymbol{x}\boldsymbol{x}^\top;\,\text{logistic}} + \tilde{J}_2^{\boldsymbol{x}\boldsymbol{x}^\top;\,\text{logistic}} = \int \boldsymbol{x}\boldsymbol{x}^\top \pi_1(\boldsymbol{x})[1 - \pi_1(\boldsymbol{x})]\, dF(\boldsymbol{x}),
$$

matching (16).

E.6 Verifying the Asymptotic Distribution of Each Estimator

By the theorem on p. 145 of Serfling [1980, Section 4.2.2, multidimensional generalization on p. 148], the result holds if we can verify three regularity conditions. Before we do, we will define the set $\boldsymbol{\Theta}$ of feasible parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Since

$$
\alpha_1 < \alpha_2 < \ldots < \alpha_{K-1},
$$

where the strict inequality follows from our assumption that no class has probability 0 for any $x \in [-1, 1]$, define the set $\mathcal{A} \subset \mathbb{R}^{K-1}$ of points satisfying this constraint. Then define $\boldsymbol{\Theta} := \mathcal{A} \times \mathbb{R}^p$, so that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \boldsymbol{\Theta}$.

Now we state and verify the needed regularity conditions.

1. (R1) *The third derivatives of the log likelihood with respect to each parameter* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *exist for all* $\boldsymbol{x} \in \mathcal{S}$. This condition holds for both the proportional odds model and logistic regression because every entry of the Hessian matrices in Section (E.2) are differentiable in every parameter for any $K \geq 1$.

2. (R2) *For each* $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \in \boldsymbol{\Theta}$, *for all* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *in a neighborhood of* $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ *it holds that (i) the element-wise absolute values of the gradients and Hessians of the likelihood are bounded by functions of* $\boldsymbol{x}$ *with finite integral over* $\boldsymbol{x} \in \mathcal{S}$, *and (ii) the element-wise*

absolute values of the third derivatives of the log likelihood is bounded by a function of $\boldsymbol{x}$ *with finite expectation with respect to* $\boldsymbol{X}$. Because $\boldsymbol{X}$ has bounded support, for these integrals and expectations to be finite it is enough for the bounding functions to over $\mathcal{S}$ to be finite constants—that is, it is enough to find upper bounds on the absolute values of the gradients and Hessians of the likelihoods and the third derivatives of the log likelihoods. The logistic regression likelihood

$$\prod_{i=1}^{n} \frac{\exp\left\{-\mathbb{1}\left\{y_i = 1\right\}(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}_i)\right\}}{1 + \exp\{-(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}_i)\}}$$

has continuous second derivatives and therefore its gradient and Hessian both have finite element-wise absolute value. The same is true of the proportional odds likelihood (35) when all outcomes have positive probability for all $\boldsymbol{x} \in \mathcal{S}$, that is, $\alpha_1 < \ldots < \alpha_{K-1}$. Finally, examining again the Hessian matrices in Section E.2, we see that they have continuous derivatives in every parameter for any $K \geq 2$, so the third derivatives of the log likelihoods are bounded for any $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \in \boldsymbol{\Theta}$ for all $\boldsymbol{x} \in \mathcal{S}$.

3. (R3) *The Fisher information matrices exist and are finite and positive definite.* One can see that both of the Fisher information matrices are finite entrywise for all $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \boldsymbol{\Theta}$ by examining the matrices and noting that the probabilities for all of the classes are strictly greater than 0 over $\mathcal{S}$ by assumption. To verify the positive definiteness of the Fisher information matrices

$$-\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \boldsymbol{\theta}^\top} \mathcal{L}(\boldsymbol{\theta})\right],$$

it is enough to show that the log likelihood for each model is strictly concave, which implies that $\frac{\partial^2}{\partial \boldsymbol{\theta} \boldsymbol{\theta}^\top} \mathcal{L}(\boldsymbol{\theta})$ is almost surely negative definite (since the log likelihood is twice differentiable). Strict concavity of the logistic regression log likelihood

$$\sum_{i=1}^{n} \left[-\mathbb{1}\left\{y_i = 1\right\}(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}_i) - \log\left(1 + e^{-(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}_i)}\right)\right]$$

is easily seen, and Pratt [1981] provides a proof that the log likelihood for the proportional odds model is strictly concave when the intercepts $\alpha_1, \ldots, \alpha_{K-1}$ are not equal.

F Proof of Theorem 3

We prove Theorem 3 in Section F.1, and Section F.2 contains proofs of the supporting lemmas.

F.1 Proof of Theorem 3

The proof will proceed as follows. First we will show that PRESTO is a member of a class of models described by Ekvall and Bottai [2022], which means we can bound the estimation error of the parameters of PRESTO in a finite sample under their Theorem 3 once we show that its assumptions are satisfied. Their result depends on the probability of a particular event $\mathcal{C}_{\kappa,n,p_n(K-1)}$, and in Proposition 7 we prove a lower bound on $\mathbb{P}\left(\mathcal{C}_{\kappa,n,p_n(K-1)}\right)$ that tends towards 1 as $n \to \infty$. This leads to the consistency of PRESTO.

In the notation of Ekvall and Bottai [2022], we can express the objective function for the PRESTO estimator from (7) as $R\{b(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\} - R\{a(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\} + \lambda_n \|\boldsymbol{\theta}\|_1$, where $R(\cdot) = F(\cdot)$, the logistic cumulative distribution function; $(a(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}); b(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}))^\top = \boldsymbol{Z}_i^\top \boldsymbol{\theta} + \boldsymbol{m}_i$ where

$$
\boldsymbol{Z}_i := \begin{pmatrix} \mathbb{1}\{y_i \geq 2\}\boldsymbol{x}_i & \mathbb{1}\{y_i \leq K-1\}\boldsymbol{x}_i \\ \mathbb{1}\{y_i \geq 3\}\boldsymbol{x}_i & \mathbb{1}\{2 \leq y_i \leq K-1\}\boldsymbol{x}_i \\ \vdots & \vdots \\ \mathbb{1}\{y_i = K\}\boldsymbol{x}_i & \mathbb{1}\{y_i = K-1\}\boldsymbol{x}_i \end{pmatrix} \in \mathbb{R}^{p_n(K-1)\times 2},
$$

$$
\boldsymbol{\theta} := \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\psi}_2 \\ \vdots \\ \boldsymbol{\psi}_{K-1} \end{pmatrix} \in \mathbb{R}^{p_n(K-1)}, \qquad \text{and}
$$

$$
\boldsymbol{m}_i := \begin{pmatrix} \begin{cases} -\infty, & y_i = 1, \\ \alpha_{k-1}, & y_i = k \ (k \geq 2) \end{cases} \\ \begin{cases} \alpha_k, & y_i = k \ (k < K-1) \\ \infty, & y_i = K \end{cases} \end{pmatrix} \in \overline{\mathbb{R}}^2,
$$

where $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ denotes the extended real number system; and, as elsewhere in the paper, $\boldsymbol{\psi}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}$ for $k \in \{2, \ldots, K-1\}$. Observe that this is a special case of model (1) of Ekvall and Bottai [2022]. Now, since $\boldsymbol{\Theta} = \mathbb{R}^{p_n(K-1)}$ is open and the standard logistic density $r(t) = \exp\{-t\}/\left(1 + \exp\{-t\}\right)^2$ is strictly log-concave, strictly positive, and continuously differentiable on $\mathbb{R}$, assumptions (a) and (b) of Theorem 3 in Ekvall and Bottai [2022] are satisfied, assumption $S(C_4)$ is sufficient for assumption (c), and assumption (e) is satisfied for $C_3 = c_2$.

We now discuss Assumption 1 from Ekvall and Bottai [2022]. Note that a decision boundary crosses at some $\boldsymbol{x} \in \mathcal{S}$ if and only if for some $k \in \{2, \ldots, K-1\}$ it holds that

$$
\begin{aligned}
F\left(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}\right) - F\left(\alpha_{k-1} + \boldsymbol{\beta}_{k-1}^\top \boldsymbol{x}\right) &\leq 0 \\
\iff \qquad b(k, \boldsymbol{x}, \boldsymbol{\theta}) - a(k, \boldsymbol{x}, \boldsymbol{\theta}) &\leq 0.
\end{aligned}
$$

Observe that for $k = 1$ it holds that $m_{i1} = -\infty$ and for $k = K$ it holds that $m_{i2} = \infty$; that is, an element of $\boldsymbol{m}_i$ is infinite. So Assumption $T(C_4)$ is sufficient for it to either hold that $\boldsymbol{Z}_i^\top \boldsymbol{\theta} + \boldsymbol{m}_i \in E$ where $E \subseteq \{\boldsymbol{t} \in \mathbb{R}^2 : t_1 < t_2\}$ or an element of $\boldsymbol{m}_i$ is infinite. To satisfy Assumption 1, it only remains to show that there exists such an E that is compact. Note that $\|\boldsymbol{\theta}_*\|_\infty$ is bounded under Assumption $S(s, C_4)$ and $\|\boldsymbol{\theta}_*\|_0 \leq s$ under our sparsity assumption, so $\|\boldsymbol{\theta}_*\|_1$ is bounded due to Hölder's inequality. Lastly $\boldsymbol{m}_i$ is bounded because $\max_{k \in \{1,\ldots,K-1\}} |\alpha_k| \leq C_4$ under Assumption $T(C_4)$, so we have shown that we can find such an E that is bounded. Choose one that is closed and we have E compact.

We have shown that the assumptions of Theorem 3 of Ekvall and Bottai [2022] are satisfied, so we have that for n large enough that $p_n \geq K - 1$, with probability at least $\mathbb{P}\left(\mathcal{C}_{\kappa,n,p_n(K-1)}\right) - [p_n(K-1)]^{-c_3} \geq \mathbb{P}\left(\mathcal{C}_{\kappa,n,p_n(K-1)}\right) - (p_n^2)^{-c_3}$

$$\left\|\hat{\boldsymbol{\theta}}^{\lambda_n} - \boldsymbol{\theta}_*\right\|_1 \leq c_5 \sqrt{\frac{\log(p_n(K-1))}{n}} \leq c_5 \sqrt{\frac{\log(p_n^2)}{n}} = c_5 \sqrt{2} \sqrt{\frac{\log p_n}{n}}, \tag{36}$$

where c_3 and c_5 are constants from Ekvall and Bottai [2022], and $\mathcal{C}_{\kappa,n,p_n(K-1)}$ is defined as follows. For a set $\mathcal{A} \subseteq \{1, \ldots, p_n(K-1)\}$, define $\boldsymbol{\theta}_\mathcal{A} \in \mathbb{R}^{p_n(K-1)}$ to have j^{th} entry

$$(\boldsymbol{\theta}_\mathcal{A})_j := \begin{cases} \boldsymbol{\theta}_j, & j \in \mathcal{A}, \\ 0, & j \notin \mathcal{A}, \end{cases}$$

and define $\boldsymbol{\theta}_{\mathcal{A}^c} := \boldsymbol{\theta} - \boldsymbol{\theta}_\mathcal{A}$. Then for an s-sparse (in the sense of Assumption $S(s, c)$) $\boldsymbol{\theta}_*$ with support set $\mathcal{S}$, define

$$\mathbb{C}(\mathcal{S}) := \left\{\boldsymbol{\theta} \in \mathbb{R}^{p_n(K-1)} : \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_1 \leq 3 \|\boldsymbol{\theta}_\mathcal{S}\|_1\right\}.$$

We interpret $\mathbb{C}(\mathcal{S})$ to be a set of approximately s-sparse vectors (the vectors would be exactly s-sparse if $\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_1 = 0$). Then for $\kappa > 0$ and $n \in \mathbb{N}$, define

$$\mathcal{C}_{\kappa,n,p_n(K-1)} := \left\{(\boldsymbol{X}, \boldsymbol{y}) : \inf_{\{\boldsymbol{\theta} \in \mathbb{C}(\mathcal{S}) : \|\boldsymbol{\theta}\|_2 = 1\}} \left\{\boldsymbol{\theta}^\top \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\top\right) \boldsymbol{\theta}\right\} \geq \kappa\right\}.$$

If the set of $\boldsymbol{\theta}$ over which this condition must hold were $\mathbb{R}^{p_n(K-1)}$, this would be a minimum eigenvalue condition on $\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\top$. This condition is sometimes called a *restricted eigenvalue condition* [Bickel et al., 2009]. We bound $\mathbb{P}\left(\mathcal{C}_{\kappa,n,p_n(K-1)}\right)$ in Proposition 7.

Proposition 7. Suppose the assumptions of Theorem 3 hold. Let

$$\pi_{\text{rare,min}} := \inf_{\boldsymbol{x} \in \mathcal{S}, k \in \{1,\ldots,K\}} \{\mathbb{P}(y_i = k \mid \boldsymbol{x})\},$$

and observe that $\pi_{\text{rare,min}} > 0$ under Assumptions $X(\mathbb{R}^{p_n})$, $S(s, C_4)$, and $T(C_4)$. Assume n is large enough so that

$$n\pi_{\text{rare, min}} > \max\left\{2\left(C\sqrt{p_n} + \sqrt{\frac{a}{\lambda_{\min}^*}}\right)^2, 2\right\} \tag{37}$$

for some $a > 0$ (recall from the statement of Theorem 3 that we assumed $\lambda^*_{\min} := \min_{k\in\{1,\ldots,K\}} \lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^\top \mid y_i\right.\right.$ b for a fixed $b > 0$). Then

$$\mathbb{P}\left(\mathcal{C}_{a\pi_{\text{rare,min}}/4,n,p_n(K-1)}\right) \geq 1-2K\exp\left\{-\frac{1}{2}\pi^2_{\text{rare,min}}n\right\}-2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\}.$$

Proof. Provided in Section F.2. □

Observe that since $p_n \leq C_1 n^{C_2}$ for some $C_1 > 0, C_2 \in (0,1)$ this probability tends to 1 as $n \to \infty$. Lemma 8, below, along with (36) then shows that with probability at least

$$1 - p_n^{-C_5} - 2K\exp\left\{-\frac{1}{2}\pi^2_{\text{rare,min}}n\right\} - 2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\}$$

it holds that

$$\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2 \leq \left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_1 \leq C_6\sqrt{\frac{\log p_n}{n}},$$

where $C_5 := 2c_3$ and $C_6 := c_5\sqrt{2}(K-1)$ (where c_5 depends on the sparsity level s).

Lemma 8. $\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2 \leq \left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_1 \leq (K-1)\left\|\hat{\boldsymbol{\theta}}^{\lambda_n} - \boldsymbol{\theta}_*\right\|_1.$

Proof. Provided in Section F.2. □

Finally, we can now show consistency by showing that the random variable $\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2$ converges in probability to 0. For any $\epsilon > 0$,

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2 < \epsilon\right)\\
\overset{(*)}{\geq}\ &\lim_{n\to\infty}\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2 < C_6\sqrt{\frac{\log p_n}{n}}\right)\\
\geq\ &\lim_{n\to\infty}\left(1 - p_n^{-C_5} - 2K\exp\left\{-\frac{1}{2}\pi^2_{\text{rare,min}}n\right\} - 2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\}\right)\\
=\ &1,
\end{aligned}$$

where $(*)$ follows because for large enough n, $\epsilon > C_6\sqrt{\log(p_n)/n}$. This establishes Theorem 3. All that remains is to provide proofs for the supporting lemmas and proposition, which we do in the following section.

F.2 Supporting Results for Proof of Theorem 3

Proof of Proposition 7. We will show that there is a high probability event such that for $a > 0$ it holds that

$$\inf_{\{\boldsymbol{\theta} \in \mathbb{C}(\mathcal{S}) : \|\boldsymbol{\theta}\|_2 = 1\}} \left\{ \boldsymbol{\theta}^\top \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \right) \boldsymbol{\theta} \right\} > \frac{1}{2} a \pi_{\text{rare,min}}.$$

Let

$$\boldsymbol{A}^{(1)} = \begin{pmatrix} \mathbf{0}_{K-1} & \boldsymbol{e}_1 \end{pmatrix} \in \mathbb{R}^{(K-1)\times 2}$$

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} \mathbf{1}_{k-1} & \mathbf{1}_{k-1} \\ 0 & 1 \\ \mathbf{0}_{K-k-1} & \mathbf{0}_{K-k-1} \end{pmatrix} = \begin{pmatrix} \sum_{k'=1}^{k-1} \boldsymbol{e}_{k'} & \sum_{k'=1}^{k} \boldsymbol{e}_{k'} \end{pmatrix} \in \mathbb{R}^{(K-1)\times 2}, \; k \in \{2, \ldots, K-1\};$$

and

$$\boldsymbol{A}^{(K)} = \begin{pmatrix} \mathbf{1}_{K-1} & \mathbf{0}_{K-1} \end{pmatrix} = \begin{pmatrix} \sum_{k'=1}^{K-1} \boldsymbol{e}_{k'} & \mathbf{0}_{K-1} \end{pmatrix} \in \mathbb{R}^{(K-1)\times 2},$$

where $\mathbf{0}_n$ and $\mathbf{1}_n$ are n-vectors of zeroes and ones (respectively) and $\boldsymbol{e}_k$ is the standard basis vector in $\mathbb{R}^{K-1}$ with a 1 in the k^{th} entry and zeroes elsewhere, and $\boldsymbol{A}^{(k)} \in \mathbb{R}^{(K-1)\times 2}$ for all k. Note that

$$\boldsymbol{Z}_i = \boldsymbol{A}^{(y_i)} \otimes \boldsymbol{x}_i. \tag{38}$$

Let

$$\boldsymbol{B} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_1 & \sum_{k'=1}^{2} \boldsymbol{e}_{k'} & \cdots & \sum_{k'=1}^{K-1} \boldsymbol{e}_{k'} \end{pmatrix} \in \mathbb{R}^{(K-1)\times(K-1)};$$

that is; the columns $\boldsymbol{B}_k = \sum_{\ell=1}^{k} \boldsymbol{e}_\ell$. We will make use of the following lemmas, which we prove later in this section:

Lemma 9.

$$\sum_{k=1}^{K} \frac{n_k}{n} \boldsymbol{A}^{(k)} \left(\boldsymbol{A}^{(k)} \right)^\top = \boldsymbol{B} \boldsymbol{D} \boldsymbol{B}^\top,$$

where

$$\boldsymbol{D} := \text{diag}\left(\frac{n_1 + n_2}{n}, \ldots, \frac{n_{K-1} + n_K}{n} \right) \in \mathbb{R}^{(K-1)\times(K-1)}.$$

Lemma 10. $\sigma_{\min}^2(\boldsymbol{B}) \geq 1/2$.

Lemma 11. For arbitrary matrices $\overline{\boldsymbol{A}}$, $\overline{\boldsymbol{B}}$, and $\overline{\boldsymbol{C}}$, if $\overline{\boldsymbol{B}} \succeq \overline{\boldsymbol{C}}$ and $\overline{\boldsymbol{A}} \succeq \mathbf{0}$ then $\overline{\boldsymbol{A}} \otimes \overline{\boldsymbol{B}} \succeq \overline{\boldsymbol{A}} \otimes \overline{\boldsymbol{C}}$.

Lemma 12. With probability at least $1-K\exp\left\{-\frac{1}{2}\pi_{\text{rare,min}}^2 n\right\}$ it holds that $\min\{n_1,\ldots,n_K\} > \pi_{\text{rare,min}} n/2$, where $n_k := \sum_{i=1}^n \mathbb{1}\{y_i = k\}$ is the number of observations in class k.

Lemma 13. Under the assumptions of Proposition 7, there exist constants $c, C > 0$ such that $\lambda_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \geq a$ for all $k \in \{1,\ldots,K-1\}$ with probability at least

$$1 - 2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\} - K\exp\left\{-\frac{1}{2}\pi_{\text{rare,min}}^2 n\right\},$$

where n_k is the number of observations in class k and $\lambda^*_{\min} := \min_{k\in\{1,\ldots,K\}}\left\{\lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{X}^\top \boldsymbol{X} \mid y = k\right]\right)\right\}$.

Using a union bound, there exists an event with probability at least

$$1 - 2K\exp\left\{-\frac{1}{2}\pi_{\text{rare,min}}^2 n\right\} - 2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\}$$

on which the conclusions of all of the above lemmas hold. On this event we have

$$\begin{aligned}
\inf_{\{\boldsymbol{\theta}\in\mathbb{C}(\mathcal{S}):\|\boldsymbol{\theta}\|_2=1\}} \left\{\boldsymbol{\theta}^\top \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)\boldsymbol{\theta}\right\} &\geq \inf_{\{\boldsymbol{\theta}\in\mathbb{R}^{pn(K-1)}:\|\boldsymbol{\theta}\|_2=1\}} \left\{\boldsymbol{\theta}^\top \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)\boldsymbol{\theta}\right\} \\
&= \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right) \\
&\overset{(a)}{=} \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{A}^{(y_i)}\otimes\boldsymbol{x}_i\right)\left(\boldsymbol{A}^{(y_i)}\otimes\boldsymbol{x}_i\right)^\top\right) \\
&= \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{A}^{(y_i)}\left(\boldsymbol{A}^{(y_i)}\right)^\top\right)\otimes\left(\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right) \\
&= \lambda_{\min}\left(\frac{1}{n}\sum_{k=1}^K\sum_{i:y_i=k} \left(\boldsymbol{A}^{(y_i)}\left(\boldsymbol{A}^{(y_i)}\right)^\top\right)\otimes\left(\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right) \\
&= \lambda_{\min}\left(\frac{1}{n}\sum_{k=1}^K \left(\boldsymbol{A}^{(k)}\left(\boldsymbol{A}^{(k)}\right)^\top\right)\otimes\left(\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right) \\
&= \lambda_{\min}\left(\sum_{k=1}^K \left(\frac{n_k}{n}\boldsymbol{A}^{(k)}\left(\boldsymbol{A}^{(k)}\right)^\top\right)\otimes\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right) \\
&\overset{(b)}{\geq} \lambda_{\min}\left(\sum_{k=1}^K \left(\frac{n_k}{n}\boldsymbol{A}^{(k)}\left(\boldsymbol{A}^{(k)}\right)^\top\right)\otimes\left(a\boldsymbol{I}_p\right)\right) \\
&\overset{(c)}{=} \lambda_{\min}\left(\sum_{k=1}^K \frac{n_k}{n}\boldsymbol{A}^{(k)}\left(\boldsymbol{A}^{(k)}\right)^\top\right)\lambda_{\min}\left(a\boldsymbol{I}_p\right) \\
&\overset{(d)}{=} \lambda_{\min}\left(\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top\right)\lambda_{\min}\left(a\boldsymbol{I}_p\right) \\
&\overset{(e)}{\geq} a\min_{k\in\{1,\ldots,K-1\}}\left\{\frac{n_k+n_{k+1}}{n}\right\}\lambda_{\min}\left(\boldsymbol{B}\boldsymbol{B}^\top\right) \\
&\overset{(f)}{\geq} \frac{a}{4}\min_{k\in\{1,\ldots,K-1\}}\left\{\frac{n_k+n_{k+1}}{n}\right\} \\
&\overset{(g)}{>} \frac{a\pi_{\text{rare,min}}}{4},
\end{aligned}$$

where in (a) we used (38), (b) holds with high probability by Lemmas 11 and 13 because $\lambda_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right) \geq a$ implies $\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top \succeq a\boldsymbol{I}_p$ for all k, in (c) we used the fact that for matrices $\boldsymbol{M},\boldsymbol{N}$ the eigenvalues of $\boldsymbol{M}\otimes\boldsymbol{N}$ are the products of the eigenvalues of $\boldsymbol{M}$ and $\boldsymbol{N}$ and both of the factor matrices are positive semidefinite, in (d) we applied

Lemma 9, (e) follows from $\boldsymbol{BDB}^\top \succeq \min_{k\in\{1,\ldots,K-1\}}\left\{\frac{n_k+n_{k+1}}{n}\right\}\boldsymbol{BB}^\top$, in (f) we applied Lemma 10, and (g) follows from Lemma 12.

□

Proof of Lemma 8. For convenience, denote $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$ and $\boldsymbol{\theta}_k = \boldsymbol{\psi}_k$, $k \in \{2,\ldots,K-1\}$. Let

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}^{\lambda_n} &= \left(\left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},1}^{\lambda_n}\right)^\top, \left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},2}^{\lambda_n}\right)^\top, \ldots, \left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},K-1}^{\lambda_n}\right)^\top\right)^\top \\
&:= \left(\left(\hat{\boldsymbol{\theta}}_1^{\lambda_n} - \boldsymbol{\theta}_1\right)^\top, \left(\hat{\boldsymbol{\theta}}_2^{\lambda_n} - \boldsymbol{\theta}_2\right)^\top, \ldots, \left(\hat{\boldsymbol{\theta}}_{K-1}^{\lambda_n} - \boldsymbol{\theta}_{K-1}\right)^\top\right)^\top.
\end{aligned}$$

Note that $\boldsymbol{\beta}_k = \sum_{k'\le k}\boldsymbol{\theta}_{k'}$. Let $\boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_{K-1}^\top)^\top$, and let

$$\hat{\boldsymbol{\beta}}_k^{\lambda_n} := \sum_{k'\le k}\hat{\boldsymbol{\theta}}_{k'}, \qquad k \in \{1,\ldots,K-1\}$$

be the estimates of $\boldsymbol{\beta}$ yielded from the estimates of $\boldsymbol{\theta}$. Let

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta}}^{\lambda_n} &= \left(\left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta},1}^{\lambda_n}\right)^\top, \left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta},2}^{\lambda_n}\right)^\top, \ldots, \left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta},K-1}^{\lambda_n}\right)^\top\right)^\top \\
&:= \left(\left(\hat{\boldsymbol{\beta}}_1^{\lambda_n} - \boldsymbol{\beta}_1\right)^\top, \left(\hat{\boldsymbol{\beta}}_2^{\lambda_n} - \boldsymbol{\beta}_2\right)^\top, \ldots, \left(\hat{\boldsymbol{\beta}}_{K-1}^{\lambda_n} - \boldsymbol{\beta}_{K-1}\right)^\top\right)^\top,
\end{aligned}$$

and observe that

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta}}^{\lambda_n} &= \left(\left(\hat{\boldsymbol{\beta}}_1^{\lambda_n} - \boldsymbol{\beta}_1\right)^\top, \left(\hat{\boldsymbol{\beta}}_2^{\lambda_n} - \boldsymbol{\beta}_2\right)^\top, \ldots, \left(\hat{\boldsymbol{\beta}}_{K-1}^{\lambda_n} - \boldsymbol{\beta}_{K-1}\right)^\top\right)^\top \\
&= \left(\left(\hat{\boldsymbol{\theta}}_1^{\lambda_n} - \boldsymbol{\theta}_1\right)^\top, \left(\sum_{k'\le 2}\hat{\boldsymbol{\theta}}_{k'} - \sum_{k'\le 2}\boldsymbol{\theta}_{k'}\right)^\top, \ldots, \left(\sum_{k'\le K-1}\hat{\boldsymbol{\theta}}_{k'} - \sum_{k'\le K-1}\boldsymbol{\theta}_{k'}\right)^\top\right)^\top \\
&= \left(\left(\hat{\boldsymbol{\theta}}_1^{\lambda_n} - \boldsymbol{\theta}_1\right)^\top, \left(\sum_{k'\le 2}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k'}^{\lambda_n}\right)^\top, \ldots, \left(\sum_{k'\le K-1}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k'}^{\lambda_n}\right)^\top\right)^\top. \qquad (39)
\end{aligned}$$

Then

$$\begin{aligned}
\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta},1}^{\lambda_n}\right\|_1 &\overset{(a)}{\le} \sum_{k=1}^{K-1}\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\beta},k}^{\lambda_n}\right\|_1 \overset{(b)}{=} \sum_{k=1}^{K-1}\left\|\sum_{k'\le k}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k'}^{\lambda_n}\right\|_1 \overset{(c)}{\le} \sum_{k=1}^{K-1}\sum_{k'\le k}\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k'}^{\lambda_n}\right\|_1 = \sum_{k=1}^{K-1}(K-k)\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k}^{\lambda_n}\right\|_1 \\
&\le (K-1)\sum_{k=1}^{K-1}\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta},k}^{\lambda_n}\right\|_1 = (K-1)\left\|\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}^{\lambda_n}\right\|_1.
\end{aligned}$$

where in (a) and (c) we used the triangle inequality and in (b) we used (39). Lastly, $\left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_2 \leq \left\|\hat{\boldsymbol{\beta}}^{\lambda_n} - \boldsymbol{\beta}\right\|_1$ is a property of the ℓ_1 and ℓ_2 norms.

□

Proof of Lemma 9.

$$\begin{aligned}
\sum_{k=1}^{K} \frac{n_k}{n} \boldsymbol{A}^{(k)} \left(\boldsymbol{A}^{(k)}\right)^\top &= \frac{n_1}{n} \boldsymbol{A}^{(1)} \left(\boldsymbol{A}^{(1)}\right)^\top + \sum_{k=2}^{K-1} \frac{n_k}{n} \boldsymbol{A}^{(k)} \left(\boldsymbol{A}^{(k)}\right)^\top + \frac{n_K}{n} \boldsymbol{A}^{(K)} \left(\boldsymbol{A}^{(K)}\right)^\top \\
&= \frac{n_1}{n} \begin{pmatrix} \boldsymbol{0}_{K-1} & \boldsymbol{e}_1 \end{pmatrix} \begin{pmatrix} \boldsymbol{0}_{K-1}^\top \\ \boldsymbol{e}_1^\top \end{pmatrix} \\
&\quad + \sum_{k=2}^{K-1} \frac{n_k}{n} \begin{pmatrix} \sum_{k'=1}^{k-1} \boldsymbol{e}_{k'} & \sum_{k'=1}^{k} \boldsymbol{e}_{k'} \end{pmatrix} \begin{pmatrix} \sum_{k'=1}^{k-1} \boldsymbol{e}_{k'}^\top \\ \sum_{k'=1}^{k} \boldsymbol{e}_{k'}^\top \end{pmatrix} \\
&\quad + \frac{n_K}{n} \begin{pmatrix} \sum_{k'=1}^{K-1} \boldsymbol{e}_{k'} & \boldsymbol{0}_{K-1} \end{pmatrix} \begin{pmatrix} \sum_{k'=1}^{K-1} \boldsymbol{e}_{k'}^\top \\ \boldsymbol{0}_{K-1}^\top \end{pmatrix} \\
&= \frac{n_1}{n} \left(\boldsymbol{0}_{K-1}\boldsymbol{0}_{K-1}^\top + \boldsymbol{e}_1\boldsymbol{e}_1^\top\right) \\
&\quad + \sum_{k=2}^{K-1} \frac{n_k}{n} \left(\sum_{k'=1}^{k-1} \boldsymbol{e}_{k'} \sum_{k'=1}^{k-1} \boldsymbol{e}_{k'}^\top + \sum_{k'=1}^{k} \boldsymbol{e}_{k'} \sum_{k'=1}^{k} \boldsymbol{e}_{k'}^\top \right) \\
&\quad + \frac{n_K}{n} \left(\sum_{k'=1}^{K-1} \boldsymbol{e}_{k'} \sum_{k'=1}^{K-1} \boldsymbol{e}_{k'}^\top + \boldsymbol{0}_{K-1}\boldsymbol{0}_{K-1}^\top \right) \\
&= \frac{n_1}{n} \boldsymbol{B}_1 \boldsymbol{B}_1^\top + \sum_{k=2}^{K-1} \frac{n_k}{n} \left(\boldsymbol{B}_{k-1}\boldsymbol{B}_{k-1}^\top + \boldsymbol{B}_k \boldsymbol{B}_k^\top \right) \\
&\quad + \frac{n_K}{n} \boldsymbol{B}_{K-1}\boldsymbol{B}_{K-1}^\top \\
&= \sum_{k=1}^{K-1} \frac{n_k + n_{k+1}}{n} \boldsymbol{B}_k \boldsymbol{B}_k^\top \\
&= \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top.
\end{aligned}$$

□

Proof of Lemma 10. $\boldsymbol{B}$ is full rank with inverse

$$\boldsymbol{B}^{-1} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

We have

$$\|\boldsymbol{B}^{-1}\|_{\mathrm{op}} \le \sqrt{\|\boldsymbol{B}^{-1}\|_1 \|\boldsymbol{B}^{-1}\|_\infty} = \sqrt{2 \cdot 2} = 2,$$

where we have used that the ℓ_1 norm of each column and row is at most 2. Then the result follows from observing that $\sigma_{\min}(\boldsymbol{B}) = 1/\|\boldsymbol{B}^{-1}\|_{\mathrm{op}}$. □

Proof of Lemma 11. The matrix $\overline{\boldsymbol{A}} \otimes (\overline{\boldsymbol{B}} - \overline{\boldsymbol{C}})$ is the Kronecker product of two positive semidefinite matrices and is therefore positive semidefinite. □

Proof of Lemma 12. First we state a lemma we will use.

Lemma 14. Suppose that for all $\boldsymbol{x} \in \mathcal{S}$ and $k \in \{1, \ldots, K\}$ it holds that $\mathbb{P}(y_i = k \mid \boldsymbol{x}) \ge \pi_{\text{rare,min}}$. Let n_k be the number of observations in class k from a data set of size n. Then for any $q \in \{1, \ldots, \lfloor n\pi_{\text{rare,min}} \rfloor\}$,

$$\mathbb{P}\left(\bigcap_{k=1}^{K} \{n_k > q\}\right) \ge 1 - K \exp\left\{-2n\left(\pi_{\text{rare,min}} - \frac{q}{n}\right)^2\right\}.$$

Proof. Provided later in this section. □

Note that $n\pi_{\text{rare,min}}/2 \le \lfloor n\pi_{\text{rare,min}} \rfloor$ because $x/2 \le \lfloor x \rfloor$ for all $x \ge 2$ and $n\pi_{\text{rare,min}} \ge 2$ due to assumption (37). So the assumptions of Lemma 14 are satisfied for $q = n\pi_{\text{rare,min}}/2$, and there are more than $n\pi_{\text{rare,min}}/2$ observations in each class with probability at least

$$1 - K \exp\left\{-2n\left(\pi_{\text{rare,min}} - \pi_{\text{rare,min}}/2\}\right)^2\right\} = 1 - K \exp\left\{-\frac{1}{2}\pi_{\text{rare,min}}^2 n\right\}.$$

□

Proof of Lemma 13. We will prove the result by using a concentration inequality on the minimum singular value of a random matrix (which will correspond to the square root of the minimum eigenvalue of $\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i \boldsymbol{x}_i^\top$; recall that the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ are the squares of the singular values of $\boldsymbol{A}$). However, the result we use applies only to random matrices with isotropic second moment matrices, so we need to standardize $\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ by its inverse square root second moment matrix. We can relate this quantity to the minimum eigenvalue of $\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ by our claim (that we will verify later) that for a symmetric positive semidefinite matrix $\boldsymbol{S}$ and symmetric positive definite $\boldsymbol{\Sigma}$ it holds that

$$\lambda_{\min}(\boldsymbol{S}) \ge \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\right)\lambda_{\min}(\boldsymbol{\Sigma}). \tag{40}$$

Then substituting $\boldsymbol{S} = \frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k = \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^\top \mid y_i = k\right]$ (which we assumed has a strictly postive minimum eigenvalue and is therefore invertible) into (40) yield that on an event where $n_k = \sum_{i=1}^n \mathbb{1}\{y_i = k\} \geq 1$, we have

$$\begin{aligned}
\lambda_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right) &\geq \lambda_{\min}\left(\frac{1}{n_k}\boldsymbol{\Sigma}_k^{-1/2}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}\right)\lambda_{\min}\left(\boldsymbol{\Sigma}_k\right) \\
&\geq \lambda_{\min}\left(\frac{1}{n_k}\boldsymbol{\Sigma}_k^{-1/2}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}\right)\lambda_{\min}^*, \qquad (41)
\end{aligned}$$

where the second line uses our assumption from the statement of Theorem 3 and the fact that $\frac{1}{n_k}\boldsymbol{\Sigma}_k^{-1/2}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}$ is almost surely positive semidefinite. Next we lower-bound the minimum eigenvalue of the random matrix $\frac{1}{n_k}\boldsymbol{\Sigma}_k^{-1/2}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}$ using a result from Vershynin [2012]. Observe that the ψ_2 norm of a bounded random vector $\boldsymbol{T} \in \mathbb{R}^p$ can be upper-bounded as follows:

$$\begin{aligned}
\|\boldsymbol{T}\|_{\psi_2} &= \sup_{\boldsymbol{v}:\boldsymbol{v}\in\mathbb{R}^p,\|\boldsymbol{v}\|_2=1}\left\{\|\boldsymbol{T}^\top\boldsymbol{v}\|_{\psi_2}\right\} \\
&= \sup_{\boldsymbol{v}:\boldsymbol{v}\in\mathbb{R}^p,\|\boldsymbol{v}\|_2=1}\left\{\inf\left\{t>0:\mathbb{E}\exp\left(\frac{\left[\boldsymbol{T}^\top\boldsymbol{v}\right]^2}{t^2}\right)\leq 2\right\}\right\} \\
&\leq \inf\left\{t>0:\exp\left(\frac{\|\boldsymbol{T}\|_\infty^2}{t^2}\right)\leq 2\right\} \\
&= \frac{\|\boldsymbol{T}\|_\infty}{\sqrt{\log 2}}.
\end{aligned}$$

Therefore under our assumptions $\boldsymbol{x}_i \mid \boldsymbol{y}$ is bounded and therefore subgaussian, with ψ_2 norm at most $\|\boldsymbol{x}_i\|_{\psi_2} \leq \|\boldsymbol{x}\|_\infty/\sqrt{\log 2} = C_4/\sqrt{\log 2}$ for all k. So for any $k \in \{1,\ldots,K\}$, we have from Theorem 5.39 in Vershynin [2012] that for any $t \geq 0$ there exists $c > 0$ such that the event

$$\mathbb{P}\left(\sigma_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\right)\geq\sqrt{n_k}-C\sqrt{p_n}-t\mid\boldsymbol{y}\right)\geq 1-2\exp\{-ct^2\}$$

holds almost surely for $C = C_4^2/(\log 2)\sqrt{\log(9)/c_1}$, where C_4 is as defined in the statement of Theorem 3, c_1 is a constant from Vershynin [2012], $\sigma_{\min}(\cdot)$ denotes the minimum singular value, and if the set $\{i : y_i = k\}$ is empty we define $\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i$ to equal $\mathbf{0}_p$ (and note that the inequality then trivially holds with probability one in this case, because $n_k = 0$ so the right side is nonpositive). For $k \in \{1,\ldots,K\}$ and $t \geq 0$, define the event $\mathcal{E}_k(t)$ by

$$\mathcal{E}_k(t) := \left\{\sigma_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\right)\geq\sqrt{n_k}-C\sqrt{p_n}-t\right\},$$

and observe that

$$\mathbb{P}(\mathcal{E}_k(t)) \geq 1 - 2\exp\{-ct^2\} \tag{42}$$

by applying Theorem 5.39 and marginalizing over $\boldsymbol{y}$. Now we consider a particular choice of t. From our assumption (37) we have

$$\begin{aligned} n\pi_{\text{rare, min}} &> \; 2\left(C\sqrt{p_n} + \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2 \\ \iff \qquad \sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}} &> \; 0, \end{aligned}$$

so we can choose $t = \sqrt{n\pi_{\text{rare,min}}/2} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}$, yielding that on

$$\mathcal{E}_k\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right) \cap \{n_k \geq 1\}$$

we have

$$\begin{aligned} \sigma_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\right) &\geq \; \sqrt{n_k} - C\sqrt{p_n} - \sqrt{\frac{n\pi_{\text{rare,min}}}{2}} + C\sqrt{p_n} + \sqrt{\frac{a}{\lambda^*_{\min}}} \\ &= \; \sqrt{n_k} - \sqrt{\frac{n\pi_{\text{rare,min}}}{2}} + \sqrt{\frac{a}{\lambda^*_{\min}}}. \end{aligned} \tag{43}$$

That is, on this event we can lower-bound the minimum eigenvalue of $\frac{1}{n_k}\boldsymbol{\Sigma}_k^{-1/2}\sum_{i:y_i=k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}$ provided that n_k is at least 1 and large enough that the right side of (43) is nonnegative. Next we work on lower-bounding n_k with high probability. Consider the event

$$\mathcal{N} := \; \left\{n_k \geq \frac{n\pi_{\text{rare,min}}}{2} \qquad \forall k \in \{1, \ldots, K\}\right\}.$$

Notice that on $\mathcal{N}$ we have that the lower bound in (43) is at least $\sqrt{a/\lambda^*_{\min}}$ and $n_k \geq 1$ for all k due to (37). That is, for any k, on $\mathcal{N} \cap \mathcal{E}_k\left(\sqrt{n\pi_{\text{rare,min}}/2} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)$ inequality (43) holds and yields $\sigma_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\right) \geq \sqrt{a/\lambda^*_{\min}}$. Since the eigenvalues of $\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}$ are the squares of the singular values of $\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i$, on $\mathcal{N} \cap \mathcal{E}_k\left(\sqrt{n\pi_{\text{rare,min}}/2} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)$ we have

$$\lambda_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k}\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_k^{-1/2}\right) \geq \frac{a}{\lambda^*_{\min}} > 0.$$

Finally, substituting this into (41) we have that on $\left(\bigcap_{k=1}^K \mathcal{E}_k\left(\sqrt{n\pi_{\text{rare,min}}/2} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\text{min}}}}\right)\right) \cap \mathcal{N}$

$$\lambda_{\min}\left(\frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right) \geq \frac{a}{\lambda^*_{\min}}\lambda^*_{\min} = a \qquad \forall k \in \{1,\dots,K\}.$$

Using a union bound, this holds with probability at least

$$\begin{aligned}
&\mathbb{P}\left(\left(\bigcap_{k=1}^K \mathcal{E}_k\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)\right) \cap \mathcal{N}\right) \\
\geq\ & 1 - \sum_{k=1}^K \mathbb{P}\left(\mathcal{E}_k^c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)\right) - \mathbb{P}\left(\mathcal{N}^c\right) \\
\geq\ & 1 - 2K\exp\left\{-c\left(\sqrt{\frac{n\pi_{\text{rare,min}}}{2}} - C\sqrt{p_n} - \sqrt{\frac{a}{\lambda^*_{\min}}}\right)^2\right\} - K\exp\left\{-\frac{1}{2}\pi^2_{\text{rare,min}}n\right\},
\end{aligned}$$

where in the last step we applied (42) and Lemma 12.

Finally, we show (40). Observe that for any $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_2 = 1$

$$\begin{aligned}
\boldsymbol{v}^\top\boldsymbol{S}\boldsymbol{v} =\ & \left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\right)^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\right) \\
\geq\ & \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\right)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2^2 \\
=\ & \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v} \\
\geq\ & \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\right)\lambda_{\min}\left(\boldsymbol{\Sigma}\right),
\end{aligned}$$

proving the claim.

□

Proof of Lemma 14. By Hoeffding's inequality we have that for any $k \in \{1,\dots,K\}$ and any $q \in \{1,\dots,\lfloor n\pi_{\text{rare,min}}\rfloor\}$, $\mathbb{P}\left(n_k \leq q\right) \leq \exp\left\{-2n\left(\pi_{\text{rare,min}} - \frac{q}{n}\right)^2\right\}$. Then using a union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k=1}^K \{n_k > q\}\right) =\ & 1 - \mathbb{P}\left(\bigcup_{k=1}^K \{n_k \leq q\}\right) \\
\geq\ & 1 - \sum_{k=1}^K \mathbb{P}\left(n_k \leq q\right) \\
\geq\ & 1 - K\exp\left\{-2n\left(\pi_{\text{rare,min}} - \frac{q}{n}\right)^2\right\}.
\end{aligned}$$

□

G Estimating PRESTO

In this section we will use the notation and terminology of Wurm et al. [2021], with the exception of continuing to use our convention of K total categories. We fit PRESTO by reparameterizing the efficient coordinate descent algorithm used to estimate ℓ_1-penalized ordinal regression models in the R `ordinalNet` [Wurm et al., 2021] package, in much the same way that the generalized lasso can be implemented using reparameterization; see Section 4.5.1.1 of Hastie et al. [2015] for a textbook-level discussion. (See Section 4.5.1.1 of Hastie et al. 2015 for more details on how this is done.) The `ordinalNet` package implements ℓ_1-penalized ordinal regression, including an ℓ_1-penalized relaxation of the proportional odds model. The parameters we seek to model are $\boldsymbol{\eta}_i \in \mathbb{R}^{K-1}$, which model the probabilities of the K outcomes by the relation

$$\boldsymbol{\eta}_i = g(\boldsymbol{p}_i),$$

where $\boldsymbol{p}_i \in \mathcal{S}^{K-1}$ (where $\mathcal{S}^{K-1} := \{\boldsymbol{p} : \boldsymbol{p} \in (0,1)^{K-1}, \|p\|_1 < 1\}$ are the probabilities of outcomes $\{1, \ldots, K-1\}$ (in particular, $p_{ik} = \mathbb{P}(y_i = k)$ for $k \in \{1, \ldots, K-1\}$ and $\mathbb{P}(y_i = K) = 1 - \sum_{k=1}^{K-1} p_{ik}$) and $g : \mathcal{S}^{K-1} \to \mathbb{R}^{K-1}$ is an invertible function. For our model, the forwards cumulative probability model, where $p_{ik} = \mathbb{P}(y_i \leq k)$,

$$\left[g(\boldsymbol{p}_i)\right]_k = \log\left(\frac{\sum_{k'=1}^{k} p_{ik'}}{1 - \sum_{k'=1}^{k} p_{ik'}}\right), \qquad k \in \{1, \ldots, K\}.$$

We choose the nonparallel model

$$\boldsymbol{\eta}_i = \boldsymbol{c} + \boldsymbol{B}^\top \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i$ is the i^{th} row of $\boldsymbol{X}$, $\boldsymbol{c} \in \mathbb{R}^{K-1}$ is a vector of intercepts, and

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{\cdot 1} & \cdots & \boldsymbol{B}_{\cdot K-1} \end{bmatrix}$$

is a $p \times (K-1)$ matrix of coefficients. Observe that we can simply write

$$\boldsymbol{\eta}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{\beta},$$

for $\boldsymbol{\beta} \in \mathbb{R}^{(K-1)(p+1)}$ defined by

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{B}_{\cdot 1} \\ \vdots \\ \boldsymbol{B}_{\cdot K-1} \end{pmatrix}$$

and $\tilde{\boldsymbol{X}}_i \in \mathbb{R}^{(K-1)\times(K-1)(p+1)}$ defined by

$$\tilde{\boldsymbol{X}}_i = \begin{pmatrix} & \boldsymbol{x}_i^\top & \boldsymbol{0}_p^\top & \cdots & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top \\ & \boldsymbol{0}_p^\top & \boldsymbol{x}_i^\top & \cdots & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top \\ \boldsymbol{I}_{K-1} & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top & \cdots & \boldsymbol{x}_i^\top & \boldsymbol{0}_p^\top \\ & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top & \cdots & \boldsymbol{0}_p^\top & \boldsymbol{x}_i^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{K-1} & \boldsymbol{I}_{K-1} \otimes \boldsymbol{x}_i^\top \end{pmatrix}.$$

The R `ordinalNet` package solves the convex [Wurm et al., 2017, Pratt, 1981] optimization problem

$$\arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ell_i \left(g^{-1} \left(\tilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} \right) \right) + \lambda \sum_{q=K}^{(K-1)(p+1)} |\beta_q| \right\}$$

where

$$\ell_i(\boldsymbol{p}_i) = \sum_{k=1}^{K-1} \mathbb{1}\{y_i = k\} \log p_{ik} + \mathbb{1}\{y_i = K\} \log \left(1 - \sum_{k=1}^{K-1} p_{ik} \right).$$

We would like to place an ℓ_1 penalty on the first differences $B_{j,k+1} - B_{jk}$ for all $k \in \{1, \dots, K-2\}$. We can do this through the parameterization $\boldsymbol{\Psi} \in \mathbb{R}^{p \times (K-1)}$ defined by

$$\Psi_{jk} = \begin{cases} B_{j1}, & j \in [p], k = 1, \\ B_{jk} - B_{j,k-1}, & j \in [p], k \in \{2, \dots, K-1\}. \end{cases}$$

Observe that these matrices are related by

$$B_{jk} = \sum_{k'=1}^{k} \Psi_{jk'},$$

so

$$\eta_{ik} = c_{0k} + \boldsymbol{B}_{\cdot k}^\top \boldsymbol{x}_i = c_{0k} + \sum_{k'=1}^{k} \boldsymbol{\Psi}_{\cdot k'}^\top \boldsymbol{x}_i, \qquad k \in \{1, \dots, K-1\}.$$

Therefore we can simply write

$$\boldsymbol{\eta}_i = \tilde{\boldsymbol{X}}_i' \boldsymbol{\beta}',$$

for $\boldsymbol{\beta}' \in \mathbb{R}^{(K-1)(p+1)}$ defined by

$$\boldsymbol{\beta}' = \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{\Psi}_{\cdot 1} \\ \vdots \\ \boldsymbol{\Psi}_{\cdot K-1} \end{pmatrix}$$

and $\tilde{\boldsymbol{X}}_i' \in \mathbb{R}^{(K-1) \times (K-1)(p+1)}$ defined by

$$\tilde{\boldsymbol{X}}_i' = \begin{pmatrix} & \boldsymbol{x}_i^\top & \boldsymbol{0}_p^\top & \cdots & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top \\ & \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top & \cdots & \boldsymbol{0}_p^\top & \boldsymbol{0}_p^\top \\ \boldsymbol{I}_{K-1} & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top & \cdots & \boldsymbol{x}_i^\top & \boldsymbol{0}_p^\top \\ & \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top & \cdots & \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top \end{pmatrix}. \tag{44}$$

We then seek to solve the slightly modified optimization problem (modification highlighted)

$$\arg\min_{\boldsymbol{\beta}'} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ell_i \left(g^{-1} \left(\underbrace{\left(\tilde{\boldsymbol{X}}_i' \right)^\top}_{(*)} \boldsymbol{\beta}' \right) \right) + \lambda \sum_{q=K}^{(K-1)(p+1)} \left| \beta_q' \right| . \right\}$$

Though this can not be implemented within the framework of the existing `ordinalNet` package, the above modification only requires changing a handful of lines of the publicly available source code in the `ordinalNet` package. We will make our code publicly available in order to show our implementation before the camera-ready deadline. Though our implementation is simple, using the modified design matrix (the change above) could cause convergence of parameter estimation to be slow, because the resulting lasso problem effectively has highly correlated features, which slows the convergence of coordinate descent. See Section 4.5.1.1 of Hastie et al. [2015] for a textbook-level discussion of this point.