Binary Outcome Models with Extreme Covariates: Estimation and Prediction[*]

Laura Liu

University of Pittsburgh

Yulong Wang

Syracuse University

First version: July 7, 2024
This version: February 21, 2025

Abstract

This paper presents a novel semiparametric method to study the effects of extreme events on binary outcomes and subsequently forecast future outcomes. Our approach, based on Bayes' theorem and regularly varying (RV) functions, facilitates a Pareto approximation in the tail without imposing parametric assumptions beyond the tail. We analyze cross-sectional as well as static and dynamic panel data models, incorporate additional covariates, and accommodate the unobserved unit-specific tail thickness and RV functions in panel data. We establish consistency and asymptotic normality of our tail estimator, and show that our objective function converges to that of a panel Logit regression on tail observations with the log extreme covariate as a regressor, thereby simplifying implementation. The empirical application assesses whether small banks become riskier when local housing prices sharply decline, a crucial channel in the 2007–2008 financial crisis.

Keywords: Binary outcome model, heavy tail, Pareto approximation, panel data, partial effects, forecast

JEL classification: C14, C21, C23, C25, C53

1 Introduction

Binary outcome models are widely used in both empirical microeconomics and empirical macroeconomics research. For example, microeconomic studies may be interested in the determinants of individuals' labor force participation decisions, and macroeconomic analyses may seek to forecast the probability of recessions or country defaults. Recently, extreme events, such as the Covid-19 pandemic and its aftermath, sustained periods of high inflation, and increasingly frequent extreme weather, have highlighted the importance of studying the effect of extreme covariates on binary outcomes.

As an illustration, consider a simple example with cross-sectional data, where we have a random sample of $\{Y, X\}$ with binary outcome $Y \in \{0, 1\}$ and continuous covariate $X \in \mathbb{R}$. Here and throughout, we follow the convention that uppercase letters denote random variables, while lowercase letters represent their realized values. Depending on the empirical context, one may be interested in the conditional probability

$$\pi(x) = \mathbb{P}\left(Y = 1 | X = x\right), \tag{1}$$

the partial effect $\partial \pi(x)/\partial x$, and the elasticity, when x takes extreme values. Given a random sample with N observations, we characterize the extremeness by letting $x \to \infty$ as $N \to \infty$.[1,2] In many cases, $\pi(x) \to 0$ or 1 as $x \to \infty$, so we combine both scenarios and define a unified measure of extreme elasticity as

$$\delta(x) = \frac{\partial \left(\pi(x)\left(1 - \pi(x)\right)\right)}{\partial x} \frac{x}{\pi(x)\left(1 - \pi(x)\right)}. \tag{2}$$

Further details are provided in Proposition 2.1 and Remark 2.3. We will use sovereign debt and country default as a running example for intuitive explanations. In this context, the outcome Y indicates whether the country defaults or not, the covariate X is the debt-to-GDP ratio, and the conditional probability $\pi(x)$ represents the probability of default when the debt-to-GDP ratio is particularly high, which can be viewed as a counterfactual probability or a predictive probability.

---

[1]Here we focus on the right tail without loss of generality. In practice, one can conduct a similar analysis for the left tail, and we allow for different tail thickness in each tail.

[2]In finite sample, our approach still shows significant improvement for moderately large x, roughly below the 20th or above 80th percentile of the X distribution, and the corresponding $\pi(x)$ may not be very close to 0 or 1 either: see Figure 3, for example.

Existing methods face limitations when dealing with extreme values, a common feature in economic data that manifests as heavy tails (Gabaix, 2009, 2016). Parametric methods, such as Logit and Probit, assume threshold-crossing models with thin-tailed error distributions, which can lead to significant misspecification bias, particularly in the tails. Conversely, non-parametric methods, such as kernel, sieve, and spline estimators, may encounter difficulties due to limited information in the tail, resulting in highly inefficient estimators with large variances. This inefficiency can lead to imprecise forecasts, especially for extreme values.

To address these challenges, we propose a novel semiparametric approach based on Bayes' theorem and RV functions, which offer a flexible framework that encompasses a wide range of distributions. For heavy-tailed distributions, the RV condition naturally leads to a Pareto approximation in the tail (see Section 2.1 for details), while making no parametric assumptions on the relationship between covariates and outcomes beyond the tail region.[3] This is crucial because patterns in the tail and middle can differ substantially, as illustrated in the log-log plots in Figures 1 and 5. Specifically, by Bayes' rule, we have that

$$\mathbb{P}\left(Y=1|X=x\right)=\frac{f_{X|Y}\left(x|1\right)\mathbb{P}\left(Y=1\right)}{f_{X|Y}\left(x|1\right)\mathbb{P}\left(Y=1\right)+f_{X|Y}\left(x|0\right)\mathbb{P}\left(Y=0\right)}, \tag{3}$$

where $f_{X|Y}\left(\cdot|y\right)$ denotes the conditional density function of X given $Y=y$. When x takes extreme values, the RV properties allow us to approximate $f_{X|Y}\left(\cdot|y\right)$ by $x^{-\alpha^{(y)}-1}$ up to some constant, where $\alpha^{(y)}$ is the Pareto exponent for $y\in\{0,1\}$. Then, the extreme elasticity $\delta(x)$ converges to $-\left|\alpha^{(1)}-\alpha^{(0)}\right|$, as $x\to\infty$. We establish formal asymptotic results in a more general setting with additional non-extreme covariates $Z=z$, where we hold z fixed and let $x\to\infty$. Note that given the limited data available in the tail for extreme x values, the convergence rate of the tail estimator is slower than the standard $\sqrt{N}$-rate.

Furthermore, our method not only addresses the limited information in the tail but also effectively handles unobserved unit-specific tail parameters and unit-specific RV functions in panel data. The heterogeneity in the RV functions naturally accommodates unit-specific scale parameters as well. We show that, under regularity conditions, our objective function is asymptotically equivalent to that of a panel Logit regression on tail observations, using the log extreme covariate as a regressor. This equivalence simplifies the estimation procedure

---

[3]To see the semiparametric nature of our tail estimator, note that traditional kernel methods focus on local observations close to a specific x value of interest, leaving the model far from x unspecified. Similarly, our tail estimator focuses on a specific local region, namely the extreme tail where $x\to\infty$, leaving the model for the middle part unspecified.

and makes the method more convenient for empirical studies.

Here we consider panel data with large N and either small or large T. For small T, we can use the conditional MLE to eliminate unit-specific tail effects. The proof nontrivially extends Wang and Tsai (2009) to panel data with independent but not necessarily identically distributed (i.n.i.d.) observations. For large T, we can directly estimate the unit-specific parameters using existing methods for large-T panel Logit and further provide unit-specific forecasts. Finally, we expand our discussion to dynamic panel data models, incorporating lagged outcomes into the analysis.

We assess the finite-sample performance of our tail estimator through Monte Carlo simulations under various specifications. Experiment 1 focuses on estimation accuracy in cross-sectional data, and Experiment 2 examines out-of-sample forecasts in panel data with large N and large T. Results show that our tail estimator outperforms parametric and nonparametric alternatives, particularly in capturing heavy-tailed behavior, reducing misspecification bias, and providing more accurate forecasts.

In our empirical application, we analyze the impact of local housing price declines on the riskiness of small banks, which played an important role in the 2007–2008 financial crisis. We construct a panel dataset of loan charge-off rates of small banks, local housing prices, and unemployment rates, and again find that the tail estimator yields the most accurate pseudo out-of-sample forecasts among all methods considered. In addition, this empirical analysis also helps reveal interesting heterogeneity patterns in riskiness among small banks, varying across geographic regions and bank characteristics.

Related literature. Our work draws on a wide range of econometric literature, including binary outcome models, panel data, extreme value theory, and forecasting.

First, our work builds on a large literature on the binary outcome models, especially those for panel data. However, we focus on the analysis of extreme events instead of mid-sample properties as in the existing literature.

In binary cross-sectional models, various methods have been well studied in the literature, including parametric approaches (e.g., Logit and Probit), nonparametric approaches (e.g., Matzkin (1992)), and semiparametric approaches, such as the single-index model (Klein and Spady, 1993) and the maximum score estimator (Manski, 1985; Horowitz, 1992). Also see Horowitz and Savin (2001) for a review. Our approach falls within the semiparametric framework, approximating the heavy tails using RV functions while making no parametric

assumptions in the middle. Notably, much of this existing literature assumes threshold-crossing models, whereas our approach does not require such a structure: see Section 2.1.1 for a detailed comparison.

In contrast to common methods that directly work with $\mathbb{P}(Y = 1|X = x)$, we employ Bayes' theorem to "reverse" the conditioning set. Similar transformations based on Bayes' theorem have also been applied in the discriminant analysis (Amemiya, 1985) with Gaussian $X|Y = y$, as well as in the semiparametric single-index model in Klein and Spady (1993). In our context of extreme events, this transformation offers a novel way to disentangle tail behavior. Unlike those previous works focusing on mid-sample properties, we analyze tail properties with only tail observations, necessitating the development of nonstandard asymptotic theory.

In binary panel data models, the unobserved unit-specific heterogeneity can be challenging to handle due to the incidental parameter problem in the presence of nonlinear model structures. In addition to parametric, nonparametric, and semiparametric approaches, the binary panel literature can also be categorized by assumptions on unit-specific heterogeneity. See Chapter 15 in Wooldridge (2010) and Chapter 7 in Hsiao (2022) for textbook discussions. In a random effects setup, the unit-specific heterogeneity is assumed to be drawn from a common underlying distribution. In a correlated random effects setup, this distribution could depend on observed covariates. In a fixed effects setup, the unit-specific heterogeneity is treated as fixed parameters unique to each unit without imposing any distributional assumptions, implicitly allowing for an arbitrary correlation between the heterogeneity and covariates. For logistic errors, Chamberlain (1980), among others, employs the conditional MLE to eliminate the unit-specific intercept and estimate the common parameters; for general errors, Manski (1987) extends the maximum score estimator to panel data.

Since we allow for heterogeneity in the RV functions without explicitly modeling it, our approach aligns with a fixed effects framework, where unit-specific tail thickness and RV functions can flexibly depend on the additional covariates Z. We show that our objective function is asymptotically equivalent to a panel Logit regression on tail observations with the log extreme covariate as a regressor. This equivalence facilitates a comparable conditional MLE analysis to estimate common parameters in small-T panels. For large-T panels, bias correction methods, such as those developed by Fernández-Val and Weidner (2018) and Stammann, Heiss, and McFadden (2016), can be utilized to further estimate unit-specific parameters. Additionally, we extend our approach to dynamic binary panel data models,

where the conditional MLE construction is related to the work of Honoré and Kyriazidou (2000).

Second, our work contributes to the heavy tail and extreme value theory literature. For a comprehensive review, please refer to de Haan and Ferreira (2006) and Gabaix (2009, 2016). The literature typically focuses on continuous outcomes with heavy tails, such as the classic Hill (1975) estimator, as well as estimators proposed by Smith (1987) and Gabaix and Ibragimov (2011). See recent reviews by Gomes and Guillou (2015) and Fedotenkov (2020) for over a hundred estimators. Wang and Tsai (2009) incorporate covariates and develop a tail index regression model, and Nicolau, Rodrigues, and Stoykov (2023) further extend this framework to time-series data under strongly mixing conditions. In contrast to the existing literature, we focus on binary outcomes and utilize Bayes' theorem to construct a novel estimator. To incorporate covariates, we build upon the tail index regression model and significantly extend the theoretical framework to handle panel data and i.n.i.d. random variables.

Third, our work further contributes to the recent literature on unit-specific forecasts in panel data setups. To the best of our knowledge, this paper is the first investigation into unit-specific forecasts in panel data settings with binary outcomes and extreme covariates. In linear models, Liu, Moon, and Schorfheide (2020) and Liu (2023) develop empirical Bayes and full Bayesian approaches for point and density forecasts, respectively. In nonlinear models, Christensen, Moon, and Schorfheide (2020) propose efficient robust forecasts for discrete outcome models, and Liu, Moon, and Schorfheide (2023) construct set and density forecasts for Tobit models with censored outcomes. Also see Baltagi (2013) for the best linear unbiased predictor, Giacomini, Lee, and Sarpietro (2023) for robust forecasts, and Qu, Timmermann, and Zhu (2023) for forecasting comparison. Most existing methods focus on continuous outcomes (except for Christensen, Moon, and Schorfheide (2020)) and mid-sample properties, and thus are not suitable for our purpose. Also note that panel data are particularly valuable for analyses involving extreme events, as the limited tail information makes it even more beneficial to combine information across cross-sectional units.

Finally, our work is also related to the growing interest in studying extreme events and their economic consequences, partly motivated by the 2007–2008 financial crisis and the Covid-19 pandemic. Existing literature addresses extreme events in various ways: some exclude them (e.g., Schorfheide and Song (2024)), some adapt models to accommodate both extreme and ordinary observations within a unified framework (e.g., Carriero, Clark, Mar-

cellino, and Mertens (2022), and Lenza and Primiceri (2022)), and some employ quantile regressions (e.g., Adrian and Brunnermeier (2016), Plagborg-Møller, Reichlin, Ricco, and Hasenzagl (2020), and Adrian, Grinberg, Liang, Malik, and Yu (2022)). Our work differs in several ways. First, their approaches are specific for continuous outcomes, not for binary ones. Second, our approach focuses on cross-sectional and panel data rather than time-series data, allowing us to exploit information across units.[4] Third, we incorporate a semiparametric approach that models extreme events separately from the middle of the data, acknowledging potentially distinct patterns in normal and extreme environments. While the value-at-risk and quantile regression papers also consider the second and third points, our approach is specifically designed for binary outcomes, with a different model structure and estimation procedure.

The remainder of this paper is organized as follows. Sections 2 and 3 specify our methodology for cross-sectional and panel data, respectively, and derive their asymptotic properties. Section 4 extends our estimator to various contexts, such as a dynamic panel data model. Section 5 conducts Monte Carlo experiments to examine the finite-sample properties of our estimators. Section 6 employs our panel data estimator to analyze how local housing price declines affected the riskiness of small banks during the 2007–2008 financial crisis. Finally, Section 7 concludes. Appendix A provides the proofs for all propositions and theorems, and Appendix B contains additional tables and figures.

2   Cross-sectional data

In this section, we continue the discussion from the introduction for cross-sectional data. Section 2.1 focuses on intuition and illustrates our main idea without additional covariates. It also compares our method with the classic threshold-crossing model. Section 2.2 details the estimator and establishes its asymptotic properties in a general setup with additional covariates.

---

[4]We also consider dynamic panels under conditional stationarity, and our framework could be extended to accommodate strongly mixing conditions in the time-series dynamics as in Nicolau, Rodrigues, and Stoykov (2023), though this extension is beyond the scope of this paper.

2.1 Baseline models and RV functions

Recall that in the introduction, we presented a simple cross-sectional model with binary outcome Y and continuous covariate X. Without loss of generality, let $X \in \mathbb{R}^+$, and $x \to \infty$ as $N \to \infty$.[5] We focus on three potential objects of interest: the conditional probability $\pi(x)$, the partial effect $\partial \pi(x)/\partial x$, and the extreme elasticity $\delta(x)$, as defined in (1) and (2).

Heavy-tailed distributions are well-documented in economic and financial data, e.g., Gabaix (2009, 2016), and the literature has suggested various methods to assess their presence. First, a log-log plot provides a visual assessment of tail behavior by plotting the threshold x against the probability of exceeding x, both on a log scale, highlighting the relative prevalence of large values in the data. If tail observations align around a downward-sloping line, this suggests a heavy tail pattern with the slope being approximately $-\alpha$, where α is the Pareto exponent or tail index. In contrast, a vertical alignment indicates a relatively thin tail. See Figures 1 and 5 for examples with simulated and empirical data. Second, a more rigorous evaluation involves estimating the tail index and conducting statistical tests: see for example, Clauset, Shalizi, and Newman (2009). Finally, for a data-driven evaluation, techniques like cross-validation and pseudo out-of-sample forecasting can be employed as well, as demonstrated in our Monte Carlo Experiment 2 and empirical example.

Existing methods encounter difficulties in handling extreme values. Parametric approaches, such as Logit and probit, may incur substantial misspecification bias in the tail, while nonparametric methods, such as kernel, sieve, and spline, may be highly inefficient due to a limited number of tail observations. To overcome these challenges, we propose a semiparametric approach based on Bayes' theorem and RV functions as follows.

First, we observe extreme values in the covariate X and seek to understand the behavior of Y under extreme X, so we essentially aim to analyze their comovement in the tail. To facilitate this tail analysis, we use Bayes' theorem to "reverse" the conditioning set in the conditional probability (1), resulting in equation (3). Note that the Bayes' theorem representation is simply an alternative characterization of the data, so we are agnostic about the causal direction between X and Y. A similar Bayes' theorem representation has been employed in the discriminant analysis (Chapter 9.2.8 in Amemiya (1985)), where normal dis-

---

[5]Directly incorporating multidimensional extreme covariates poses theoretical challenges. However, a practical solution is to replace X with $v(X; \gamma)$, where $v(\cdot; \gamma)$ is a known function with unknown finite-dimensional parameters γ, and γ can be consistently estimated with sufficiently fast convergence rate. For simplicity, we will treat X as a scalar in the rest of the paper, but most of our discussions can be extended to the case of $v(X; \gamma)$.

tributions of $X|Y = y$ leads to a quadratic Logit form of $\mathbb{P}(Y = 1|X = x)$. Klein and Spady (1993) also use a similar Bayes' theorem transformation in their semiparametric single-index estimator.

Next, as $x \to \infty$, the conditional pdfs $f_{X|Y}(x|y)$ for $y \in \{0, 1\}$ are dominated by their tail behaviors. If $X|Y = y$ exhibits a heavy tail, its distribution can be well approximated by a Pareto distribution (Smith, 1987). Therefore, we adopt a more general concept of RV functions that leads to a Pareto approximation in the tail.[6]

We first introduce the following definitions. Let $X \in \mathbb{R}^+$ be a generic random variable with a heavy-tailed distribution. We say the upper tail probability (or survival function) $1 - F_X(x)$ is *Regularly Varying (RV)* at infinity with index $-\alpha$ for some $\alpha > 0$, if for all $x > 0$, as $\eta \to \infty$,

$$\frac{1 - F_X(\eta x)}{1 - F_X(\eta)} \to x^{-\alpha}, \tag{4}$$

which is denoted as $1 - F_X \in RV_{-\alpha}$. Equivalently, by Karamata's characterization theorem, we can write

$$1 - F_X(x) = x^{-\alpha} \mathcal{L}(x),$$

where $\mathcal{L}(\cdot)$ is a *slowly varying function* such that for all $x > 0$, $\frac{\mathcal{L}(\eta x)}{\mathcal{L}(\eta)} \to 1$, as $\eta \to \infty$.

Let us highlight a few key points. First, the RV condition (4) implies that as $\underline{x} \to \infty$, for $x \geq \underline{x}$,

$$1 - F_X(x) = (1 - F_X(\underline{x})) \left(\frac{x}{\underline{x}}\right)^{-\alpha} (1 + o(1)). \tag{5}$$

Thus, the tail distribution of X is asymptotically proportional to a Pareto distribution as the first-order approximation. Second, the parameter α is also referred to as the *Pareto exponent*, which characterizes the tail thickness of $1 - F_X(x)$. In particular, a smaller α indicates a heavier tail, meaning that the upper tail probability decays to zero more slowly; as $\alpha \to \infty$, the distribution becomes thin-tailed, such as logistic and normal distributions. Third, the RV condition is relatively mild and satisfied by many commonly used heavy-tailed distributions, including the Student-t, F, and Cauchy distributions. For example, in a Student-t distribution, α is equal to the degrees of freedom. A list of distributions and their corresponding values of α can be found in Gabaix (2016). Fourth, the moments of X, $\mathbb{E}[|X|^r]$, are finite

---

[6]Our method can be adapted to the generalized Pareto distribution, which also encompasses thin-tailed and bounded-support distributions. However, these distributions are often easily distinguishable from heavy-tailed distributions given empirical data. Moreover, the simplicity of our Pareto-based estimator offers a practical advantage over the more involved generalized Pareto estimator.

up to order $r < \alpha$. Note that we can handle cases with very heavy tails where $\alpha \in (0,1)$ and no moments exist. Finally, following from Proposition B.1.9(11) in de Haan and Ferreira (2006), if the pdf $f_X(x)$ is non-increasing for sufficiently large x, we have the following approximation that as $x \to \infty$,

$$\frac{x f_X(x)}{1 - F_X(x)} \to \alpha. \tag{6}$$

Back to our binary outcome model, we assume that the distribution of X conditional on $Y = y$ satisfies the RV condition (4), that is, for $y \in \{0,1\}$, $1 - F_{X|Y}(\cdot|y) \in RV_{-\alpha^{(y)}}$. The Pareto exponent $\alpha^{(y)}$ is indexed by y, allowing for different tail thickness for $y \in \{0,1\}$. If Y were continuous, estimating $\alpha^{(y)}$ as a function of y would be challenging without parametric assumptions. The binary outcome simplifies this, as $\alpha^{(y)}$ takes only two values, $\alpha^{(0)}$ and $\alpha^{(1)}$, which can be easily estimated using methods such as the classic Hill (1975) estimator in this simple model. Further estimation details are provided in Section 2.2. Also note that our approach is semiparametric, as we impose no parametric assumptions on the relationship between X and Y outside the tail region.

After estimating $\alpha^{(y)}$, we can proceed with the conditional probability $\pi(x)$ as well as other objects of interest. Based on Bayes' theorem (3) and the pdf approximation (6), it follows that

$$\pi(x) \sim \frac{1}{1 + \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{1 - F_{X|Y}(x|0)}{1 - F_{X|Y}(x|1)}}, \tag{7}$$

where $f(x) \sim g(x)$ denotes $\lim_{x \to \infty} f(x)/g(x) = 1$ for generic functions $f(x)$ and $g(x)$. By the RV condition (5), given some large $\underline{x}^{(y)}$, the term in the denominator becomes

$$\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{1 - F_{X|Y}(x|0)}{1 - F_{X|Y}(x|1)} \sim \frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{\left(\underline{x}^{(0)}\right)^{\alpha^{(0)}}}{\left(\underline{x}^{(1)}\right)^{\alpha^{(1)}}} x^{\alpha^{(1)} - \alpha^{(0)}}. \tag{8}$$

Then, $\pi(x)$ can be consistently estimated by plugging in $\hat{\alpha}^{(y)}$ and the sample analog of $\frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \approx \frac{N^{(0)}}{N^{(1)}}$, where $N^{(y)} = \sum_{i=1}^{N} \mathbf{1}\left\{X_i^{(y)} \geq \underline{x}^{(y)}, Y_i = y\right\}$ is the number of tail observations in the subsample with $Y_i = y$.[7]

---

[7]The consistency result was established in a previous version of this paper and is available upon request.

Remark 2.1 Based on (7) and (8), we define $A = \frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{\left(\underline{x}^{(0)}\right)^{\alpha^{(0)}}}{\left(\underline{x}^{(1)}\right)^{\alpha^{(1)}}}$, a constant not changing over x. Then, $\pi(x)$ is simplified to:

$$\pi(x) \sim \frac{1}{1 + A \cdot x^{\alpha^{(1)} - \alpha^{(0)}}}, \tag{9}$$

and we can directly estimate A and $\alpha^* = \alpha^{(1)} - \alpha^{(0)}$ via the MLE. However, for the cross-sectional case, we slightly prefer the estimator based on (7) and (8) due to its easier implementation in the simple setup; moreover, when incorporating additional covariates Z (see Section 2.2), the parameter A could be a complicated function of Z and difficult to estimate.

Furthermore, letting $\tilde{A} = -\log A$ yields $\pi(x) \sim \frac{1}{1 + \exp\left(-\tilde{A} + \alpha^* \log x\right)}$, so our objective function asymptotically resembles that of a Logit regression on tail observations, using the log extreme covariate as the regressor. This asymptotic equivalence is particularly relevant in panel data cases with unobserved unit-specific heterogeneity: see Remark 3.1.

Remark 2.2 From (9), we see that between the two subsamples with $Y_i = y \in \{0, 1\}$, the one with the heavier tail (i.e., smaller Pareto exponent $\alpha^{(y)}$) will ultimately dominate the conditional probability, that is, as $x \to \infty$,

$$\pi(x) \sim \frac{1}{1 + A \cdot x^{\alpha^{(1)} - \alpha^{(0)}}} \to \begin{cases} 1, & \text{if } \alpha^{(1)} < \alpha^{(0)}, \\ 0, & \text{if } \alpha^{(1)} > \alpha^{(0)}, \\ \frac{1}{1+A}, & \text{if } \alpha^{(1)} = \alpha^{(0)}. \end{cases}$$

This Pareto approximation captures the first-order behavior, and higher-order terms, as detailed in Assumption 2.2, can further refine $\pi(x)$. Also note that, in finite samples, our method still provides significant improvements for moderately large x with $\pi(x)$ not very close to 0 or 1, as discussed in footnote 2.

After estimating $\pi(x)$, the partial effect $\partial \pi(x)/\partial x$ can be obtained as a by-product through either analytical or numerical differentiation. It may also be of interest to consider averages of the partial effects over certain covariate values, which relate to commonly used measures, such as the average partial effects (APE) and the average marginal effects (AME) in panel data models: see Abrevaya and Hsu (2021) for a survey of various partial effects in panels, and Davezies, D'Haultfoeuille, and Laage (2024) for recent partial identification results of average causal effects in short-T panel Logit models with fixed effects. Proposition

A.1 in the Appendix establishes the existence of the tail average of partial effects.

For extreme elasticity (2), given the Pareto approximation in the tail, the extreme elasticity is solely determined by the difference in tail indices, $\alpha^* = \alpha^{(1)} - \alpha^{(0)}$.

Proposition 2.1 (Cross-sectional data: extreme elasticity) *Suppose we have: (a)* $1 - F_{X|Y}(\cdot|y) \in RV_{-\alpha^{(y)}}$, *for* $y \in \{0,1\}$; *(b)* $f_{X|Y}(x|y)$ *and* $f'_{X|Y}(x|y)$ *are non-increasing in* $x \geq \underline{x}$, *for some* $\underline{x} > 0$; *and (c)* $0 < \mathbb{P}(Y = 1) < 1$. *Then, as* $x \to \infty$, $\delta(x) \to -|\alpha^*|$.

Remark 2.3 From Remark 2.2, we see that $\pi(x)$ approaches either 0 or 1 as $x \to \infty$ when $\alpha^{(1)} \neq \alpha^{(0)}$. The extreme elasticity defined in (2) provides a unified characterization of both cases. Rewriting (2), we obtain that

$$\delta(x) = \frac{\partial \pi(x)}{\partial x} \frac{x}{\pi(x)} + \frac{\partial (1 - \pi(x))}{\partial x} \frac{x}{1 - \pi(x)} = \delta_\pi(x) + \delta_{1-\pi}(x). \tag{10}$$

For example, when $\alpha^{(1)} < \alpha^{(0)}$, we have $\pi(x) \to 1$, $\delta_\pi(x) \to 0$, and thus $\delta(x) \sim \delta_{1-\pi}(x) \to \alpha^{(1)} - \alpha^{(0)}$. Note that the extreme elasticity is inherently non-positive due to the RV structure. Moreover, as $\pi(x)(1 - \pi(x)) = \mathbb{V}(Y|X = x)$, the extreme elasticity can also be interpreted as the elasticity of risk in the tail.

2.1.1 Comparison with a threshold-crossing model

We now compare our approach with the classic threshold-crossing model, which has been extensively studied in the literature:

$$Y = \mathbf{1}\{X - \varepsilon \geq 0\},$$

where ε is the error term. Note that our method does not require a threshold-crossing structure. However, the following proposition reveals a direct relationship between our key parameters $\alpha^{(y)}$ and the Pareto exponents of X and ε in the threshold-crossing model.

Proposition 2.2 (Threshold-crossing model) *Suppose we have: (a)* $1 - F_X \in RV_{-\alpha_X}$, *and* $1 - F_\varepsilon \in RV_{-\alpha_\varepsilon}$, *for some* $\alpha_X, \alpha_\varepsilon > 0$; *(b)* $f_X(x)$ *is non-increasing in* $x \geq \underline{x}$, *for some* $\underline{x} > 0$; *and (c)* $\varepsilon \perp X$. *Then, for* $y \in \{0,1\}$, $1 - F_{X|Y}(\cdot|y) \in RV_{-\alpha^{(y)}}$ *with*

$$\alpha^{(0)} = \alpha_X + \alpha_\varepsilon, \text{ and } \alpha^{(1)} = \alpha_X.$$

Furthermore, $\alpha_\varepsilon = \alpha^{(0)} - \alpha^{(1)}$.

There are three points worth noting. First, the threshold-crossing structure and the *unconditional* RV tails provide a sufficient condition for our *conditional* RV tails. The comovement between X and ε in the tail is captured by the comovement between X and Y in the tail.

Second, the difference in the conditional Pareto exponents, $\alpha^{(0)} - \alpha^{(1)}$, equals the Pareto exponent of the unobserved error term. As the tail of the error term becomes thinner with larger α_ε, this difference increases. The extreme case of a very thin-tailed error with $\alpha_\varepsilon \to \infty$ (e.g., close to the Logit) corresponds to $\alpha^{(0)} \to \infty$, $\alpha^{(1)} = \alpha_X$, and $\pi(x) \to 1$: see also the first case in Remark 2.2.

Finally, directly estimating the threshold-crossing model with RV errors ε could be challenging due to two related issues: first, ε is not observable, making it difficult to determine which observations are in the tail; second, using the full data could lead to significant bias, as the middle part may deviate substantially from the Pareto distribution. In contrast, our method circumvents this issue by using cutoffs of observed X, and offers an asymptotic equivalent estimator that is relatively easy to implement and scalable to more complicated setups, such as those involving additional covariates and panel data.

2.2 Estimation and asymptotic properties

In this subsection, we first extend the simple cross-sectional model in Section 2.1 to incorporate additional covariates, which is more empirically relevant. We then derive the asymptotic properties of our proposed estimator.

Let Z be a d_Z-dimensional vector of non-extreme covariates in addition to X. Z can be either discrete or continuous. For example, consider the context of sovereign debts and country defaults, where institutional characteristics could be included as covariates given their potential impact on default risk. Now we rewrite our conditional probability and the Bayes' theorem representation by adding Z to all conditioning sets:

$$\pi(x, z) = \mathbb{P}(Y = 1 | X = x, Z = z)$$
$$= \frac{f_{X|Y,Z}(x|1, z)\, \mathbb{P}(Y = 1 | Z = z)}{f_{X|Y,Z}(x|1, z)\, \mathbb{P}(Y = 1 | Z = z) + f_{X|Y,Z}(x|0, z)\, \mathbb{P}(Y = 0 | Z = z)}.$$

Note that here we hold z constant and let x tend to infinity. Accordingly, we assume that $1 - F_{X|Y,Z}(\cdot|y, z) \in RV_{-\alpha^{(y)}(z)}$, where the Pareto exponent $\alpha^{(y)}(z)$ is a function of the

additional covariates z.

While $\alpha^{(y)}(\cdot)$ could depend on z in a complicated nonlinear way, it is difficult to non-parametrically estimate it due to the data limitation that only large values of X are used for tail analysis. To sidestep this issue, we assume a pseudo-linear structure:

$$\alpha^{(y)}\left(z; \theta^{(y)}\right) = z'\theta^{(y)}, \tag{11}$$

where $\theta^{(y)}$ denotes the pseudo-parameters for $y \in \{0, 1\}$, and $\alpha^{(y)}(z) = \alpha^{(y)}\left(z; \theta_0^{(y)}\right)$ represents the tail index function evaluated at the pseudo-true parameter values. This simplification is also in line with the literature, such as Wang and Tsai (2009).[8] Additionally, we require that $Z_i'\theta^{(y)} > 0$ almost surely, with sufficient conditions discussed after Assumption 2.2.

From the RV condition, for $y \in \{0, 1\}$, given some large threshold $\underline{x}_N^{(y)}$, if the conditional pdf is non-increasing for sufficiently large x, it is asymptotically equivalent to a Pareto distribution,

$$f_{X|Y,Z}\left(x \,\Big|\, y, z, x \geq \underline{x}_N^{(y)}\right) \sim \alpha^{(y)}(z) \left(\frac{x}{\underline{x}_N^{(y)}}\right)^{-\alpha^{(y)}(z)-1}.$$

Let $\Xi_{N,i}^{(y)} = \left\{X_i^{(y)} \geq \underline{x}_N^{(y)}, Y_i = y\right\}$, then $N^{(y)} = \sum_{i=1}^N \mathbf{1}\left\{\Xi_{N,i}^{(y)}\right\}$ is the total number of tail observations in the subsample with $Y_i = y$. The asymptotic Pareto distribution above leads to the following pseudo-MLE

$$\hat{\theta}^{(y)} = \arg\max_{\theta^{(y)} \in \Theta^{(y)}} \sum_{i=1}^N \left(\log Z_i'\theta^{(y)} - Z_i'\theta^{(y)} \log \frac{X_i}{\underline{x}_N^{(y)}}\right) \mathbf{1}\left\{\Xi_{N,i}^{(y)}\right\}, \tag{12}$$

where $\Theta^{(y)} \subset \mathbb{R}^{d_Z}$ is a convex cone as discussed after Assumption 2.2, and the pseudo-true parameter value $\theta_0^{(y)} \in \text{int}\left(\Theta^{(y)}\right)$, the interior of $\Theta^{(y)}$. Since the objective function is concave and the domain is convex, this MLE is a convex programming problem that can be easily solved. In the simple model without Z, the FOC of (12) results in $\hat{\alpha}^{(y)} = \left[\frac{1}{N^{(y)}} \sum_{i=1}^N \left(\log X_i - \log \underline{x}_N^{(y)}\right) \mathbf{1}\left\{\Xi_{N,i}^{(y)}\right\}\right]^{-1}$, which coincides with the classic Hill (1975) estimator, as the indicator function $\mathbf{1}\left\{\Xi_{N,i}^{(y)}\right\}$ effectively selects the largest order statistics.

To establish the asymptotic properties, we impose the following assumptions.

---

[8]We adopt a pseudo-linear approximation to better align with the panel data case in Section 3, rather than an exponential approximation in Wang and Tsai (2009).

Assumption 2.1 (Cross-sectional data: model assumptions) *Suppose we have:*

(a) $\{Y_i, X_i, Z_i\}$ *is i.i.d.*

(b) $0 < \mathbb{P}\left(Y_i = 1 | Z_i\right) < 1$ *almost surely in* Z_i.

Condition (b) ensures non-degenerate probabilities, thus guaranteeing a non-trivial Bayes' theorem representation.

Assumption 2.2 (Cross-sectional data: tail approximation) *Let* $\mathbb{P}_Z$ *be the probability measure induced by* Z. *For* $y \in \{0,1\}$, *and for* $\mathbb{P}_Z$-*almost all* $z \in supp(Z)$:

(a) *The conditional cdf* $F_{X|Y,Z}\left(x|y,z\right)$ *satisfies that*

$$1 - F_{X|Y,Z}\left(x|y,z\right) = C^{(y)}\left(z\right) x^{-\alpha^{(y)}(z)} \left(1 + D^{(y)}\left(z\right) x^{-\beta^{(y)}(z)} + r^{(y)}\left(x,z\right)\right),$$

as $x \to \infty$, *where* $C^{(y)}\left(z\right) > 0$, $\left|D^{(y)}\left(z\right)\right| \leq \overline{D} < \infty$, $\alpha^{(y)}\left(z\right) > 0$ *satisfying* (11), *and* $\beta^{(y)}\left(z\right) \geq \underline{\beta} > 0$, *for some constants* $\overline{D}$ *and* $\underline{\beta}$.

(b) *The remainder term satisfies that* $\left|x^{\beta^{(y)}(z)} r^{(y)}\left(x,z\right)\right| \leq \bar{r}(x)$, *where* $\bar{r}(x) \to 0$ *as* $x \to \infty$.

(c) *The conditional pdf* $f_{X|Y,Z}\left(x|y,z\right)$ *is non-increasing in* $x \geq \underline{x}$, *for some* $\underline{x} > 0$.

First, the expression in condition (a) implies the RV condition and resembles equation (2.2) in Wang and Tsai (2009). In the absence of additional covariates Z, it simplifies to the well-studied second-order condition: see for example, Hall (1982) and Chapter 2 in de Haan and Ferreira (2006). The first-order term $C^{(y)}\left(z\right) x^{-\alpha^{(y)}(z)}$ is proportional to a Pareto distribution, while the second-order term $D^{(y)}\left(z\right) x^{-\beta^{(y)}(z)}$ captures deviations from the ideal Pareto form and is essential for characterizing the asymptotic distribution. Second, to guarantee the existence of $\theta^{(y)}$ such that $\alpha^{(y)}(Z_i) = Z_i'\theta^{(y)} > 0$ almost surely (i.e., the existence of the convex cone), a sufficient condition is that each element of Z_i has a domain strictly above or below zero. This can be achieved through a monotonic transformation of Z_i, such as the exponential or probability integral transforms. For example, if $Z_i \in \mathbb{R}_{++}^{d_Z}$ (the set of vectors with strictly positive elements) almost surely, then we can set $\Theta^{(y)} = \mathbb{R}_+^{d_Z} \setminus \{0\}$ (the set of vectors with non-negative elements, excluding the zero vector). Third, the bounds on $D^{(y)}\left(z\right)$ and $\beta^{(y)}\left(z\right)$, as well as the supremum condition (b) on the remainder term, allow us to ignore higher order terms in the asymptotic analysis. Finally, the relatively mild condition

(c) of a non-increasing conditional probability density function ensures that the tail behavior of the pdf can be inferred from the corresponding cdf, according to Proposition B.1.9(11) in de Haan and Ferreira (2006).

To derive the asymptotic properties, note that $N^{(y)}$ is a random variable, so we need to carefully account for its randomness when applying the LLN and CLT. We further define

$$\xi_N^{(y)} = \mathbb{E}\left[C^{(y)}(Z_i) \left(\underline{x}_N^{(y)} \right)^{-\alpha^{(y)}(Z_i)} \right] \mathbb{P}(Y_i = y), \tag{13}$$

a non-random sequence representing the asymptotic proportion of tail observations for $Y_i = y$, and utilize it to characterize the asymptotic behavior of our estimator.

Assumption 2.3 (Cross-sectional data: estimation) *For $y \in \{0, 1\}$, suppose that:*

(a) $N\xi_N^{(y)} \to \infty$, *as $N \to \infty$.*

(b) $\sqrt{\frac{N}{\xi_N}}\mathbb{E}\left[Z_i \frac{\beta^{(y)}(Z_i)C^{(y)}(Z_i)D^{(y)}(Z_i)}{\alpha^{(y)}(Z_i)\left(\alpha^{(y)}(Z_i)+\beta^{(y)}(Z_i)\right)} \left(\underline{x}_N^{(y)} \right)^{-\alpha^{(y)}(Z_i)-\beta^{(y)}(Z_i)} \right] \to 0$, *as $N \to \infty$.*

(c) $H_{N0}^{(y)} = \mathbb{E}\left[\left. \frac{Z_i Z_i'}{\left(\alpha^{(y)}(Z_i)\right)^2} \right| \Xi_{N,i}^{(y)} \right]$ *is finite and positive definite, for sufficiently large N.*

Conditions (a) and (b) jointly impose upper and lower bounds on the rate at which the tuning parameter $\underline{x}_N^{(y)}$ tends to infinity. It implies that $N^{(y)}$ goes to infinity at a slower rate than N. Also note that in condition (b), we select a larger $\underline{x}_N^{(y)}$ to eliminate the asymptotic bias, albeit at the expense of a slower convergence rate. This is close in spirit to choosing an undersmoothing bandwidth in kernel regressions. Condition (c) is a mild regularity condition that ensures the invertibility of the Hessian matrix.

The following theorem establishes the asymptotic result, building on Wang and Tsai (2009).

Theorem 2.1 (Cross-sectional data: parameter estimation) *Suppose $\theta_0^{(y)} \in int\left(\Theta^{(y)}\right)$, where $\Theta^{(y)} \subset \mathbb{R}^{d_Z}$ is a convex cone. Suppose Assumptions 2.1–2.3 hold. Then, for $y \in \{0, 1\}$, we have that*

$$\sqrt{N\xi_N^{(y)}} \left(H_{N0}^{(y)} \right)^{1/2} \left(\hat{\theta}^{(y)} - \theta_0^{(y)} \right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{d_Z}\right),$$

as $N \to \infty$. In addition, $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(0)}$ are asymptotically independent.

The convergence rate is slower than the standard parametric $\sqrt{N}$-rate, since the tail estimator only accounts for observations in the tail and the number of tail observations

16

$N^{(y)} = N\xi_N^{(y)}(1 + o_p(1))$ increases at a slower rate than the total number of observations N. The convergence rate depends on the magnitude of the second-order term: larger $\beta^{(y)}(z)$ allows for a smaller cutoff $\underline{x}_N^{(y)}$ according to Assumption 2.3(b), and thus a larger $\xi_N^{(y)}$ based on equation (13), leading to a faster convergence rate. Especially, when $\beta^{(y)}(z) \to \infty$, the distribution approaches the ideal Pareto case, and the convergence rate is close to the $\sqrt{N}$-rate.

For implementation, we adopt a common practice to select the threshold $\underline{x}_N^{(y)}$: taking empirical quantiles (e.g., 90th and 95th percentiles) as potential thresholds, and checking them through graphic diagnostics like log-log plots. While there are various threshold estimation methods in the literature,[9] our simulations and empirical analysis suggest that parameter estimates are relatively robust within a range of threshold values. Intuitively, in a log-log plot, such as Figure 1, the threshold $\underline{x}_N^{(y)}$ acts as the starting point for slope estimation, and small variations in the threshold within the downward-sloping region would only minimally affect the estimated slope.

Given $\hat{\theta}^{(y)}$, we can obtain the conditional probability $\pi(x, z)$ similar to (7) and (8) in the simple model. It is more practical to estimate $\mathbb{P}\left(Y = y, X \geq \underline{x}_N^{(y)} | Z = z\right)$ parametrically, due to limited tail data. Analogous to Proposition 2.1, if we further assume that $f'_{X|Y,Z}(x|y,z)$ is non-increasing for sufficiently large x, then as $x \to \infty$, the extreme elasticity is now

$$\delta(x, z) = \frac{(\partial \pi(x,z)(1 - \pi(x,z)))}{\partial x} \frac{x}{\pi(x,z)(1 - \pi(x,z))} \to -\left|z'\left(\theta^{(1)} - \theta^{(0)}\right)\right|.$$

A consistent estimator is obtained by substituting $\hat{\theta}^{(y)}$ in place of $\theta^{(y)}$.

3 Panel data

Our method is particularly useful in panel data analysis, specifically in addressing unobserved individual heterogeneity. For instance, in the context of country defaults, different countries could have various cultural and historical backgrounds that might not be fully captured by observed data. In the analysis of extreme events, this unobserved heterogeneity could

---

[9]For example, Guillou and Hall (2001) choose the threshold by minimizing the asymptotic MSE, Clauset, Shalizi, and Newman (2009) select the threshold by maximizing the marginal likelihood or minimizing the distance between the power-law distribution and empirical data, and Wang and Tsai (2009) further consider a discrepancy measure for a tail index regression model with covariates.

manifest as unobserved unit-specific tail thickness and RV functions.[10]

This section focuses on panel data with large N and small T. To build intuition, we begin in Section 3.1 by considering a simple case without additional covariates Z_i. We then incorporate these covariates in Section 3.2 and derive the asymptotic for the common parameters and extreme elasticity. Later on, we will extend our discussions to models with large T in Section 4.1, time-varying additional covariates in Section 4.2, and dynamic panel data in Section 4.3.

3.1 Baseline model and conditional MLE

Suppose we observe a panel dataset $\{Y_{it}, X_{it}\}$ for $i = 1, ..., N$ and $t = 1, ..., T$. For illustrative purposes, let $T = 2$ (though our method is applicable to any $T \geq 2$), and X_{it} be a scalar extreme covariate.

As X_{it} approaches infinity, unobserved individual heterogeneity could reflect in unit-specific tail thickness and RV functions. Assume the unit-specific tail indices take the following additive form

$$\tilde{\alpha}_i^{(y)} = \alpha^{(y)} + \lambda_i, \tag{14}$$

where $\alpha^{(y)}$ represents the common component depending on $y \in \{0, 1\}$, and λ_i denotes the unit-specific component. For example, the tail indices of stock market returns exhibit heterogeneity across regions (Jondeau and Rockinger, 2003), as do Covid-19 cases and deaths across countries (Einmahl and He, 2023). The additive form helps cancel out the unobserved unit-specific tail thickness, as will soon be demonstrated.

Let $\mathbb{P}_i$ be the probability measure given unit-specific quantities. Specifically, $\mathbb{P}_i(\cdot) = \mathbb{P}\left(\cdot \; ; \lambda_i, \left\{\mathcal{L}_i^{(y)}\right\}\right)$, where λ_i is the unit-specific tail thickness as defined above, and $\mathcal{L}_i^{(y)}(\cdot)$ is the unit-specific slowly varying function as specified below. Any functions subscripted with i are implicitly conditioned on these unit-specific quantities. Note that we are working within a fixed effects framework where $\left\{\lambda_i, \left\{\mathcal{L}_i^{(y)}\right\}\right\}$ are considered as fixed for each unit i and can be arbitrarily correlated with the covariates of i, making the panel data analysis richer and more challenging than the cross-sectional case.

Based on a simplified version of Assumption 3.1 in Section 3.2 on the model setup, we assume that $\{Y_{it}, X_{it}\}$ are independent across units and stationary across time. The

---

[10]For panel data with observed heterogeneity only, a pooled estimator with observed covariates can be employed, similar to the cross-sectional case in Section 2.2.

stationarity condition enables us to eliminate unit-specific tail effects, and the conditional independence condition ensures that for $y_1, y_2 \in \{0, 1\}$,

$$\mathbb{P}_i\left(Y_{i1} = y_1, Y_{i2} = y_2 | X_{i1}, X_{i2}\right) = \mathbb{P}_i\left(Y_{i1} = y_1 | X_{i1}\right) \cdot \mathbb{P}_i\left(Y_{i2} = y_2 | X_{i2}\right). \tag{15}$$

We rewrite the conditional probability of unit i using Bayes' theorem similar to the cross-sectional case.

$$\begin{aligned}\pi_i(x) &= \mathbb{P}_i\left(Y_{it} = 1 | X_{it} = x\right) \\ &= \frac{f_{i,X_{it}|Y_{it}}(x|1)\mathbb{P}_i\left(Y_{it} = 1\right)}{f_{i,X_{it}|Y_{it}}(x|1)\mathbb{P}_i\left(Y_{it} = 1\right) + f_{i,X_{it}|Y_{it}}(x|0)\mathbb{P}_i\left(Y_{it} = 0\right)}.\end{aligned}$$

Here the subscript i in $\pi_i(x)$ indicates potential heterogeneity across i, and the absence of a t index reflects the stationarity condition.

To proceed, we again apply the Pareto approximation at the tail, assuming that $1 - F_{i,X_{it}|Y_{it}}(\cdot|y) \in RV_{-\tilde{\alpha}_i^{(y)}}$, or equivalently, $1 - F_{i,X_{it}|Y_{it}}(x|y) = x^{-\tilde{\alpha}_i^{(y)}}\mathcal{L}_i^{(y)}(x)$, for some slowly varying function $\mathcal{L}_i^{(y)}(x)$. Given the stationarity condition, $\mathcal{L}_i^{(y)}(x)$ does not change over time. From the RV condition and the pdf approximation in (6), we have that

$$\pi_i(x) \sim \frac{1}{1 + \frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}} \frac{1-F_{i,X_{it}|Y_{it}}(x|0)}{1-F_{i,X_{it}|Y_{it}}(x|1)}} = \frac{1}{1 + \frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}} \frac{\mathcal{L}_i^{(0)}(x)}{\mathcal{L}_i^{(1)}(x)} \frac{x^{-\tilde{\alpha}_i^{(0)}}}{x^{-\tilde{\alpha}_i^{(1)}}}}.$$

Let us look into each term one by one. First, note that $\frac{x^{-\tilde{\alpha}_i^{(0)}}}{x^{-\tilde{\alpha}_i^{(1)}}} = x^{\alpha^{(1)}-\alpha^{(0)}}$, since λ_i enters additively into the tail indices (14) and is thus canceled out in the ratio. Second, we have that $\frac{\mathcal{L}_i^{(0)}(x)}{\mathcal{L}_i^{(1)}(x)} \to \mathcal{L}_i^*$, which does not depend on x asymptotically due to the slowly varying nature of $\mathcal{L}_i^{(y)}(x)$. Third, the term $\frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}}$ could depend on $\left\{\lambda_i, \left\{\mathcal{L}_i^{(y)}\right\}\right\}$ in a complicated, possibly nonlinear manner, as they enter into the conditioning sets. Combining all three terms, let $A_i = \frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}}\mathcal{L}_i^*$, and the conditional probability admits the following asymptotic approximation

$$\pi_i(x) \sim \frac{1}{1 + A_i \cdot x^{\alpha^{(1)}-\alpha^{(0)}}}. \tag{16}$$

Without further assumptions, A_i cannot be consistently estimated in small-T panels due to the incidental parameter problem. However, we can eliminate A_i by conditioning on the

19

event $Y_{i1} + Y_{i2} = 1$. Following from the conditional independence (15), we obtain that as $x_1, x_2 \to \infty$,

$$\mathbb{P}_i \left(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 1, X_{i1} = x_1, X_{i2} = x_2 \right) \tag{17}$$

$$= \frac{1}{1 + \frac{\mathbb{P}_i(Y_{i1}=0|X_{i1}=x_1)\mathbb{P}_i(Y_{i2}=1|X_{i2}=x_2)}{\mathbb{P}_i(Y_{i1}=1|X_{i1}=x_1)\mathbb{P}_i(Y_{i2}=0|X_{i2}=x_2)}} = \frac{1}{1 + \frac{\pi_i(x_2)(1-\pi_i(x_1))}{\pi_i(x_1)(1-\pi_i(x_2))}} \sim \frac{1}{1 + \left(\frac{x_1}{x_2}\right)^{\alpha^{(1)}-\alpha^{(0)}}}.$$

The key idea here parallels the panel data Logit model, which we will further compare in Remark 3.1 below.

Let $\alpha^* = \alpha^{(1)} - \alpha^{(0)}$, then the extreme elasticity is given by

$$\delta_i(x) = \frac{\partial \left(\pi_i(x) \left(1 - \pi_i(x) \right) \right)}{\partial x} \frac{x}{\pi_i(x) \left(1 - \pi_i(x) \right)} \to -|\alpha^*|,$$

which implies that, under the RV condition together with the additive form (14), all units share the same extreme elasticity, despite having unit-specific tail thickness and RV functions. We can estimate α^* via the conditional MLE on tail observations

$$\hat{\alpha}^* = \arg\max_{\alpha^*} \sum_{i=1}^{N} \left[\alpha^*(1 - Y_{i1}) \log \frac{X_{i1}}{X_{i2}} - \log \left(1 + \left(\frac{X_{i1}}{X_{i2}} \right)^{\alpha^*} \right) \right] \mathbf{1}\{\Xi_{N,i}\},$$

where event $\Xi_{N,i} = \{ Y_{i1} + Y_{i2} = 1, X_{it} \geq \underline{x}_N, \text{ for } t = 1, 2 \}$, and the tail threshold $\underline{x}_N \to \infty$ as $N \to \infty$. While in principle the tail threshold could be specific to both the period t and the value of y, we set it to a common $\underline{x}_N$ for simplicity in both theory and implementation.

Remark 3.1 We now compare our setup with the classic panel Logit model. The following discussion is also in line with Remark 2.1 for the cross-sectional case. For instance, equation (1) in Honoré and Kyriazidou (2000), with a slight change in notation, can be represented as:

$$\mathbb{P} \left(Y_{it} = 1 | X_{it} = x; C_i \right) = \frac{1}{1 + \exp \left(-(x\beta + C_i) \right)}, \tag{18}$$

and accordingly

$$\mathbb{P} \left(Y_{i1} = 1 | X_{i1} = x_1, X_{i2} = x_2, Y_{i1} + Y_{i2} = 1 \right) = \frac{1}{1 + \exp \left(-(x_1 - x_2)\beta \right)}. \tag{19}$$

Comparing the expressions in (16) and (17) with those in (18) and (19) leads to the following useful observations.

First, the classic Logit model is characterized by the exponential function of x, while ours is by a power function of x. However, by applying a log transformation to x, our method essentially adopts the same conditional likelihood function as the panel Logit model, but on tail observations. Our parameter of interest α^* corresponds to $-\beta$ in the panel Logit model, and as shown in Section 4.1 for large T, our $-\log A_i$ corresponds to their C_i as well.

Our derivation thus provides a theoretical justification for, and clarifies the underlying assumptions behind, the intuitive practice of taking the log of X_{it} when handling extreme observations. In practice, we can estimate α^* by conducting the conditional MLE for panel Logit models on tail observations, using $\log X_{it}$ as a regressor, where the negative of its coefficient provides the estimate for α^*.

Second, as discussed in Section 2.1, our method is a semiparametric approach, focusing on tail behavior. Especially, our approach avoids imposing parametric assumptions on the entire error distribution, and thus is more robust to potential misspecification. Of course, this robustness comes at the cost of reduced sample size, as we only utilize the large X_{it} observations.

3.2 Estimation and asymptotics

In this subsection, we derive the asymptotic normality of our estimator in the panel data setup. We now extend the baseline model in Section 3.1 and include additional covariates Z_i to capture potential observed heterogeneity. The unit-specific tail thickness becomes

$$\tilde{\alpha}_i^{(y)}\left(z; \theta^{(y)}\right) = z'\theta^{(y)} + \lambda_i > 0, \tag{20}$$

and $\tilde{\alpha}_i^{(y)}(z) = \tilde{\alpha}_i^{(y)}\left(z; \theta_0^{(y)}\right)$ denotes the tail index functions evaluated at the pseudo-true parameter values. The corresponding RV condition is $1 - F_{i,X_{it}|Y_{it},Z_i}(\cdot|y,z) \in RV_{-\tilde{\alpha}_i^{(y)}(z)}$. We focus on the time-invariant Z_i in this subsection, and the case with time-varying Z_{it} is discussed in Section 4.2.

Let $\mathcal{C}_i$ denote the unit-specific collection of functions $\left\{\beta_i^{(y)}(\cdot), C_i^{(y)}(\cdot), D_i^{(y)}(\cdot)\right\}$ in the unit-specific distribution $F_{i,X_{it}|Y_{it},Z_i}$: see Assumption 3.2 for more details. As $\{\lambda_i, \mathcal{C}_i\}$ are fixed for each i, we denote $\mathbb{P}_i$ as the probability measure given $\{\lambda_i, \mathcal{C}_i\}$, that is, $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot \ ; \ \lambda_i, \mathcal{C}_i)$. Similarly, we denote $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot \ ; \ \lambda_i, \mathcal{C}_i]$.

First, we adopt the following assumptions on the model setup.

Assumption 3.1 (Panel data: model assumptions) *Suppose we have:*

(a) $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}, Z_i\}$ *are independent across* i.

(b) *For each* i, $\{Y_{it}, X_{it}\}$ *are stationary across* $t = 1, 2$.

(c) *For each* i, *given* $\{\lambda_i, \mathcal{C}_i\}$, $Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, Z_i$.

(d) *For each* i, $\mathbb{P}_i (Y_{it} = 1 | X_{i1}, X_{i2}, Z_i) = \mathbb{P}_i (Y_{it} = 1 | X_{it}, Z_i)$ *for* $t = 1, 2$.

(e) *For each* i, $0 < \mathbb{P}_i(Y_{it} = 1 | Z_i) < 1$ *almost surely in* Z_i.

In condition (a), the covariates and outcomes are i.n.i.d. across units due to the unobserved unit-specific heterogeneity as fixed effects. Condition (b) ensures stationarity that helps eliminate the unit-specific tail effects. Conditions (c) and (d) imply conditional independence, thus for $y_1, y_2 \in \{0, 1\}$,

$$\mathbb{P}_i (Y_{i1} = y_1, Y_{i2} = y_2 | X_{i1}, X_{i2}, Z_i) = \mathbb{P}_i (Y_{i1} = y_1 | X_{i1}, Z_i) \cdot \mathbb{P}_i (Y_{i2} = y_2 | X_{i2}, Z_i). \quad (21)$$

Condition (e) guarantees non-degenerate probabilities for all units.

Second, we assume the following tail conditions.

Assumption 3.2 (Panel data: tail approximation) *For* $i = 1, \cdots, N$, *for* $y \in \{0, 1\}$, *and for* $\mathbb{P}_{Z_i}$-*almost all* $z \in \text{supp}(Z_i)$:

(a) *The conditional cdf* $F_{i, X_{it} | Y_{it}, Z_i} (x | y, z)$ *satisfies that*

$$1 - F_{i, X_{it} | Y_{it}, Z_i} (x | y, z) = C_i^{(y)} (z) x^{-\tilde{\alpha}_i^{(y)}(z)} \left(1 + D_i^{(y)} (z) x^{-\beta_i^{(y)}(z)} + r_i^{(y)} (x, z) \right),$$

as $x \to \infty$, *where* $C_i^{(y)} (z) > 0$, $\left| D_i^{(y)} (z) \right| \leq \overline{D} < \infty$, $\tilde{\alpha}_i^{(y)} (z) \geq \underline{\alpha} > 0$, *and* $0 < \underline{\beta} \leq \beta_i^{(y)}(z) \leq \bar{\beta} < \infty$, *for some constants* $\overline{D}$, $\underline{\alpha}$, $\underline{\beta}$, *and* $\bar{\beta}$.

(b) *The remainder term satisfies that* $\left| x^{\beta_i^{(y)}(z) + k} \frac{\partial^k r_i^{(y)}(x, z)}{\partial x^k} \right| \leq \bar{r}(x)$, *where* $\bar{r}(x) \to 0$ *as* $x \to \infty$, *for* $k = 0, 1$.

(c) *The conditional pdf* $f_{i, X_{it} | Y_{it}, Z_i} (x | y, z)$ *is non-increasing in* $x \geq \underline{x}$, *for some* $\underline{x} > 0$.

Please refer to the discussion after Assumption 2.2 for a detailed explanation. Compared to Assumption 2.2, we now have all these functions as unit-specific, indexed by i, and condition

(b) further bounds the derivative of the remainder term for the i.n.i.d. case. Especially, we can accommodate a unit-specific scale parameter, given by $\left[C_i^{(y)}(Z_i)\right]^{1/\tilde{\alpha}^{(y)}(Z_i)}$, in the Pareto tail approximation. These unit-specific functions will be absorbed into A_i in (22) below, and then differenced out for small T or estimated for large T.

Then, $\mathcal{L}_i^{(y)}(x,z) = C_i^{(y)}(z)\left(1 + D_i^{(y)}(z)x^{-\beta_i^{(y)}(z)} + r_i^{(y)}(x,z)\right)$ is the corresponding slowly varying function, and $\frac{\mathcal{L}_i^{(0)}(x,z)}{\mathcal{L}_i^{(1)}(x,z)} \to \mathcal{L}_i^*(z)$. Define $\theta^* = \theta^{(1)} - \theta^{(0)}$, and $A_i = \frac{\mathbb{P}_i(Y_{it}=0|Z_i)}{\mathbb{P}_i(Y_{it}=1|Z_i)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}} \mathcal{L}_i^*(Z_i)$. As $x \to \infty$,

$$\pi_i(x,z) = \mathbb{P}_i(Y_{it}=1|X_{it}=x, Z_i) \sim \frac{1}{1 + A_i \cdot x^{Z_i'\theta^*}}. \tag{22}$$

Note that as $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot\ ;\ \lambda_i, \mathcal{C}_i)$, A_i can be viewed as the fixed effects that can potentially depend on the observed heterogeneity Z_i and unobserved heterogeneity $\{\lambda_i, \mathcal{C}_i\}$ in an arbitrary way.

For small T, we again eliminate A_i by conditioning on $Y_{i1} + Y_{i2} = 1$ and construct the conditional MLE as follows:

$$\hat{\theta}^* = \arg\max_{\theta^* \in \Theta^*} \sum_{i=1}^{N} \left[Z_i'\theta^*(1 - Y_{i1})\log\frac{X_{i1}}{X_{i2}} - \log\left(1 + \left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}\right)\right] \mathbf{1}\{\Xi_{N,i}\}, \tag{23}$$

where $\Theta^* \subset \mathbb{R}^{d_Z}$ is a convex cone and the pseudo-true parameter value $\theta_0^* \in \text{int}(\Theta^*)$. Again, this can be implemented via the conditional MLE for panel Logit models, applied to tail observations and using $Z_i \log X_{it}$ as regressors.

Finally, analogous to the cross-sectional case, the number of units contributing to the conditional likelihood, $N_\Xi = \sum_{i=1}^{N} \mathbf{1}\{\Xi_i\}$, is a random variable, and we define the following non-random sequence to capture the asymptotic proportion of these units:

$$\xi_N = \frac{1}{N}\sum_{i=1}^{N} \xi_{N,i}, \text{ where } \xi_{N,i} = 2\mathbb{E}_i\left[\prod_{y\in\{0,1\}} \underline{x}_N^{-\tilde{\alpha}_i^{(y)}(Z_i)} C_i^{(y)}(Z_i)\, \mathbb{P}_i(Y_{it}=y|Z_i)\right]. \tag{24}$$

Also define the Hessian terms $H_{N0,i} = \mathbb{E}_i\left[\frac{\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta_0^*}}{\left(1+\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta_0^*}\right)^2}\left(\log\frac{X_{i1}}{X_{i2}}\right)^2 Z_iZ_i'\Bigg|\Xi_{N,i}\right]$ and $H_{N0} = \sum_{i=1}^{N} H_{N0,i}\xi_{N,i}/\sum_{i=1}^{N}\xi_{N,i}$. Now we assume the following conditions for estimation.

Assumption 3.3 (Panel data: estimation) *For $y \in \{0,1\}$, suppose we have:*

(a) $N\xi_N \to \infty$ *and* $N\xi_N = o\left((\underline{x}_N)^{\beta \frac{2+\kappa}{1+\kappa}}\right)$ *as* $N \to \infty$*, for some* $\kappa \geq 0$.

For sufficiently large N *and all* i:

(b) $\lambda_{\min}(H_{N0,i}) \geq \underline{\sigma}_N^2$*, where* $\underline{\sigma}_N^2 (N\xi_N)^{\kappa/(2+\kappa)} \to \infty$ *as* $N \to \infty$.

(c) $\mathbb{E}_i\left[\left|\log \frac{X_{i1}}{X_{i2}}\right|^4 \|Z_i\|^4 \middle| \Xi_{N,i}\right] < M$ *for some* $M < \infty$.

This assumption is in line with Assumption 2.3 for the cross-sectional case and accommodates i.n.i.d. observations across i. Condition (a) ensures that the number of observations contributing to the likelihood increases with the cross-sectional sample size, while employing undersmoothing to eliminate asymptotic bias. Conditions (b) and (c) guarantee the validity of the LLN and Lindeberg-Feller CLT for i.n.i.d. observations.

Under these assumptions, we establish the asymptotic normality of our estimator for panel data models with i.n.i.d. random variables. As the number of tail observations $N_\Xi = N\xi_N (1 + o_p(1))$ increases at a slower rate than N, the convergence rate is slower than the parametric rate.

Theorem 3.1 (Panel data: common parameters) *Suppose that* $\theta_0^* \in int(\Theta^*)$*, where* $\Theta^* \subset \mathbb{R}^{d_Z}$ *is a convex cone. Suppose Assumptions 3.1–3.3 hold. Then, as* $N \to \infty$,

$$\sqrt{N\xi_N} \left(H_{N0}\right)^{1/2} \left(\hat{\theta}^* - \theta_0^*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{d_Z}\right).$$

Moreover, the unit-specific extreme elasticity depends only on observed heterogeneity,

$$\delta_i(x, z) = \frac{\partial \left(\pi_i(x, z)(1 - \pi_i(x, z))\right)}{\partial x} \frac{x}{\pi_i(x, z)\left(1 - \pi_i(x, z)\right)} \to -\left|z'\theta^*\right|,$$

provided that $f'_{i,X_{it}|Y_{it},Z_i}(x|y, z)$ is non-increasing for sufficiently large x. In practice, we can consistently estimate this elasticity by substituting $\hat{\theta}^*$ for θ^*.

4 Extensions

4.1 Panel data with large T

Our previous panel data analysis has focused on small-T panels. In such cases, unit-specific parameters cannot be consistently estimated, and hence we eliminate them by conditioning

on the sum of Y_{it} over time, which is a sufficient statistic for the individual parameters. Now, with a large T, we are able to estimate these unit-specific parameters.

Based on (22) in Section 3.2, let $\tilde{A}_i = -\log A_i$, and we have that, as $x \to \infty$,

$$\mathbb{P}_i\left(Y_{it} = 1 | X_{it} = x, Z_i\right) \sim \frac{1}{1 + A_i \cdot x^{Z_i'\theta^*}} = \frac{1}{1 + \exp\left(-\tilde{A}_i + \log x \cdot Z_i'\theta^*\right)}. \tag{25}$$

Given this asymptotic equivalence, we can resort to panel Logit estimators for large N and large T. For example, as shown in Example 2 in Fernández-Val and Weidner (2018), the MLE estimates of $\{\theta^*, \{\tilde{A}_i\}\}$ are jointly consistent and $\hat{\theta}^* - \theta_0^*$ is asymptotically distributed as $\mathcal{N}\left(\frac{B}{N}, \frac{H^{-1}}{NT}\right)$, as $N, T \to \infty$. B denotes the asymptotic bias due to the incidental parameter, and analytical or jackknife methods can be applied for bias correction. See also Stammann, Heiss, and McFadden (2016). Once $\{\theta^*, \{\tilde{A}_i\}\}$ are consistently estimated, we can proceed to estimate the APE and provide unit-specific forecasts of $\mathbb{P}_i\left(Y_{i,T+1} = 1 | X_{i,T+1} = x, Z_i\right)$ based on (25).[11]

4.2 Panels with time-varying covariates Z_{it}

In empirical studies, some additional covariates Z may vary over time. For instance, in the context of country defaults, factors such as global business cycles could play a significant role. Then, the unit-specific tail indices in (20) becomes $\tilde{\alpha}_i^{(y)}(z_{it}) = z_{it}'\theta^{(y)} + \lambda_i$, and A_i in (22) becomes $A_{it} = \frac{\mathbb{P}_i(Y_{it}=0|Z_{it})}{\mathbb{P}_i(Y_{it}=1|Z_{it})} \frac{\tilde{\alpha}_i^{(0)}(Z_{it})}{\tilde{\alpha}_i^{(1)}(Z_{it})} \mathcal{L}_i^*(Z_{it})$, so we cannot difference out A_{it} as in (22).

However, under regularity conditions, θ^* can be estimated using a local (conditional) likelihood estimator (Tibshirani and Hastie, 1987; Honoré and Kyriazidou, 2000). Intuitively, it introduces an additional kernel weight term that controls the distances across Z_{it}. Assuming $T = 2$ for simplicity, the estimator for θ^* is given by

$$\hat{\theta}^* = \arg\max_{\theta^* \in \Theta^*} \sum_{i=1}^{N} k_h^{(d_Z)}(Z_{i1}, Z_{i2}) \left\{ (1 - Y_{i1}) \log \frac{X_{i1}}{X_{i2}} \cdot Z_{i1}'\theta^* + \log\left(1 + \left(\frac{X_{i1}}{X_{i2}}\right)^{Z_{i1}'\theta^*}\right) \right\} \mathbf{1}\{\Xi_{N,i}\}$$

where $k_h^{(d_Z)}(z_{i1}, z_{i2}) = \frac{1}{h^{d_Z}} \prod_{d=1}^{d_Z} k\left(\frac{z_{i1,d} - z_{i2,d}}{h}\right)$ is a multidimensional kernel with $k(\cdot)$ being

---

[11]Introducing time-varying thickness $\theta_t^{(y)}$ could be possible with additional parametric structure, such as combining our estimator with the autoregressive conditional density models in Hansen (1994). However, this changes A_i to time-varying $A_{it} = \frac{\mathbb{P}_i(Y_{it}=0|Z_i)}{\mathbb{P}_i(Y_{it}=1|Z_i)} \frac{\tilde{\alpha}_{it}^{(0)}(Z_i)}{\tilde{\alpha}_{it}^{(1)}(Z_i)} \mathcal{L}_i^*(Z_i)$, which can no longer be differenced out even if $\{\lambda_i, \mathcal{C}_i\}$ does not change over time. To address this, we need to impose further parametric assumptions on A_{it} or employ a local likelihood estimator as in Section 4.2.

the kernel function and h being the bandwidth.

4.3 Dynamic panel data model

In this subsection, we extend our analysis to dynamic panel data, incorporating predetermined variables such as lagged outcomes to capture potential persistence.

For small T, w.l.o.g., we assume the first order Markov property: for each i,

$$X_{it}, Y_{it} \mid X_{i,1:t-1}, Y_{i,1:t-1}, Z_i = X_{it}, Y_{it} \mid Y_{i,t-1}, Z_i,$$

given $\{\lambda_i, \mathcal{C}_i\}$, where $\mathcal{C}_i$ comprises unit-specific functions in $F_{i,X_{it}|Y_{it},Y_{i,t-1},Z_i}$, similar to the tail approximation in Assumption 2.2. The first order Markov property implies the stationarity of the conditional joint distribution $Y_{it}, Y_{i,t-1}, X_{it} \mid Z_i$.

Recall that previously we applied Bayes' theorem by partitioning the data into two subsets based on $Y_{it} = 0$ and 1. With dynamics, we now further partition the data according to transition dynamics. In the first order Markov case, there are four transition patterns: $0 \to 0$, $0 \to 1$, $1 \to 0$, and $1 \to 1$. Therefore, we partition the data into four subsets by $(Y_{it}, Y_{i,t-1}) = (y, y_-)$, and characterize Bayes' theorem as follows

$$
\begin{aligned}
&\mathbb{P}_i \left(Y_{it} = y | X_{it} = x, Y_{i,t-1} = y_-, Z_i = z \right) \\
&= \frac{f_{i,X_{it}|Y_{it},Y_{i,t-1},Z_i} \left(x | y, y_-, z \right) \mathbb{P}_i \left(Y_{it} = y | Y_{i,t-1} = y_-, Z_i = z \right)}{\sum_{y,y_-} f_{i,X_{it}|Y_{it},Y_{i,t-1},Z_i} \left(x | y, y_-, z \right) \mathbb{P}_i \left(Y_{it} = y | Y_{i,t-1} = y_-, Z_i = z \right)}.
\end{aligned}
$$

Similar to (20), we introduce unit-specific tail thickness as

$$\tilde{\alpha}_i^{(yy-)}(z) = z'\theta^{(yy-)} + \lambda_i.$$

Then, we have $1 - F_{i,X_{it}|Y_{it},Y_{i,t-1},Z_i} \left(\cdot | y, y_-, z \right) \in RV_{-\tilde{\alpha}_i^{(yy-)}(z)}$.

In the previous static panel analysis, our exercise was equivalent to normalizing $\theta^{(0)} = \mathbf{0}$ and estimating $\theta^* = \theta^{(1)}$. Here, we similarly normalize $\theta^{(00)} = 0$ to ensure identification. With four transition patterns, a minimum of five periods of data $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})'$ are required. The following four events cover all transition patterns and help construct the conditional likelihood:

$$E_1 : \mathbf{Y}_i = (0, 0, 1, 1, 0)', \quad E_2 : \mathbf{Y}_i = (0, 1, 1, 0, 0)',$$

$$E_3 : \mathbf{Y}_i = (1, 1, 0, 0, 1)', \quad E_4 : \mathbf{Y}_i = (1, 0, 0, 1, 1)'.$$

Given the stationarity of the conditional joint distribution $Y_{it}, Y_{i,t-1}, X_{it} \mid Z_i$ implied by the first order Markov assumption, we have that as $x_t \to \infty$ for $t = 1, \cdots, 5$,

$$\mathbb{P}_i\left(E_1 \big| \cup_{e=1}^{4} E_e, \{X_{it} = x_t\}_{t=1}^{5}, Z_i = z\right)$$

$$\sim \frac{x_3^{-z'\theta^{(01)}} x_4^{-z'\theta^{(11)}} x_5^{-z'\theta^{(10)}}}{\left(\begin{array}{l} x_3^{-z'\theta^{(01)}} x_4^{-z'\theta^{(11)}} x_5^{-z'\theta^{(10)}} + x_2^{-z'\theta^{(01)}} x_3^{-z'\theta^{(11)}} x_4^{-z'\theta^{(10)}} \\ + x_2^{-z'\theta^{(11)}} x_3^{-z'\theta^{(10)}} x_5^{-z'\theta^{(01)}} + x_2^{-z'\theta^{(10)}} x_4^{-z'\theta^{(01)}} x_5^{-z'\theta^{(11)}} \end{array}\right)}.$$

Then, the conditional MLE for $\left(\theta^{(01)}, \theta^{(10)}, \theta^{(11)}\right)$ can be constructed similarly to the static panel case in Section 3.2. Note that unit i contributes to the conditional likelihood only if X_{it} appears in the tail for at least five periods. This data requirement may be challenging, so the method would be more suitable for datasets with a larger N, and a fixed but slightly larger T.

For large T, the bias correction in Fernández-Val and Weidner (2018) remains applicable to dynamic panel data models, so the estimator in Section 4.1 remains valid.

5 Monte Carlo simulations

We conduct two sets of Monte Carlo simulation experiments. Experiment 1 examines cross-sectional data and focuses on estimation performance, providing intuitions into when and how the proposed estimator outperforms the alternatives. Experiment 2 investigates panel data with large N and large T,[12] and focuses on pseudo out-of-sample forecasting performance, aligning more closely with the empirical example of bank loan charge-off rates.

5.1 Alternative estimators

We compare the proposed estimator with four alternatives: a Logit estimator using all observations (Logit, all X), a Logit estimator using only tail observations (Logit, tail X), a local linear estimator, and a local Logit estimator, where the first two are parametric estimators and the last two are nonparametric ones.

---

[12]We also conducted Monte Carlo simulations in panel data with large N and small T in a previous version of this paper. Results from these simulations are available upon request.

For cross-sectional data, let $\beta = (\beta_0, \beta_1)'$. First, the Logit estimator is defined as

$$Y_i = \mathbf{1}(\beta_0 + \beta_1 X_i - \varepsilon_i \geq 0), \quad \varepsilon_i \sim \text{ standard logistic.}$$

Second, the local linear estimator is given by

$$\hat{\beta}(x) = \arg\min_{\beta} \sum_{i=1}^{N} k_h(X_i - x) \left[Y_i - \beta_0 - \beta_1(X_i - x)\right]^2, \text{ and } \hat{\mathbb{E}}[Y_i | X_i = x] = \hat{\beta}_0(x),$$

where $k_h(\cdot)$ is a kernel function with bandwidth h. We employ a Gaussian kernel here, and choose the bandwidth based on Silverman's rule of thumb $h \approx 1.06\hat{\sigma}N^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of the X_is. The results are robust with respect to a range of bandwidth choices. Finally, the local Logit estimator is a flexible nonparametric estimator that specifically accounts for binary outcomes Y_i. Let $p_i(\beta, x) = \frac{1}{1+\exp[-(\beta_0 + \beta_1(X_i - x))]}$. The local Logit maximizes the locally weighted log-likelihood function

$$\hat{\beta}(x) = \arg\max_{\beta} \sum_{i=1}^{N} k_h(X_i - x) \left[Y_i \log p_i(\beta, x) + (1 - Y_i) \log\left(1 - p_i(\beta, x)\right)\right],$$

$$\hat{\mathbb{E}}[Y_i | X_i = x] = p_i(\hat{\beta}(x), x).$$

For panel data, the Logit estimator is given by

$$Y_{it} = \mathbf{1}(\beta_0 + \beta_1 X_{it} + C_i - \varepsilon_{it} \geq 0), \quad \varepsilon_{it} \sim \text{standard logistic,}$$

where C_i captures unobserved individual heterogeneity. In panels with large N and large T, $\{\beta_0, \beta_1, \{C_i\}\}$ can again be jointly estimated using a fixed effects estimator with bias corrections (Fernández-Val and Weidner, 2018; Stammann, Heiss, and McFadden, 2016). For the local linear and local Logit estimators, the setup is more flexible

$$Y_{it} = g\left(X_{it}, C_i, \varepsilon_{it}\right),$$

where the function g and the distributions of C_i and ε_{it} could be unknown. We incorporate a correlated random effects structure, which allows for the unobserved individual heterogeneity to be correlated with the covariates X_{it} (and Z_i) and thus may help enhance the performance of these alternatives. More specifically, suppose there is a sufficient statistic V_i which could

Table 1: Monte Carlo design - Experiment 1

Model:	$Y_i = \mathbf{1}\left(X_i - \varepsilon_i \geq \text{med}_X - \text{med}_\varepsilon\right)$
Covariate:	$X_i \sim |t_{\alpha_X}|,\ \alpha_X = 0.5, 1, 1.5, 2$
Error term:	$\varepsilon_i \sim |t_{\alpha_\varepsilon}|,\ \alpha_\varepsilon = 0.5, 1, 1.5, 2$
Sample Size:	$N = 10000$
# Repetitions:	$N_{sim} = 1000$

be multidimensional, such that $C_i|X_{i,1:T} = C_i|V_i$. One commonly used example of V_i is the time sum of X_{it}, i.e., $V_i = \sum_t X_{it}$. Then, V_i can be included in the conditioning set for the local linear and local Logit estimators, so we essentially run a nonparametric regression to estimate the conditional mean $\mathbb{E}[Y_i|X_i, V_i]$: see details in Liu, Poirier, and Shiu (2024) for general nonlinear panel data models. Without loss of generality, let V_i be a scalar for notation simplicity. Now $\beta = (\beta_0,\ \beta_1,\ \beta_2)'$, $p_{it}(\beta, x) = \frac{1}{1+\exp[-(\beta_0+\beta_1(X_{it}-x)+\beta_2(V_i-v))]}$, and

$$\text{Local linear: } \min_\beta \sum_{i=1}^N k_h(X_{it} - x)k_h(V_i - v)\left[Y_{it} - \beta_0 - \beta_1(X_{it} - x) - \beta_2(V_i - v)\right]^2,$$

$$\text{Local Logit: } \max_\beta \sum_{i=1}^N k_h(X_{it} - x)k_h(V_i - v)\left[Y_{it}\log p_{it}(\beta, x) + (1 - Y_{it})\log\left(1 - p_{it}(\beta, x)\right)\right].$$

Furthermore, for all these estimators, we can also incorporate additional covariates Z_i, and the formulas are similar to those above.

5.2   Experiment 1: cross-sectional data

Experiment 1 is based on cross-sectional data, where we focus on comparing estimation performance across estimators.

The Monte Carlo design is summarized in Table 1. The data are generated from a threshold-crossing model with both X_i and ε_i following Student-t distributions.[13] Here we consider a range of α values from 0.5 to 2. As elaborated in Proposition 2.2, in the tail, $\alpha^{(0)} = \alpha_X + \alpha_\varepsilon$ ranges from 1 to 4, $\alpha^{(1)} = \alpha_X$ varies from 0.5 to 2, and the extreme elasticity $-\left|\alpha^{(1)} - \alpha^{(0)}\right| = -\alpha_\varepsilon$ spans from -2 to -0.5.[14] The sample size $N = 10000$ is directly

---

[13]We use the difference in medians ($\text{med}_X - \text{med}_\varepsilon$) as the threshold to keep the samples more balanced between $Y_i = 0$ and 1.

[14]In many economic datasets, the tail index α typically falls between 1 and 2. For example, in a review paper, Gabaix (2009) mentions that "it seems that the tail exponent of wealth is rather stable, perhaps around 1.5," referencing Klass, Biham, Levy, Malcai, and Solomon (2006). Also, Clauset, Shalizi, and

Figure 1: Log-log plot - Experiment 1, $\alpha_X = 1$, $\alpha_\varepsilon = 1$
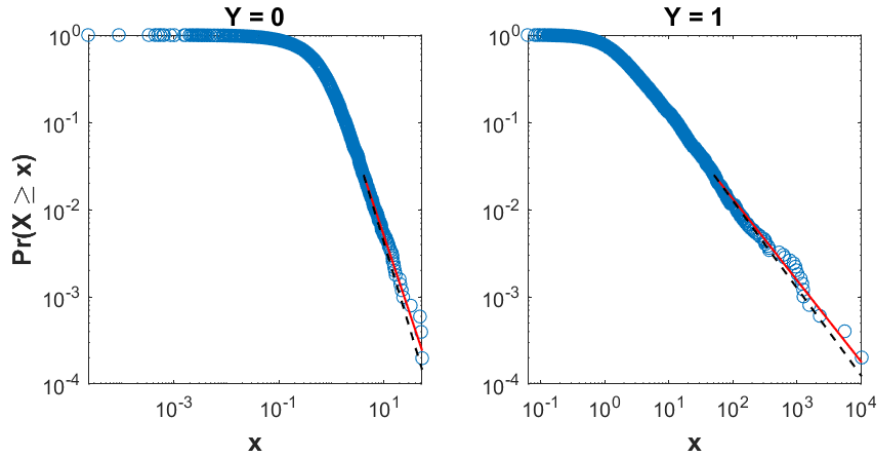
Notes: In the tail, $\alpha^{(0)} = 2$ and $\alpha^{(1)} = 1$ are indicated by the black dash lines. The estimated $\hat{\alpha}^{(0)}$ and $\hat{\alpha}^{(1)}$ are indicated by the red solid lines.

comparable with our empirical data sets on bank loan charge-off rates, which comprises 8538 banks in the baseline sample. For each experimental setup, we execute 1000 Monte Carlo simulations.

Our tail estimator for $\alpha^{(y)}$ is based on the "rank-1/2" estimator in Gabaix and Ibragimov (2011), which can be viewed as a refinement of the classic Hill estimator and often performs well in finite samples. Then, we estimate the conditional probability $\pi(x)$ using the sample analog of (7) and (8). We also compare with the alternative estimators described in Section 5.1. For both the tail estimator and the "Logit, tail X" estimator, we set the cutoffs $\underline{x}_N^{(y)}$ at the 97.5th percentile of the distributions of X for $Y = 0$ and 1 separately.[15] In the main text, we plot the comparisons regarding estimated parameters, conditional probability, and extreme elasticity, for the specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$. For detailed results across all model specifications, please refer to Tables A.1–A.4 in the Appendix. The main messages are similar across specifications with different tail index values.

Figure 1 depicts a log-log plot from one of the 1000 Monte Carlo simulations, where the tail observations align around a downward sloping line, contrasting with the flatter non-tail region. This distinct pattern between the tail and non-tail region is common in many empirical datasets as well, and suggests that the proposed tail estimator would be able

Newman (2009) remark that the tail index usually lies between 1 and 2 (note that the α in their notation corresponds to $\alpha - 1$ in ours).

[15]The tail estimator is robust with respect to a range of cutoffs $\underline{x}_N^{(y)}$. As evident from the log-log plot in Figure 1, the tail exhibits a pronounced downward pattern, with the slope remaining stable across small variations in $\underline{x}_N^{(y)}$.

Figure 2: Parameter estimation - Experiment 1, $\alpha_X = 1$, $\alpha_\varepsilon = 1$
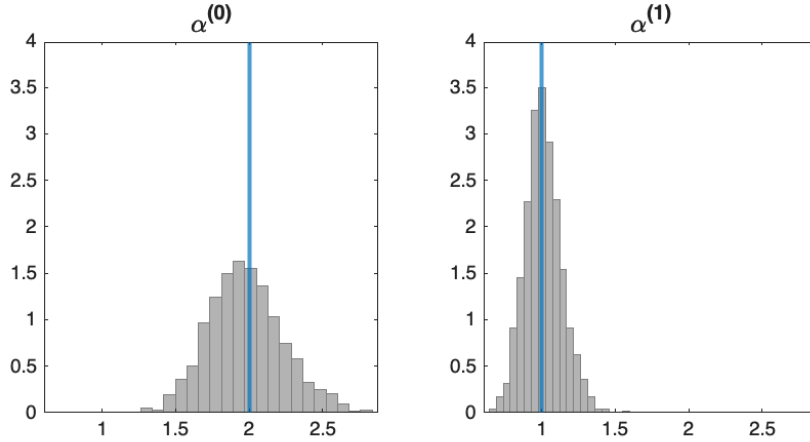

Notes: In the tail, $\alpha^{(0)} = 2$ and $\alpha^{(1)} = 1$ are indicated by the blue vertical lines.

to effectively capture the heavy tail behavior. The estimated tail indices (red solid lines) closely match their true asymptotic values (black dashed lines). Figure 2 further shows the distributions of the parameter estimates from all 1000 Monte Carlo simulations. The distributions are both bell-shaped and centered around the true asymptotic values indicated by the blue vertical lines. Notably, $\alpha^{(1)}$ is more precisely estimated than $\alpha^{(0)}$, as the former corresponds to a thicker tail.

Figure 3 presents the estimated conditional probabilities for X at the 90th, 95th, 97.5th, and 99th percentiles of its distribution. The blue vertical lines indicate the true conditional probability $\mathbb{P}(Y = 1 \mid X = x)$, which increases with the evaluation point x. The proposed tail estimator dominates all other alternatives, centered around the true probabilities with lower variance. The variance of the tail estimator decreases as the evaluation point x increases, reflecting reduced estimation uncertainty as the conditional probability approaches one.

In contrast, the Logit estimators are largely biased due to the model misspecification that fails to account for the heavy tail. The direction of the bias depends on the relative positions of the estimation sample and the evaluation points. When the estimation sample includes all observations, the "Logit, all X" estimator shows an upward bias. This occurs because the estimation sample is overweighted by non-tail observations, and thus the tail evaluation points are too extreme given the misspecified thin-tailed logistic distributions, which leads to an overestimation of the probability of $Y = 1$ in the tail evaluation points. Conversely, when the estimation sample includes tail observations only, the "Logit, tail X" estimator tends to exhibit a downward bias, except at the 99th percentile. The reason is that given the logistic

31

Figure 3: $\hat{\mathbb{P}}(Y = 1|X = x)$ - Experiment 1, $\alpha_X = 1$, $\alpha_\varepsilon = 1$
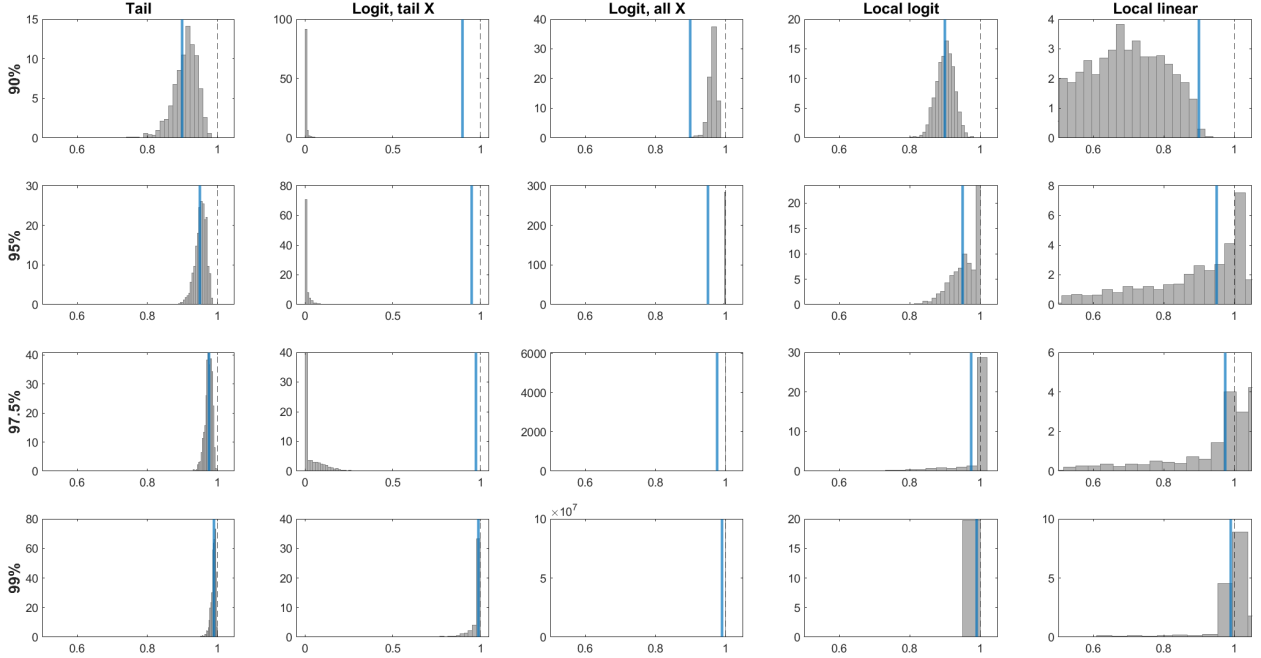
Notes: Each row evaluates the probability at a specific percentile of the distribution of X (90th, 95th, 97.5th, and 99th). The true $\mathbb{P}(Y = 1|X = x)$ is indicated by the blue vertical lines.

model's misspecified assumption of a thinner tail, lower evaluation points appear relatively moderate compared to the tail observations for estimation, resulting in an underestimation of the probability of $Y = 1$.

The local Logit estimator performs reasonably well for less extreme evaluation points, such as when x is at the 90th percentile, but it exhibits large bias and variance at more extreme points. The local linear estimator is even more flexible, and thus yields even larger variance across all evaluation points. Due to their poor estimation performance and relatively long computation times, we omit these nonparametric estimators in Experiment 2 and the empirical example below, except for the baseline case in Table 4.

Across all comparisons, we see that our tail estimator enjoys the best of both worlds as a semiparametric estimator that puts parametric assumptions only on the tail, the region of primary interest, while remaining agnostic in the non-tail region. Furthermore, as mentioned before, our approach offers significant improvement in finite samples for moderately large x where $\pi(x)$ may not be very close to 1.

Figure 4 plots the estimated extreme elasticities evaluated at the 97.5th percentile of the distribution of X, and the key messages are similar—the Logit estimators induce large bias,

Figure 4: Extreme elasticity estimation - Experiment 1, $\alpha_X = 1$, $\alpha_\varepsilon = 1$
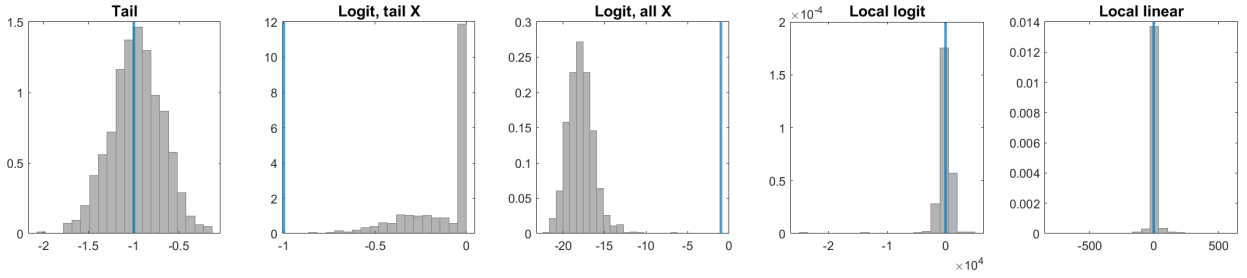
Notes: The extreme elasticity, $-\left|\alpha^{(1)} - \alpha^{(0)}\right| = -\alpha_\varepsilon = -1$, is indicated by the blue vertical lines. Evaluated at the 97.5th percentile of the distribution of X.

Table 2: Monte Carlo design - Experiment 2

Model:	$Y_{it} = \mathbf{1}\left(X_{it} - \varepsilon_{it} \geq \mathrm{med}_X - \mathrm{med}_\varepsilon\right)$
Covariate:	$X_{it} \sim \lvert t_{\lambda_i}\rvert$, $\lambda_i = \max\{\alpha_X + \tilde{\lambda}_i, 0\}$,
	$\alpha_X = 1, 1.5, 2$, $\tilde{\lambda}_i \sim 0.2N(-0.25, 0.1^2) + 0.8N(0.25, 0.1^2)$
Error term:	$\varepsilon_{it} \sim \lvert t_{\alpha_\varepsilon}\rvert$, $\alpha_\varepsilon = 1, 1.5, 2$
Sample Size:	$N = 10000$, $T = 100$
# Repetitions:	$N_{sim} = 1000$

the nonparametric estimators exhibit high variance, whereas the proposed tail estimator flexibly and efficiently capture the tail behavior and yields the most accurate estimates.

5.3 Experiment 2: panel data

Experiment 2 examines panel data with large N and large T, focusing on pseudo out-of-sample forecasting performance, which is closely related to the empirical analysis of bank loan charge-off rates.

The Monte Carlo design is adapted from the cross-sectional case in Experiment 1, with specifics described in Table 2. Now we incorporate unit-specific tail thickness λ_i for X_{it}, where the underlying distribution of λ_i is bimodal. For the tail estimator, we employ bias corrections as in Fernández-Val and Weidner (2018) and Stammann, Heiss, and McFadden (2016). See Sections 4.1 and 5.1 for additional details on the tail and alternative estimators, respectively. The cutoffs for both the tail estimator and "Logit, tail X" are set at the 90th percentile of the X distributions. Recall that the tail estimator is robust over a range of cutoffs, as discussed in footnote 15.

The accuracy of density forecasts is evaluated using the log predictive score (LPS), as recommended by Amisano and Giacomini (2007). The LPS is calculated as $LPS =$

Table 3: Parameter estimation and forecast evaluation - Experiment 2

			$\alpha_X = 1$	$\alpha_X = 1.5$	$\alpha_X = 2$
$\alpha_\varepsilon = 1$	$\hat{\alpha}^*$		-1.05 (0.03)	-1.06 (0.03)	-1.07 (0.03)
	$\text{LPS} \cdot N_f^\dagger$	*Tail*	*-113*	*-235*	*-318*
		Logit, tail X	-148 ***	-47 ***	-8 ***
		Logit, all X	-480 ***	-89 ***	-38 ***
$\alpha_\varepsilon = 1.5$	$\hat{\alpha}^*$		-1.53 (0.06)	-1.53 (0.04)	-1.52 (0.04)
	$\text{LPS} \cdot N_f^\dagger$	*Tail*	*-37*	*-106*	*-172*
		Logit, tail X	-50 ***	-43 ***	-7 ***
		Logit, all X	-267 ***	-64 ***	-33 ***
$\alpha_\varepsilon = 2$	$\hat{\alpha}^*$		-2.02 (0.09)	-1.98 (0.06)	-1.95 (0.05)
	$\text{LPS} \cdot N_f^\dagger$	*Tail*	*-15*	*-55*	*-102*
		Logit, tail X	-13 ***	-35 ***	-7 ***
		Logit, all X	-135 ***	-55 ***	-20 ***

Notes: $\hat{\alpha}^*$ is estimated by the tail estimator, averaged across $N_{sim} = 1000$ repetitions. Standard errors across repetitions are shown in parentheses. For the tail estimator, the table reports the exact values of $\text{LPS} \cdot N_f^\dagger$ (averaged across $N_{sim} = 1000$ repetitions). For other estimators, the table reports their differences from the tail estimator. The tests compare other estimators with the tail estimator, with significance levels indicated by *: 10%, **: 5%, and ***: 1%.

$\frac{1}{N_f^\dagger} \sum_{i \in \mathcal{N}_f^\dagger} \log \hat{p}(y_{i,T+1}|D)$, where $y_{i,T+1}$ is the outcome at time $T+1$, and $\hat{p}(y_{i,T+1}|D)$ is the predictive likelihood based on the estimated model and observed data D. $\mathcal{N}_f^\dagger$ is the set of units forecasted, characterized by following criteria: (1) X_{it} exceeds the 90th percentile in the estimation sample, (2) Y_{it} switches values over time among these tail observations, and (3) $X_{i,T+1}$ also falls within the tail during the forecasting period. To assess the significance of the differences in the LPS, we combine the tests from Amisano and Giacomini (2007) for density forecasts and Qu, Timmermann, and Zhu (2023) for panel data.

In Table 3, the first row in each subpanel presents extreme elasticity estimates from the tail estimator, which are close to the true values, $-\alpha_\varepsilon$. Subsequent rows compare forecast accuracy among estimators, and we see that the tail estimator consistently outperforms both "Logit, tail X" and "Logit, all X" across all specifications. Although "Logit, tail X" is better than "Logit, all X," it still performs worse than the tail estimator, especially with smaller α_X and α_ε, where heavy tails are more pronounced. Therefore, it is important to distinguish the tail from the middle of the sample as well as account for heavy tail patterns. To further demonstrate this point, Figure A.1 in the Appendix shows the scatter plots of the LPS from all Monte Carlo repetitions in the setup with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$.

6 Empirical example: housing prices and bank riskiness

Charge-off rates serve as an indicator of bank losses. A bank could be riskier for a particular type of loan if its corresponding charge-off rates exceed a certain level. In our analysis, we focus on a panel of small banks with assets less than \$1 billion, similar to Liu, Moon, and Schorfheide (2023). Since the banks are small, it is reasonable to assume that they operate primarily in local markets. In this empirical example, we examine the impact of substantial local housing price declines on the riskiness of small banks, considering that this channel played a pivotal role during the 2007–2008 financial crisis.

6.1 Data and sample

In this empirical example, we focus on the setup in (25) in Section 4.1 for panel data models with large N and large T. The binary outcome Y_{it} is a risk dummy based on the loan charge-off rate for a specific loan type of bank i in quarter t. $Y_{it} = 1$ if the charge-off rate is greater than a level c. We present results for $c = 0$ in the main text and relegate robustness checks for alternative c values to the Appendix. The qualitative findings are consistent across different levels of c. This exercise aligns with policy analysis practices, where policymakers often compare current charge-off rates to historical averages to monitor bank risk: see, for example, the Federal Reserve Board's Financial Stability Report.[16] For instance, the historical average charge-off rate for Residential Real Estate (RRE) loans is around 0.1%, with the average plus two standard deviations around 0.2%. Robustness checks include these levels.

The extreme regressor X_{it} captures decreases in local housing prices. To convert the extreme values to the right tail, we define X_{it} as the deflation rate of the local housing price in the previous quarter.[17] The additional covariate Z_i is given by the average quarterly change in the local unemployment rate, accounting for local economic conditions.

Our data are obtained from the following sources. Bank balance sheet data at a quarterly frequency, such as loan charge-off rates, are constructed based on the Call Reports from the

---

[16]https://www.federalreserve.gov/publications/financial-stability-report.htm.

[17]We define the tail region as observations above the 90th percentile of X_{it}, as in Experiment 2. Given that the 90th percentiles are positive in all our samples, $\log X_{it}$ is well-defined in the tail.

Federal Reserve Bank of Chicago.[18] The local market is defined at the county level, and the local market for each bank is determined based on the annual Summary of Deposits from the Federal Deposit Insurance Corporation.[19] Housing price indices at a quarterly frequency (all transactions, not seasonally adjusted) are sourced from the Federal Housing Finance Agency, and the 3-digit zip code data are converted to the county level using the HUD USPS ZIP Code Crosswalk from the Department of Housing and Urban Development.[20] The county-level not seasonally adjusted unemployment rates are obtained from the Bureau of Labor Statistics website, and we aggregate the monthly data to a quarterly frequency by time averaging. Finally, we standardize the county-level housing price deflations and unemployment rate changes using the means and standard deviations calculated from their corresponding aggregate time series.

Our baseline sample focuses on the RRE charge-off rates. The estimation sample spans from 1999Q4 to 2009Q3, comprising $N = 8538$ small banks across 40 quarters.[21] There are $N_e^\dagger = 2642$ banks in the tail for more than one period and contributing to the likelihood, that is, X_{it} is above the 90th percentile of the estimation sample and Y_{it} switches values across time for these tail observations. We perform a pseudo out-of-sample forecast for the period of 2009Q4. There are $N_f^\dagger = 2098$ banks that satisfy the conditions for $N_e^\dagger$ and additionally have $X_{i,T+1}$ fall in the tail during the forecasting period.[22]

For robustness check, we also consider (non-farm) non-residential commercial real estate (CRE) charge-off rates, various time dimensions in the estimation samples ranging from 32 to 48 quarters, and different forecasting periods $T + 1 = 2009Q3$ and $2009Q4$. The sample statistics of these samples are reported in Table A.5 in the Appendix. The main results remain consistent across different samples.

---

[18]Please refer to Appendix D in Liu, Moon, and Schorfheide (2023) for details on constructing loan charge-off rates from the raw data.

[19]We calculate the deposits received by each bank from every county and link the bank to the county from which it received the largest amount of deposits. We also assess the robustness of our analysis by constructing weighted averages of the covariates, using deposit proportions from each county as weights. The results are very similar, which can be attributed to the concentration of deposits across counties and the similarity in covariate values among neighboring counties.

[20]We use the 2010Q1 Crosswalk, the earliest available version that is close to our empirical time frame. We also experimented with the county-level Zillow Home Value Index, but it unfortunately has more missing data within our study period, though the results are qualitatively similar.

[21]We exclude banks with: 1) average non-missing domestic total assets exceeding $1 billion, or 2) missing target charge-off rates for all periods in the sample. These criteria result in the removal of 583 and 162 banks, respectively, for the baseline RRE sample.

[22]To account for banks' endogenous exit choices, one could extend to a panel Tobit model as in Liu, Moon, and Schorfheide (2023), which is left for future research.

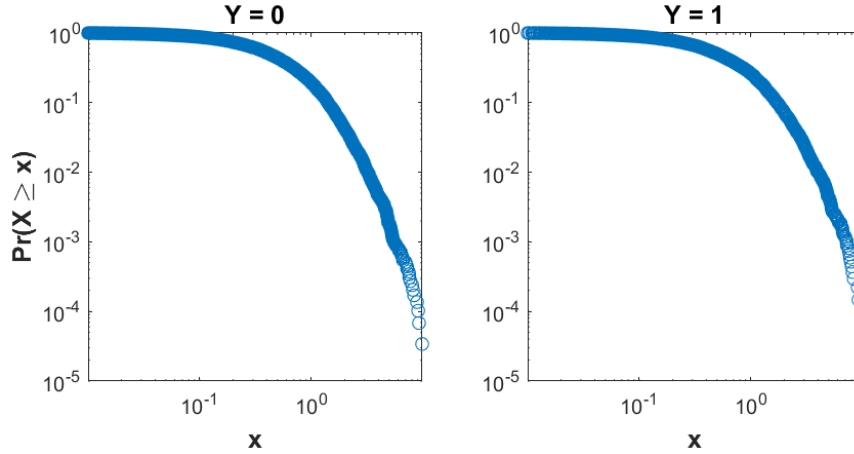Figure 5: Log-log plot - banking application, baseline sample

Table 4: Forecast evaluation - banking application, baseline samples

	RRE	CRE
Tail	*-1433.72*	*-1134.48*
Logit, tail X	-18.66 ***	-23.49 ***
Logit, all X	-160.25 ***	-54.26 ***
Local Logit	-429.20 ***	-516.43 ***

Notes: Forecasting period = 2009Q4, $T = 40$. For the tail estimator, the table reports the exact values of LPS$\cdot N_f^\dagger$. For other estimators, the table reports their differences from the tail estimator. The tests compare other estimators with the tail estimator, with significance levels indicated by *: 10%, **: 5%, and ***: 1%.

Based on the sample skewness and kurtosis in Table A.5 for all samples, as well as the log-log plot in Figure 5 for the baseline sample,[23] we see that $X_{it}|Y_{it} = y$ indeed exhibits heavy right tails, so our tail estimator would be more appealing. Furthermore, it is worth noting that in finite samples, the probability of $Y_{it} = 1$ for tail x, $\pi(x)$, is not necessarily very close to 1. This is evident from both the empirical data (see the log-log plot in Figure 5 and the histogram in Figure A.2 in the Appendix) and the estimated model (see the predictive probability in Figure 6).

6.2   Results

Table 4 compares forecasting performance across estimators. The estimators are similar to those in the Monte Carlo Experiment 2: see Sections 5.1 and 5.3 for more details. The proposed tail estimator is the overall best. "Logit, tail X" ranks second, yet still significantly

---

[23]Fitted lines, such as those in Figure 1, are not plotted here, because these lines could be unit-specific due to observed heterogeneity Z_i and unobserved heterogeneity $\{\lambda_i, C_i\}$.

worse than the tail estimator, indicating the importance of carefully addressing the heavy tail behavior. "Logit, all X" ranks third, indicating the presence of large misspecification and possible distinct pattern in the tail compared to the middle range. Local Logit yields the least accurate forecasts, suggesting that while being flexible, the nonparametric approach could be too noisy given the limited data available in the tail.

Table A.6 in the Appendix provides further details on the parameter and APE estimates based on the tail estimator. First, the coefficient on $\log X_{it}$ is always significant, with values around 0.95–1.15 for the RRE samples and around 1.2–1.4 for the CRE ones. The negative of this coefficient can be roughly viewed as the homogeneous part of the extreme elasticity, and the estimated values suggest the potential presence of heavy tails. Second, the coefficient on $Z_i \log X_{it}$ is mostly positive, being larger and significant for the RRE samples, while smaller and insignificant for the CRE ones. This difference aligns with intuition: larger increases in unemployment could directly amplify the impact of housing price drops on residential loan risk, whereas changes in unemployment may not directly affect non-residential commercial loan risk. Third, the estimated APEs are around 0.15 for the RRE samples and 0.13 for the CRE ones, which could be interpreted as that in the tail, a 1% decrease in housing prices in the previous quarter corresponds to approximately a 0.15 (0.13) increase in the probability of high risk for RRE (CRE) loans. Finally, the unobserved heterogeneity exhibits a larger dispersion than the observed heterogeneity, as the sample variances of the estimated $\tilde{A}_i$ are around 1 while the sample variances of $\hat{\theta}^*_{Z \log X} Z_i \log X_{it}$ range from 0.1–0.3.

Tables A.6 and A.7 in the Appendix also show that our results are robust with respect to the level $c = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1$,[24] to both CRE and RRE loans, to the estimation sample's time dimension $T = 32, 36, 40, 44, 48$, and to the forecasting period being 2009Q3 and 2009Q4.

Figure 6 plots the average predictive probability of high risk across counties, with darker shades indicating higher risk levels in those counties. Notably, Florida and the Great Lakes regions appear as areas with elevated risk. This heterogeneity in risk may arise from observed heterogeneity, such as variations in local housing prices $X_{i,T+1}$ and local unemployment rates Z_i, as well as unobserved heterogeneity $\{\lambda_i, \mathcal{C}_i\}$, which is captured by $\tilde{A}_i$, where larger values of $\tilde{A}_i$ are associated with higher risk: see (25).

---

[24]As the level c increases, the coefficient on $Z_i \log X_{it}$ becomes less significant, particularly becoming insignificant when $c = 1$. This is because a larger c reduces the occurrence of $Y_{it} = 1$, leading to fewer banks in the tail with Y_{it} switching values across time. Then, the effective sample size for estimation $N_e^{\dagger}$ substantially decreases, resulting in noisier estimates.

Figure 6: Predictive probability of high risk - baseline sample

☐ < 0.26   ☐ [0.26,0.36)   ☐ [0.36,0.48)   ■ [0.48,0.6)   ■ [0.6,0.75)   ■ >= 0.75
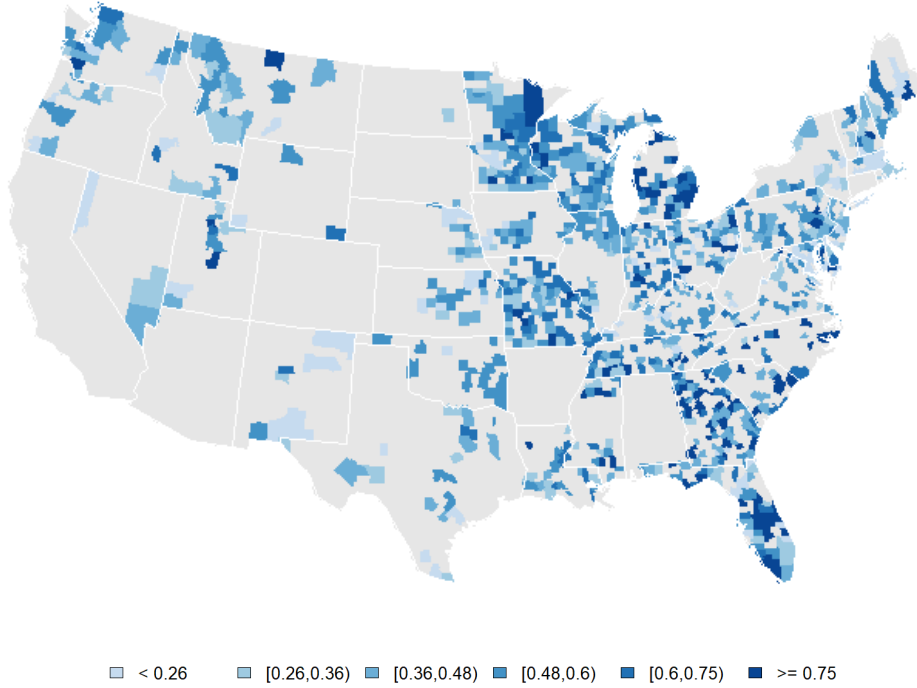
Accordingly, in Table 5, we regress the estimated $\tilde{A}_i$ on bank characteristics. The "Initial" column uses bank characteristics in the initial period 1999Q4, which are exogenous to subsequent dynamics. The "Average" column uses time-averaged bank characteristics over the estimation period, incorporating more recent information. The findings across both columns are in general consistent. We see that larger banks (measured by log assets), those that specialize in RRE loans (measured by RRE loans to total loans ratio), in lending activities (measured by loan to assets ratio), with more diversified earnings (measured by the share of non-interest income to total income), and with higher operational efficiency (measured by the negative of overhead costs to assets ratio), tend to have a higher $\tilde{A}_i$ and thus higher risk, as these characteristics could help increase banks' capacity to assume riskier RRE loans. On the other hand, higher profitability (measured by return on assets) corresponds to a lower $\tilde{A}_i$, potentially due to reverse causality where reduced risk boosts profitability. The credit quality (measured by ALLL to total loans ratio) and the capital-asset ratio do not have significant effects in the regression using initial bank characteristics.

Table 5: Heterogeneity and bank characteristics - baseline sample

	Initial		Average	
	Coef.	s.e.	Coef.	s.e.
Log assets	0.20***	(0.03)	0.19***	(0.03)
Loan frac.	0.28***	(0.10)	0.68***	(0.11)
Capital-assets	-0.65	(0.56)	-2.05***	(0.60)
Loan-assets	0.83***	(0.18)	1.77***	(0.19)
ALLL-loan	2.61	(3.18)	15.01***	(3.45)
Diversification	1.00**	(0.44)	2.79***	(0.50)
Ret. on assets	-30.53***	(9.62)	-36.51***	(11.66)
OCA	-19.34**	(9.05)	-60.19***	(12.30)
Intercept	-3.68***	(0.37)	-4.25***	(0.38)

Notes: Significance levels are indicated by *: 10%, **: 5%, and ***: 1%.

7 Conclusion

This paper proposes a novel semiparametric method based on Bayes' theorem and RV functions. It models heavy tail behavior through a Pareto approximation, while making no parametric assumptions on the relationship between the outcome and covariates outside of the tail region.

The proposed method is particularly useful in panel data models, accounting for unobserved unit-specific heterogeneity. We show that under regularity conditions, our objective function asymptotically aligns with a panel Logit regression on tail observations using $\log X_{it}$ as a regressor. Then, various established econometric techniques could be applicable, which could be convenient for empirical research. Specifically, in panels with large N and small T, the unobserved unit-specific tail thickness and RV functions can be canceled out via the conditional MLE; in panels with large N and large T, bias correction methods (Fernández-Val and Weidner, 2018; Stammann, Heiss, and McFadden, 2016) can be employed to estimate unit-specific parameters and predict unit-specific future outcomes. Furthermore, we also extend our method to dynamic panel data models that incorporate lagged outcomes.

The potential applications of our method could encompass both microeconomics and macroeconomics studies, particularly valuable in light of recent extreme events. For example, one may be interested in analyzing the effect of large sovereign debts on country default risks, or assessing the impact of extreme weather on productivity at the regional, firm, and individual levels.

References

ABREVAYA, J., AND Y.-C. HSU (2021): "Partial effects in non-linear panel data models with correlated random effects," *Econometrics Journal*, 24(3), 519–535.

ADRIAN, T., AND M. K. BRUNNERMEIER (2016): "CoVaR," *American Economic Review*, 106(7), 1705.

ADRIAN, T., F. GRINBERG, N. LIANG, S. MALIK, AND J. YU (2022): "The term structure of growth-at-risk," *American Economic Journal: Macroeconomics*, 14(3), 283–323.

AMEMIYA, T. (1985): *Advanced econometrics*. Harvard university press.

AMISANO, G., AND R. GIACOMINI (2007): "Comparing density forecasts via weighted likelihood ratio tests," *Journal of Business & Economic Statistics*, 25(2), 177–190.

BALTAGI, B. H. (2013): "Panel Data Forecasting," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2B of *Handbooks in Economics*, chap. 18, pp. 995–1024. Elsevier, Amsterdam.

CARRIERO, A., T. E. CLARK, M. MARCELLINO, AND E. MERTENS (2022): "Addressing COVID-19 outliers in BVARs with stochastic volatility," *Review of Economics and Statistics*, pp. 1–38.

CHAMBERLAIN, G. (1980): "Analysis of covariance with qualitative data," *Review of Economic Studies*, 47, 225–238.

CHRISTENSEN, T., H. R. MOON, AND F. SCHORFHEIDE (2020): "Robust forecasting," *arXiv preprint arXiv:2011.03153*.

CLAUSET, A., C. R. SHALIZI, AND M. E. NEWMAN (2009): "Power-law distributions in empirical data," *SIAM review*, 51(4), 661–703.

DAVEZIES, L., X. D'HAULTFOEUILLE, AND L. LAAGE (2024): "Identification and estimation of average marginal effects in fixed effect Logit models," *arXiv preprint arXiv:2105.00879*.

DE HAAN, L., AND A. FERREIRA (2006): *Extreme Value Theory: An Introduction*. Springer.

EINMAHL, J. H. J., AND Y. HE (2023): "Extreme value estimation for heterogeneous Data," *Journal of Business & Economic Statistics*, 41(1), 255–269.

FEDOTENKOV, I. (2020): "A review of more than one hundred Pareto-tail index estimators," *Statistica*, 80(3), 245–299.

FERNÁNDEZ-VAL, I., AND M. WEIDNER (2018): "Fixed effects estimation of large-T panel data models," *Annual Review of Economics*, 10, 109–138.

GABAIX, X. (2009): "Power laws in economics and finance," *Annual Review of Economics*, 1, 255–293.

——— (2016): "Power laws in economics: An introduction," *Journal of Economic Perspectives*, 30(1), 185–206.

GABAIX, X., AND R. IBRAGIMOV (2011): "Rank-1/2: a simple way to improve the OLS estimation of tail exponents," *Journal of Business Economics & Statistics*, 29(1), 24–39.

GIACOMINI, R., S. LEE, AND S. SARPIETRO (2023): "A robust method for microforecasting and estimation of random effects," *arXiv preprint arXiv:2308.01596*.

GOMES, M. I., AND A. GUILLOU (2015): "Extreme value theory and statistics of univariate extremes: A review," *International Statistical Review*, 83(2), 263–292.

GUILLOU, A., AND P. HALL (2001): "A diagnostic for selecting the threshold in extreme value analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 293–305.

HALL, P. (1982): "On some simple estimates of an exponent of regular variation," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(1), 37–42.

HANSEN, B. E. (1994): "Autoregressive conditional density estimation," *International Economic Review*, pp. 705–730.

HILL, B. M. (1975): "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, 3(5), 1163–1174.

HONORÉ, B. E., AND E. KYRIAZIDOU (2000): "Panel data discrete choice models with lagged dependent variables," *Econometrica*, 68(4), 839–874.

HOROWITZ, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, pp. 505–531.

HOROWITZ, J. L., AND N. SAVIN (2001): "Binary response models: Logits, probits and semiparametrics," *Journal of Economic Perspectives*, 15(4), 43–56.

HSIAO, C. (2022): *Analysis of Panel Data*. Cambridge University Press, 4 edn.

JONDEAU, E., AND M. ROCKINGER (2003): "Testing for differences in the tails of stock-market returns," *Journal of Empirical Finance*, 10(5), 559–581.

KLASS, O. S., O. BIHAM, M. LEVY, O. MALCAI, AND S. SOLOMON (2006): "The Forbes 400 and the Pareto wealth distribution," *Economics Letters*, 90(2), 290–295.

KLEIN, R. W., AND R. H. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, pp. 387–421.

LENZA, M., AND G. E. PRIMICERI (2022): "How to estimate a vector autoregression after March 2020," *Journal of Applied Econometrics*, 37(4), 688–699.

LIU, L. (2023): "Density forecasts in panel data models: A semiparametric bayesian perspective," *Journal of Business & Economic Statistics*, 41(2), 349–363.

LIU, L., H. R. MOON, AND F. SCHORFHEIDE (2020): "Forecasting with dynamic panel data models," *Econometrica*, 88(1), 171–201.

——— (2023): "Forecasting with a panel Tobit model," *Quantitative Economics*, 14(1), 117–159.

LIU, L., A. POIRIER, AND J.-L. SHIU (2024): "Identification and estimation of partial effects in nonlinear semiparametric panel models," *Journal of Econometrics*, forthcoming.

MANSKI, C. F. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55(2), 357–362.

MATZKIN, R. L. (1992): "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models," *Econometrica*, pp. 239–270.

NICOLAU, J., P. M. RODRIGUES, AND M. Z. STOYKOV (2023): "Tail index estimation in the presence of covariates: Stock returns' tail risk dynamics," *Journal of Econometrics*, 235(2), 2266–2284.

PLAGBORG-MØLLER, M., L. REICHLIN, G. RICCO, AND T. HASENZAGL (2020): "When is growth at risk?," *Brookings Papers on Economic Activity*, 2020(1), 167–229.

QU, R., A. TIMMERMANN, AND Y. ZHU (2023): "Comparing forecasting performance in cross-sections," *Journal of Econometrics*, 237(2), 105186.

SCHORFHEIDE, F., AND D. SONG (2024): "Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic," *International Journal of Central Banking*, 20(4), 275–320.

SMITH, R. L. (1987): "Estimating tails of probability distributions," *Annals of Statistics*, 15(3), 1174–1207.

STAMMANN, A., F. HEISS, AND D. MCFADDEN (2016): "Estimating fixed effects Logit models with large panel data," .

TIBSHIRANI, R., AND T. HASTIE (1987): "Local likelihood estimation," *Journal of the American Statistical Association*, 82(398), 559–567.

WANG, H., AND C.-L. TSAI (2009): "Tail index regression," *Journal of the American Statistical Association*, 104(487), 1233–1240.

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