Binary Outcome Models with Extreme Covariates: Estimation and Prediction*

Laura Liu
University of Pittsburgh

Yulong Wang
Syracuse University

July 7, 2024

Abstract

This paper presents a novel semiparametric method to study the effects of extreme events on binary outcomes and forecast future outcomes, which is particularly relevant given the recent occurrences of extreme events. Our approach, based on Bayes' theorem and regularly varying (RV) functions, facilitates a Pareto approximation in the tail while allowing for a flexible relationship between covariates and outcomes beyond the tail. We analyze cross-sectional as well as static and dynamic panel data models, incorporate additional covariates in both setups, and accommodate the unobserved unit-specific tail thickness and RV functions in the panel setup. We establish consistency and asymptotic normality of the proposed tail estimator. We demonstrate that, under regularity conditions, our objective function converges to that of a panel Logit regression on tail observations with the log of the extreme covariate as a regressor, which simplifies implementation and facilitates empirical research. We evaluate the finite-sample properties of the proposed tail estimator in Monte Carlo simulations. In the empirical application, we use a panel of small banks to study whether they become riskier when local housing prices experience a significant decline, a channel crucial in the 2007-08 financial crisis.

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

*Liu: Department of Economics, University of Pittsburgh, laura.liu@pitt.edu. Wang: Department of Economics, Syracuse University, ywang402@syr.edu. We thank Bo Honoré, Roger Klein, Ulrich Müller, Frank Schorfheide, and seminar participants at the FRB Chicago, Monash University, University of Melbourne, Reserve Bank of Australia, and University of Sydney, as well as conference participants at the Applied Time Series Econometrics Workshop at the FRB St. Louis, Dolomiti Macro Meetings, AiE Conference and Festschrift in Honor of Joon Y. Park, Midwest Econometrics Group Annual Meeting, Greater New York Econometrics Colloquium, and Women in Macroeconomics Workshop II for helpful comments and discussions. The authors are solely responsible for any remaining errors.

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 events have highlighted the importance of studying the effect of extreme covariates on binary outcomes.

Consider the simplest example with cross-sectional data, where we have a random sample of $\{Y, X\}$ with binary outcome $Y \in \{0, 1\}$ and continuous covariate $X \in \mathbb{R}$. 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 $\partial \pi\left(x\right) / \partial x$, and/or the extreme elasticity[1]

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

where x takes extreme values. Given a random sample with N observations, we characterize the extremeness by letting $x \to \infty$ as $N \to \infty$.[2] Let us 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.

Existing methods face limitations when dealing with extreme covariates, as often seen in economic data that exhibit heavy tails (Gabaix, 2009, 2016). This extremeness in covariates can be related to heavy-tailed error terms (or shocks, as referred to in macroeconomics) in the context of threshold-crossing models: see Proposition 2.2. Parametric methods, such as Logit or Probit, which assume thin-tailed error distributions, can lead to significant misspecification bias, particularly in the tails. Conversely, nonparametric

[1] Note that in the tail analysis, one of the two quantities, $\pi(x)$ and $(1 - \pi(x))$, tends to 0 and while the other tends to 1, so we consider their product $\pi(x)\left(1 - \pi(x)\right)$ instead. Further details are provided in Proposition 2.1.

[2] Here we focus on the right tail without loss of generality. In practice, a similar exercise can be applied to the left tail, allowing for different tail thickness for each tail.

methods, such as kernel or sieve 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. RV functions offer a flexible framework that encompasses a wide range of distributions, making our method more robust to potential misspecification. For heavy-tailed distributions, the RV condition naturally leads to a Pareto approximation in the tail (see Section 2.1 for details), while maintaining flexibility in the relationship between covariates and outcomes beyond the tail region.[3]

Specifically, by Bayes' rule, we have that

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

where $f_{X|Y}(\cdot|y)$ denotes the conditional density function of X given $Y=y$. When x takes extreme values, the RV properties allow us to approximate $1-F_{X|Y}(\cdot|y)$ by $x^{-\alpha^{(y)}}$ up to some constant, where $\alpha^{(y)}$, the Pareto exponent for $y \in \{0,1\}$, is the object of interest. Then, $f_{X|Y}(\cdot|y)$ is approximately proportional to $x^{-\alpha^{(y)}-1}$. Under this Pareto approximation, the extreme elasticity $\delta(x)$ asymptotically 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. We show that, under regularity conditions, our objective function is asymptotically equivalent to that of a panel Logit regression on tail observations, when we treat the log extreme covariate as a regressor. This equivalence simplifies the estimation procedure and makes the method more accessible for empirical studies.

Here we consider two scenarios. First, for panels with 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, non-identically distributed observations, as well as extends Hoadley (1971) to tail analysis. Second, for panels with large T, when

[3]Note that traditional kernel methods focus on local observations close to a specific x of interest, allowing for flexibility far from x. Similarly, our tail estimator focuses on a specific local region, namely the extreme tail where $x \to \infty$, allowing for flexibility in the middle.

the unit-specific tail behaviors or the unit-specific forecast are of interest, we can directly estimate the unit-specific parameter based on existing methods for panel Logit with large T. Finally, we expand our discussion to dynamic panel data models, incorporating lagged outcomes into the analysis.

We assessed 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 study the impact of local housing price declines on the riskiness of small banks, a crucial factor 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 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, while we focus on extreme events, our work builds on a large literature on the binary outcome models, especially those for panel data. However, existing literature typically focuses on mid-sample properties: see Chapter 15 in Wooldridge (2010) for a textbook discussion.

In binary cross-sectional models, various methods have been explored, including parametric approaches (e.g., Logit and Probit), nonparametric approaches (e.g., Matzkin (1992)), and semiparametric approaches, such as a single-index model (Klein and Spady, 1993) and a 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 allowing flexibility in the middle. Notably, much of this existing literature assumes threshold-crossing models, whereas our approach does not require such a structure. For a detailed comparison, see Remark 2.1.1.

In contrast to common methods that directly work with $P(Y = 1|X = x)$, we employ Bayes' theorem to reverse the conditioning set. This transformation is reminiscent of what is used in the discriminant analysis (see Chapter 9.2.8 in Amemiya (1985)), where assuming

normal distributions for $X|Y = y$ yields a quadratic Logit form for $P(Y = 1|X = x)$, as well as in the semiparametric single-index model in Klein and Spady (1993). In our specific context of extreme events, this transformation offers a novel way to disentangle the tail distributions. Unlike their focus on mid-sample properties using the entire sample, we analyze tail properties with only tail observations. This distinction necessitates the development of nonstandard asymptotic theory for the tails.

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. Besides the distinction among parametric, nonparametric, and semiparametric approaches, the binary panel literature can be further divided based on the assumptions made regarding unit-specific heterogeneity. In a random effects setup, the unit-specific heterogeneity is assumed to be drawn from a common underlying distribution. For example, when the distribution is fully parameterized, the common parameters can be estimated via integrated maximum likelihood (see for example, Chamberlain (1980)). In a correlated random effects setup, this underlying 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, which implicitly allows for an arbitrary correlation between the heterogeneity and covariates, without explicitly modeling it. For example, for logistic errors, common parameters can be identified and estimated by eliminating the unit-specific intercept through conditional MLE, as pioneered by Rasch (1960, 1961), Andersen (1970), and Chamberlain (1980); for general errors, Manski (1987) extended the maximum score estimator to the panel data context.

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. The heterogeneity in the RV functions naturally accommodates unit-specific scale parameters as well. In the current context of extreme covariates, we show that, under our Pareto tail approximation based on the RV condition, our objective function is asymptotically equivalent to that of a panel Logit regression on tail observations with the log of the extreme covariate as a regressor. In panels with small T, this equivalence facilitates a comparable analysis based on a similar conditional MLE; in panels with large T, we can use bias correction methods, as developed by Fernández-Val and Weidner (2018) and Stammann, Heiss, and McFadden (2016), to estimate unit-specific parameters and forecast unit-specific future outcomes.

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.[4] In contrast, we focus on binary outcomes and utilize Bayes' representation to introduce a novel estimator in this setting. To incorporate covariates, we build upon the work of Wang and Tsai (2009), significantly extending their theoretical framework to handle panel data and independent but not necessarily identically distributed (i.n.i.d.) random variables. While Hoadley (1971) provides a foundation for MLE asymptotics in the i.n.i.d. case, we extend his framework to tail analysis, where the number of tail observations increases at a slower rate than the total sample size.

Third, our work further contributes to the burgeoning literature on unit-specific forecasts in panel data setups. To the best of our knowledge, this paper is the first study to analyze unit-specific forecasts in panel data settings with binary outcomes and extreme covariates. In linear models, Liu, Moon, and Schorfheide (2020) develop an empirical Bayes approach for point forecasts, while Liu (2023) constructs a full Bayesian approach for density forecasts. In nonlinear models, Christensen, Moon, and Schorfheide (2020) study discrete outcome models and propose efficient robust forecasts, and Liu, Moon, and Schorfheide (2023) examine a Tobit model, which our empirical application builds on. Also see Fosten and Greenaway-McGrevy (2022) for panel nowcasting, 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 not suitable for forecasting under extreme events. Also note that panel data are particularly valuable in these extreme cases, 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, partly spurred by the Covid-19 pandemic and its economic consequences. Broadly, existing approaches generally focusing on time-series data, either excluding outliers (e.g., Schorfheide and Song (2021)) or adapting models to accommodate both outliers and or-

[4]In addition, numerous estimators for the Pareto exponent have been developed, including Smith (1987) and Gabaix and Ibragimov (2011), among many others. See recent reviews by Gomes and Guillou (2015) and Fedotenkov (2020) for over a hundred estimators.

dinary observations within a unified framework (e.g., Carriero, Clark, Marcellino, and Mertens (2022), and Lenza and Primiceri (2022)). Our work differs in two key ways. First, our approach focuses on cross-sectional and panel data, allowing us to exploit information across units.[5] Second, 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.

The remainder of this paper is organized as follows. Section 2 specifies our methodology for cross-sectional data and derives its asymptotic properties. Section 3 presents our estimators in panel data setups. Section 4 extends our estimator to various contexts, such as a dynamic panel data model and a model with time-varying covariates Z_{it}. Section 5 conducts Monte Carlo experiments to study the finite-sample properties of our estimators. Section 6 applies our panel data estimator to examine whether small banks became riskier when local housing prices experienced a significant decline 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 the motivation 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 the general case with additional covariates.

2.1 Baseline models and RV functions

Recall that in the introduction, we introduced the simplest cross-sectional setup with a random sample of binary outcome Y and continuous covariate X. Without loss of generality, let the support of X be $\mathbb{R}^+$ and consider $x \to \infty$ as $N \to \infty$.[6] We focused

[5] While our framework can accommodate dynamic panels under conditional stationarity, extending it to more complicated time-series models like unit roots poses theoretical challenges for extreme value analysis.

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

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 equations (1) and (2).

Heavy-tailed distributions are widely documented in economic and financial data, e.g., Gabaix (2009, 2016). Also, we can assess the presence of heavy tails using various methods in the literature. First, a simple visual inspection using the log-log plot can provide initial insights. By plotting the threshold x against the probability of exceeding x, both on a log scale, we can assess tail behavior: tail observations aligning around a downward-sloping line suggest a heavy-tailed distribution with the slope being approximately $-\alpha$, while a vertical alignment indicates a relatively thin-tailed distribution. Figures 1 and 6 illustrate this for simulated and empirical data, respectively. Note that the exact values of X may or may not be large, depending on the unit, but the log-log plot reveals the relative prevalence of large values within the range of the data. Second, a more rigorous approach involves estimating the tail index, i.e., the Pareto exponent, 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.

The presence of extreme values in the covariate (and error term in the threshold-crossing model in Section 2.1.1) poses significant challenges to analysis and limits the applicability of existing methods. Parametric approaches, such as Logit and probit, may incur substantial misspecification bias in the tail, while nonparametric methods, such as kernel or sieve, may be highly inefficient due to a limited number of tail observations. To overcome these challenges, we proposed a semiparametric approach based on Bayes' theorem and RV functions as follows.

First, we observe extreme values in the covariate X, and our interest lies in understanding the behavior of Y when X takes on extreme values, so we essentially aim to analyze the comovement in the tail between X and Y. Recall the conditional probability defined in equation (1). To facilitate this tail analysis, we can use Bayes' theorem to reverse the conditioning in this probability, resulting in equations (3), reproduced below:

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

Note that this Bayesian representation is simply an alternative characterization of the data. We make no assumptions about the causal direction between X and Y, and this representation allows for any possible causal relationship. A similar Bayes' representation has been employed in the discriminant analysis (Chapter 9.2.8 in Amemiya (1985)), where

normal distributions of $X|Y = y$ leads to a quadratic Logit form of $P(Y = 1|X = x)$. Klein and Spady (1993) also use a similar Bayes' 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. In particular, if $X|Y = y$ exhibits a heavy tail, its distribution can be well approximated by the Pareto distribution, e.g., see Smith (1987). Therefore, we adopt a more general concept of RV functions that results in a Pareto approximation in the tail.[7] Let us first introduce the following definition. For a generic function $g(\cdot)$ with a heavy tail, we say g is RV at infinity with index $-\alpha$ for some $\alpha > 0$, if for all $x > 0$,

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

as $\eta \to \infty$. This is denoted by $g \in RV_{-\alpha}$. Equivalently, by Karamata's characterization theorem, we can write

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

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 $\eta \to \infty$. The parameter α is also referred to as the Pareto exponent, which characterizes the tail thickness of $g(x)$. In particular, a smaller α indicates a heavier tail, i.e., the function $g(x)$ decays to zero more slowly. As α approaches infinity, we approach thin-tailed distributions, such as logistic and normal.[8]

Let $X \in \mathbb{R}^+$ be a generic random variable with cdf $F_X(x)$. If the upper tail probability (or survival function) $1 - F_X \in RV_{-\alpha}$, 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. 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, for the Student-t with v degrees of freedom, α equals v. A list of distributions and their corresponding values of α can be found in Gabaix (2016).

[7] While our method could be extended to the generalized Pareto distribution, encompassing thin-tailed and bounded-support distributions, these cases are often distinguishable given the empirical data. Moreover, estimators based on the generalized Pareto distribution tend to be more involved than our simpler Pareto-based approach.

[8] While the literature has considered cases where $\alpha = 0$ (slowly varying) and $\alpha < 0$ (rapidly varying) (see Appendix B in de Haan and Ferreira (2006)), our focus here is on $\alpha > 0$ for heavy tail distributions.

Additionally, the moments of X are characterized by α: $\mathbb{E}\left[|X|^r\right]$ is finite for $r < \alpha$, while infinite for $r > \alpha$. Note that we can handle cases with very heavy tails where $\alpha \in (0, 1)$ and no moments exist.

Returning to our binary outcome model, we assume that the distribution of X conditional on $Y = y$ satisfies the RV condition (4), that is, for $y \in \{0, 1\}$,

$$1 - F_{X|Y}(\cdot|y) \in RV_{-\alpha^{(y)}}.$$

Three points are worth noting. First, the Pareto exponent $\alpha^{(y)}$ is indexed by y, allowing for different tail thickness for $y \in \{0, 1\}$. Second, if Y were continuous, $\alpha^{(y)}$ would become a function of y, which would be difficult to estimate without parametric assumptions. However, the binary outcome structure simplifies this, with $\alpha^{(y)}$ taking only two distinct values, $\alpha^{(0)}$ and $\alpha^{(1)}$, which can be easily estimated, such as the classic Hill (1975) estimator in this simplest model without additional covariates. Further estimation details are provided in Section 2.2. Finally, this approach is semiparametric, as we allow for flexible relationships 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. Following from Theorem B.1.9(11) in de Haan and Ferreira (2006), if $f_{X|Y}(x|y)$ is non-increasing for sufficiently large x, as $x \to \infty$, $\frac{x f_{X|Y}(x|y)}{1-F_{X|Y}(x|y)} \to \alpha^{(y)}$, and hence

$$\pi(x) \sim \frac{1}{1 + \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{1-F_{X|Y}(x|0)}{1-F_{X|Y}(x|1)}}, \tag{6}$$

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 in equation (5), given some large $\underline{x}^{(y)}$, the term in the denominator becomes

$$\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{1-F_{X|Y}(x|0)}{1-F_{X|Y}(x|1)} \sim \frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{\left(\underline{x}^{(0)}\right)^{\alpha^{(0)}}}{\left(\underline{x}^{(1)}\right)^{\alpha^{(1)}}} x^{\alpha^{(1)}-\alpha^{(0)}}. \tag{7}$$

Then, a consistent estimator of $\pi(x)$ can be obtained by plugging in the estimated $\hat{\alpha}^{(y)}$ and the sample analog of $\frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \approx \frac{N^{(1)}}{N^{(0)}}$, 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 $Y_i = y$.[9]

[9]The theorem for consistency was established in a previous version of this paper and is available upon request.

Remark 2.1 Based on equations (6) and (7), we can define $A = \frac{\mathbb{P}\left(Y=0, X \geq \underline{x}^{(0)}\right)}{\mathbb{P}\left(Y=1, X \geq \underline{x}^{(1)}\right)} \frac{\alpha^{(0)}}{\alpha^{(1)}} \frac{\left(\underline{x}^{(0)}\right)^{\alpha^{(0)}}}{\left(\underline{x}^{(1)}\right)^{\alpha^{(1)}}}$, which can be viewed as a constant that does not change over x. This simplifies the expression for $\pi(x)$ to:

$$\pi(x) \sim \frac{1}{1 + A \cdot x^{\alpha^{(1)} - \alpha^{(0)}}}. \tag{8}$$

Then, an alternative way to estimate $\pi(x)$ would be directly estimating A and $\alpha^* = \alpha^{(1)} - \alpha^{(0)}$ via the MLE.

Furthermore, let $\tilde{A} = -\log A$, then we have

$$\pi(x) \sim \frac{1}{1 + \exp\left(-\tilde{A} + \alpha^* \log x\right)}.$$

This reveals an interesting connection: in this simplified model, our objective function becomes asymptotically equivalent to that of a Logit regression on tail observations, using the log of the extreme covariate as the regressor. Thus, our analysis offers a theoretical justification for the common practice of log-transforming covariates to address extreme values. This insight becomes particularly valuable in panel data cases with unobserved unit-specific heterogeneity: see Remark 3.1.

In terms of the estimation in the cross-sectional case, we slightly prefer the estimator based on equations (6) and (7). First, in the above simplest model, this estimator is generally easier to implement and often does not require an optimization step. Second, when incorporating additional covariates Z (see Section 2.2), the parameter A may become a complicated function of Z, making it difficult to infer.

Remark 2.2 An additional insight from equation (8) is that the component with the heavier tail (i.e., smaller Pareto exponent $\alpha^{(y)}$) will ultimately dominate the conditional probability, that is, as $x \to \infty$,

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

Note that this Pareto approximation represents only the first-order term. Higher-order terms, as detailed in Assumption 2.2, can also contribute to the behavior of $\pi(x)$. Also, from both Monte Carlo simulations and the empirical example, our method provides significant improvement even when $\pi(x)$ is not extremely close to 0 or 1. In practice, it generally works well for values of $\pi(x)$ below 0.2 or above 0.8.

Having estimated $\pi(x)$ from equations (6) and (7), the partial effect $\partial\pi(x)/\partial x$ can be obtained as a by-product through either analytical or numerical differentiation.

Finally, for extremely large x, $\pi(x)$ would approach either 0 or 1, and the partial effect would approach 0. In this case, it would be more informative to examine the extreme elasticity, as defined in equation (2) in the introduction. Interestingly, in the tail, the extreme elasticity is determined solely by the difference between coefficients $\alpha^{(y)}$.

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* $x \to \infty$,

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

Note that the extreme elasticity is inherently non-positive in the tail due to the RV structure.

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. It is worth noting that our method does not require a threshold-crossing structure. The following proposition shows that $\alpha^{(0)} - \alpha^{(1)}$ directly corresponds to the Pareto exponent of ε.

Proposition 2.2 (Threshold-crossing model) *Suppose we have: (a)* $1 - F_X \in RV_{-\alpha_X}$, *and* $1 - F_\varepsilon \in RV_{-\alpha_\varepsilon}$; *(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. With both X and ε exhibiting heavy tails, so we essentially examine the comovement in the tail between X and Y.

Second, the difference between the two conditional Pareto exponents, $\alpha^{(0)}-\alpha^{(1)}$, equals the Pareto exponent for the distribution 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 (i.e., close to the Logit) with $\alpha_\varepsilon \to \infty$ 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, one might attempt to estimate the threshold-crossing model with RV errors ε, but ε is not observable, making it difficult to determine which observations are in the tail. This could significantly bias the results as data in the middle may substantially deviate 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 simplest cross-sectional model in Section 2.1 to incorporate additional covariates, aligning it more closely with empirical research. We then derive the asymptotic properties of our estimator.

Let $Z \in \mathbb{R}^{d_z}$, where d_z is the dimension of Z, be a 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 details could be included as covariates given their potential impact on default risk. Now we rewrite our conditional probability and the Bayes' representation by introducing Z into all conditioning sets:

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

Note that here we hold z constant and let x tend towards 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.

In principle, the dependence of the Pareto exponent $\alpha^{(y)}(\cdot)$ on z could be nonlinear and complex, and could potentially be estimated nonparametrically. However, our focus on the tails requires using only large values of X, rendering nonparametric estimation of $\alpha^{(y)}(\cdot)$ challenging due to data limitations. To sidestep this issue, we assume a pseudo-linear

structure:

$$\alpha^{(y)}(z) = z'\theta^{(y)}, \tag{9}$$

with pseudo-parameters $\theta^{(y)}$ for $y \in \{0,1\}$. This simplification is also consistent with approaches used in the literature, such as Wang and Tsai (2009). Additionally, we require that $z'\theta^{(y)} > 0$ almost surely over z, with sufficient conditions discussed after Assumption 2.2.

From the RV condition, for $y \in \{0,1\}$, given some large threshold $\underline{x}_N^{(y)}$, we have the conditional cdf for values of x exceeding this threshold given by

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

By Theorem B.1.9(11) in de Haan and Ferreira (2006), if the corresponding pdf is non-increasing for sufficiently large x, then it is asymptotically equivalent to a Pareto distribution,

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

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

$$\hat{\theta}^{(y)} = \arg\max_{\theta \in \Theta^{(y)}} \sum_{i=1}^N \left(\log Z_i'\theta - Z_i'\theta \log \frac{X_i}{\underline{x}_N^{(y)}}\right) \mathbf{1}\left\{\Xi_i^{(y)}\right\}. \tag{10}$$

Since the objective function is concave, and the domain is convex (specifically, it could be a convex cone as discussed after Assumption 2.2), this MLE is a convex programming problem that can be easily solved. Also note that this MLE is closely related to the tail index regression proposed by Wang and Tsai (2009), with differences being: first, we separate the data into two subsamples based on $y \in \{0,1\}$; and second, we adopt a pseudo-linear approximation for the Pareto exponent to better align with the panel data case in Section 3, rather than an exponential approximation in their setting.

In a special case without Z, the FOC of equation (10) results in

$$\hat{\alpha}^{(y)} = \left[\frac{1}{N^{(y)}} \sum_{i=1}^N \left(\log X_i - \log \underline{x}_N^{(y)}\right) \mathbf{1}\left\{\Xi_i^{(y)}\right\}\right]^{-1},$$

for $y \in \{0,1\}$. This coincides with the classic Hill (1975) estimator, as the indicator function $\mathbf{1}\left\{\Xi_i^{(y)}\right\}$ effectively selects the largest order statistics.

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

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

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

(b) $0 < \mathbb{P}(Y_i = 1 | Z_i = z) < 1$ *almost surely for* $z \in supp(\mathcal{Z})$.

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

Assumption 2.2 (Cross-sectional data: tail approximation) *For* $y \in \{0,1\}$ *and almost surely for* $z \in supp(\mathcal{Z})$*:*

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

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

as $x \to \infty$*, where functions* $C^{(y)}(z) > 0$, $\left|D^{(y)}(z)\right| \leq \overline{D} < \infty$, $\alpha^{(y)}(z) > 0$ *satisfying equation* (9)*, and* $\beta^{(y)}(z) \geq \underline{\beta} > 0$*, for some constants* $\overline{D}$ *and* $\underline{\beta}$.

(b) The remainder term satisfies that $\sup_z \left\{x^{\beta^{(y)}(z)} o\left(x^{-\beta^{(y)}(z)}\right)\right\} \to 0$*, as* $x \to \infty$.

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

First, the equation in condition (a) implies the RV conditional and resembles equation (2.2) in Wang and Tsai (2009).[10] The first-order term $C^{(y)}(z)\, x^{-\alpha^{(y)}(z)}$ is proportional to a Pareto distribution, while the second-order term $D^{(y)}(z)\, x^{-\beta^{(y)}(z)}$ captures deviations from the ideal Pareto form. This second-order term is essential for characterizing the asymptotic distribution of our estimator. Second, to guarantee the existence of $\theta^{(y)}$ such that $\alpha^{(y)}(z) = z'\theta^{(y)} > 0$ almost surely (i.e., the existence of the convex cone), a sufficient condition is that each element of z has a domain strictly above or below zero. This can be achieved through a monotonic transformation of z, such as the exponential or probability integral transforms. Third, the bounds on $D^{(y)}(z)$ and $\beta^{(y)}(z)$, as well as the supremum condition (b) on the remainder term, allow us to ignore higher order terms

[10] In the absence of additional covariates Z, this equation simplifies to the well-studied second-order condition: see for example, Hall (1982) and Chapter 2 in de Haan and Ferreira (2006).

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 $N^{(y)} = \sum_{i=1}^{N} \mathbf{1}\{\Xi_i^{(y)}\} = \sum_{i=1}^{N} \mathbf{1}\left\{X_i^{(y)} \geq \underline{x}_N^{(y)}, Y_i = y\right\}$ is a random variable, so we need to carefully account for its randomness when applying the LLN and CLT. Then, we further introduce

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

a non-random sequence representing the asymptotic proportion of tail observations for $Y_i = y$. Specifically, $\frac{N^{(y)}}{N} = \xi_N^{(y)}(1 + o_p(1))$, and we will utilize $\xi_N^{(y)}$ to characterize the asymptotic behavior of our estimator.

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

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

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

(c) $\Sigma_N^{(y)} = \mathbb{E}\left[\frac{Z_i Z_i'}{\left(\alpha^{(y)}(Z_i)\right)^2}\middle| \Xi_i^{(y)}\right]$ *is a finite and positive definite matrix, for all N sufficiently large.*

Conditions (a) and (b) jointly impose lower and upper 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) while also accounting for the fact that $N^{(y)}$, for $y \in \{0,1\}$, are random variables.

Theorem 2.1 (Cross-sectional data: parameter estimation) *Let $\theta^{(y)} \in \Theta^{(y)}$, where $\Theta^{(y)}$ is a compact convex subset of $\mathbb{R}^{d_\theta}$ and d_θ is the dimension of $\theta^{(y)}$. Suppose Assumptions 2.1–2.3 hold. Then, for $y \in \{0,1\}$,*

$$\sqrt{N^{(y)}}\left(\Sigma_N^{(y)}\right)^{1/2}\left(\hat{\theta}^{(y)} - \theta^{(y)}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{d_\theta}\right)$$

In addition, $\hat{\theta}^{(1)}$ *and* $\hat{\theta}^{(0)}$ *are asymptotically independent.*

The convergence rate is slower than the standard parametric rate of $\sqrt{N}$, since the tail estimator only accounts for observations in the tail, where the number of tail observations increases at a slower rate than the total number of observations.

To select the threshold $\underline{x}_N^{(y)}$, we follow a common empirical practice that uses empirical quantiles (e.g., 90% or 95%) as potential thresholds, and then checks the chosen threshold through graphic diagnostics, such as a log-log plot. While the literature offers various methods for threshold estimation,[11] our experiments in the Monte Carlo simulations and empirical example 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. If two potential threshold values are close to each other within the downward-sloping region, the choice between the two would only minimally affect the estimated slope.

Given $\hat{\theta}^{(y)}$, we can obtain the conditional probability $\pi(x,z)$ similar to equations (6) and (7) in the simplest model.

$$\pi(x,z) \sim \frac{1}{1+\frac{\mathbb{P}\left(Y=0,X\geq\underline{x}^{(0)}|Z=z\right)}{\mathbb{P}\left(Y=1,X\geq\underline{x}^{(1)}|Z=z\right)}\frac{\alpha^{(0)}(z)}{\alpha^{(1)}(z)}\frac{\left(\underline{x}_N^{(0)}\right)^{\alpha^{(0)}(z)}}{\left(\underline{x}_N^{(1)}\right)^{\alpha^{(1)}(z)}}x^{\alpha^{(1)}(z)-\alpha^{(0)}(z)}}.$$

Due to the data limitation in the tail, a more practical approach might involve parametric estimation of $\mathbb{P}\left(Y=y,X\geq\underline{x}_N^{(y)}|Z=z\right)$. Nevertheless, we omit a detailed discussion here, as we focus on the panel data analysis in Section 3 below, where this term can be entirely canceled out.

Analogous to Proposition 2.1, if we further assume that $f'_{X|Y,Z}(x|y,z)$ is non-increasing for sufficiently large x, we have that 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 $\theta^{(y)}$.

[11] For example, Guillou and Hall (2001) select the threshold by minimizing the asymptotic MSE, and Clauset, Shalizi, and Newman (2009) propose methods based on maximizing the marginal likelihood, as well as minimizing the distance between the power-law model and empirical data.

3 Panel data

Our methodology offers significant advantages in panel data setups, particularly 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, leading to unobserved individual heterogeneity. In the analysis of extreme events, this heterogeneity could manifest as unobserved unit-specific tail thickness and RV functions.[12]

This section focuses on panel data with large N and small T. To build intuition, we begin in Section 3.1 by considering a simplified 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

$$\tilde{\alpha}_i^{(y)} = \alpha^{(y)} + \lambda_i, \tag{11}$$

where $\alpha^{(y)}$ represents the common component depending on $y \in \{0, 1\}$, and λ_i denotes the unit-specific component. For example, Jondeau and Rockinger (2003) show that the tail indices of stock market returns exhibit heterogeneity across regions, and Einmahl and He (2023) find cross-country heterogeneity in the tails of Covid-19 cases and deaths. The additive form helps us 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, $\mathbb{P}_i(\cdot) = \mathbb{P}\left(\cdot\ ;\ \lambda_i, \{\mathcal{L}_i^{(y)}\}\right)$, 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. Similarly, the subscript i

[12] For panel data with observed heterogeneity only, it would be essentially captured by a pooled estimator with observed covariates as the cross-sectional case in Section 2.2.

in $F_{i,X_{it}|Y_{it}}$ below indicates that the distribution is also conditioned on these unit-specific quantities. Note that we are working within a fixed effects framework where $\left\{\lambda_i, \{\mathcal{L}_i^{(y)}\}\right\}$ are considered fixed for each unit i and can be arbitrarily correlated with the covariates of i, which makes the panel data setup richer but 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. This stationarity enables us to eliminate individual effects for tail observations. Furthermore, the conditional independence condition ensures that for $y_1, y_2 \in \{0, 1\}$,

$$\mathbb{P}_i\left(Y_{i1} = y_1, Y_{i2} = y_2 | X_{i1}, X_{i2}\right) = \mathbb{P}_i\left(Y_{i1} = y_1 | X_{i1}\right) \cdot \mathbb{P}_i\left(Y_{i2} = y_2 | X_{i2}\right). \tag{12}$$

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 by assuming that $1 - F_{i,X_{it}|Y_{it}}(\cdot|y) \in RV_{-\tilde{\alpha}_i^{(y)}}$. Equivalently, we can write $1 - F_{i,X_{it}|Y_{it}}(x|y) = x^{-\tilde{\alpha}_i^{(y)}} \mathcal{L}_i^{(y)}(x)$, for some slowly varying function $\mathcal{L}_i^{(y)}(x)$. This slowly varying function does not change over t, given that $\{Y_{it}, X_{it}\}$ are stationary across t. From the RV condition and assuming that $f_{i,X_{it}|Y_{it}}(x|y)$ is non-increasing for sufficiently large x, we have the approximation $\frac{x f_{i,X_{it}|Y_{it}}(x|y)}{1 - F_{i,X_{it}|Y_{it}}(x|y)} \to \tilde{\alpha}_i^{(y)}$. It follows that as $x \to \infty$,

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

Let us look into each term one by one. First, $\frac{x^{-\tilde{\alpha}_i^{(0)}}}{x^{-\tilde{\alpha}_i^{(1)}}} = x^{\alpha^{(1)} - \alpha^{(0)}}$, since λ_i enters additively into the tail indices in equation (11) and can be canceled out. Second, $\frac{\mathcal{L}_i^{(0)}(x)}{\mathcal{L}_i^{(1)}(x)} \to \mathcal{L}_i^*$, which does not depend on x asymptotically due to the slowly varying nature of $L_i^{(y)}(x)$. Third, $\frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}}$ could depend on λ_i in a complicated, possibly nonlinear manner, as λ_i

enters into the conditioning sets. In the end, let $A_i = \frac{\mathbb{P}_i(Y_{it}=0)}{\mathbb{P}_i(Y_{it}=1)} \frac{\tilde{\alpha}_i^{(0)}}{\tilde{\alpha}_i^{(1)}} \mathcal{L}_i^*$. As $x \to \infty$, we can approximate the conditional probability $\pi_i(x)$ as follows

$$\pi_i(x) \sim \frac{1}{1 + A_i \cdot x^{\alpha^{(1)} - \alpha^{(0)}}}, \tag{13}$$

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 event $Y_{i1} + Y_{i2} = 1$. Based on the conditional independence in equation (12), as $x_1, x_2 \to \infty$,

$$\begin{aligned} &\mathbb{P}_i(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 1, X_{i1} = x_1, X_{i2} = x_2) \\ &= \frac{1}{1 + \frac{\mathbb{P}_i(Y_{i1}=0|X_{i1}=x_1)\mathbb{P}_i(Y_{i2}=1|X_{i2}=x_2)}{\mathbb{P}_i(Y_{i1}=1|X_{i1}=x_1)\mathbb{P}_i(Y_{i2}=0|X_{i2}=x_2)}} = \frac{1}{1 + \frac{\pi_i(x_2)(1-\pi_i(x_1))}{\pi_i(x_1)(1-\pi_i(x_2))}} \sim \frac{1}{1 + \left(\frac{x_1}{x_2}\right)^{\alpha^{(1)} - \alpha^{(0)}}}. \end{aligned} \tag{14}$$

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

Further suppose that $f'_{i,X_{it}|Y_{it}}(x|y)$ is non-increasing for sufficiently large x, then $\alpha^* = \alpha^{(1)} - \alpha^{(0)}$ is closely related to the extreme elasticity:

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

as $x \to \infty$. We can estimate α^* by conducting the conditional MLE on tail observations

$$\begin{aligned} \hat{\alpha}^* = \arg\max_{\alpha^*} \prod_{i=1}^{N} &\left(\frac{1}{1 + \left(\frac{X_{i1}}{X_{i2}}\right)^{\alpha^*}} \right)^{Y_{i1}} \left(\frac{\left(\frac{X_{i1}}{X_{i2}}\right)^{\alpha^*}}{1 + \left(\frac{X_{i1}}{X_{i2}}\right)^{\alpha^*}} \right)^{1 - Y_{i1}} \\ &\times \underbrace{\mathbf{1}\left\{Y_{i1} + Y_{i2} = 1, X_{i1} \geq \underline{x}_{1N}, X_{i2} \geq \underline{x}_{2N}\right\}}_{=\mathbf{1}\{\Xi_i\}}, \end{aligned}$$

where $\underline{x}_{tN}$ for $t = 1, 2$ denotes the tail threshold for each time period, such that $\underline{x}_{tN} \to \infty$ as $N \to \infty$. For simplicity, one can set $\underline{x}_{1N} = \underline{x}_{2N} = \underline{x}_N$ in 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:

$$\mathbb{P}(Y_{it} = 1 | X_{it} = x; C_i) = \frac{1}{1 + \exp\left(-(x\beta + C_i)\right)}, \tag{15}$$

and accordingly

$$\mathbb{P}\left(Y_{i1}=1|X_{i1}=x_1,X_{i2}=x_2,Y_{i1}+Y_{i2}=1\right)=\frac{1}{1+\exp\left(-(x_1-x_2)\beta\right)}. \tag{16}$$

Comparing the expressions in (13) and (14) with those in (15) and (16) leads to several 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 uses 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 elucidates the explicit assumptions underlying 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 the coefficient on $\log X_{it}$ serves as the estimate for α^*.

Second, as discussed in Section 2.1, our method adopts a semiparametric approach, focusing on tail behavior. Especially, our approach avoids imposing parametric assumptions on the entire error distribution, making it 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. Extending the baseline model in Section 3.1, we introduce additional covariates Z_i that capture potential observed heterogeneity, so the unit-specific tail thickness becomes

$$\tilde{\alpha}_i^{(y)}(z)=z'\theta^{(y)}+\lambda_i>0, \tag{17}$$

and 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 assemble 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 2.2 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, $Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, Z_i$.

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

(e) $\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i\left[\mathbb{P}_i(Y_{it} = 1 | Z_i) \in (0, 1)\right] > \underline{p}$ *for some* $\underline{p} > 0$, *for all* N *sufficiently large.*

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

$$\mathbb{P}_i\left(Y_{i1} = y_1, Y_{i2} = y_2 | X_{i1}, X_{i2}, Z_i\right) = \mathbb{P}_i\left(Y_{i1} = y_1 | X_{i1}, Z_i\right) \cdot \mathbb{P}_i\left(Y_{i2} = y_2 | X_{i2}, Z_i\right). \quad (18)$$

Condition (e), similar to Assumption 2.1 (b) for cross-sectional models, ensures non-degenerating probabilities. However, it is somewhat weaker than Assumption 2.1 (b): intuitively, while Assumption 2.1 (b) requires non-degenerate probabilities almost surely, condition (e) only requires a non-degenerate proportion of switchers, i.e., units with Y_{it} changing over time. The statement is more complicated due to the need to account for i.n.i.d. observations.

Second, we assume the following tail condition.

Assumption 3.2 (Panel data: tail approximation) *For* $i = 1, \cdots, N$, *for* $y \in \{0, 1\}$, *and almost surely for* $z \in supp(\mathcal{Z}_i)$:

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

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

as $x \to \infty$, *where functions* $C_i^{(y)}\left(z\right) > 0$, $\left|D_i^{(y)}\left(z\right)\right| \leq \overline{D} < \infty$, $\tilde{\alpha}_i^{(y)}\left(z\right) > 0$ *satisfying equation* (17), *and* $\beta_i^{(y)}\left(z\right) \geq \underline{\beta} > 0$, *for some constants* $\overline{D}$ *and* $\underline{\beta}$.

(b) The remainder term satisfies that $\sup_{z,i}\left\{x^{\beta_i^{(y)}(z)}o\left(x^{-\beta_i^{(y)}(z)}\right)\right\}\to 0$*, as* $x\to\infty$*.*

(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. Especially, we can accommodate a unit-specific scale parameter, given by $\left[C_i^{(y)}(Z_i)\right]^{1/\alpha^{(y)}(Z_i)}$, in the Pareto tail approximation. These unit-specific functions will be absorbed into A_i in equation (19) below, and then differenced out for small T or estimated for large T.

Then, $\mathcal{L}_i^{(y)}(x,z)=C_i^{(y)}(z)\left(1+D_i^{(y)}(z)\,x^{-\beta_i^{(y)}(z)}+o\left(x^{-\beta_i^{(y)}(z)}\right)\right)$ 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 $x\to\infty$,

$$\pi_i(x,z)=\mathbb{P}_i\left(Y_{it}=1|X_{it}=x,Z_i\right)\sim\frac{1}{1+A_i\cdot x^{Z_i'\theta^*}}. \tag{19}$$

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:

$$\hat{\theta}^*=\arg\max_{\theta^*\in\Theta}\prod_{i=1}^N\left[\frac{1}{1+\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}}\right]^{Y_{i1}}\left[\frac{\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}}{1+\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}}\right]^{(1-Y_{i1})}\mathbf{1}\{\Xi_i\}, \tag{20}$$

where event $\Xi_i=\{Y_{i1}+Y_{i2}=1,X_{i1}\geq\underline{x}_N,X_{i2}\geq\underline{x}_N\}$, for some threshold $\underline{x}_N\to\infty$ as $N\to\infty$. 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 (see Lemma 1 in the Appendix):

$$\xi_N=\frac{2}{N}\sum_{i=1}^N\mathbb{E}_i\left[\underline{x}_N^{-\tilde{\alpha}_i^{(1)}(Z_i)-\tilde{\alpha}_i^{(0)}(Z_i)}C_i^{(1)}(Z_i)\,C_i^{(0)}(Z_i)\right].$$

Assumption 3.3 (Panel data: estimation) *Let* $\theta^*\in\Theta$*, where* Θ *is a compact convex subset of* $\mathbb{R}^{d_\theta}$ *and* d_θ *is the dimension of* θ^**. For* $y\in\{0,1\}$*, suppose we have:*

(a) $N\xi_N \to \infty$, *as* $N \to \infty$.

(b) Let $M_i^{(y)}(Z_i) = |Z_i|\, C_i^{(1)}(Z_i)\, C_i^{(0)}(Z_i) \left| D_i^{(y)}(Z_i) \right| \left| 1 - \frac{\beta_i^{(y)}(Z_i)}{\tilde{\alpha}_i^{(y)}(Z_i)} \right|$. *Then,*
$\sqrt{\frac{N}{\xi_N}} \cdot \frac{1}{N} \sum_{i=1}^N \mathbb{E}_i \left[\left| \log \frac{X_{i1}}{X_i 2} \right| M_i^{(y)}(Z_i) \underline{x}_N^{-\tilde{\alpha}_i^{(1)}(Z_i) - \tilde{\alpha}_i^{(0)}(Z_i) - \beta_i^{(y)}(Z_i)} \right] \to 0$, *as* $N \to \infty$.

(c) $\Sigma_N = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_i \left[\frac{\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}}{\left(1+\left(\frac{X_{i1}}{X_{i2}}\right)^{Z_i'\theta^*}\right)^2} \left(\log \frac{X_{i1}}{X_{i2}} \right)^2 Z_i Z_i' \middle| \Xi_i \right] \to \Sigma(\theta^*)$, *as* $N \to \infty$, *where* $\Sigma(\theta^*)$ *is a finite positive definite matrix.*

(d) For all i, $\mathbb{E}_i \left[\left| \log \frac{X_{i1}}{X_{i2}} \right|^{2+\kappa} \|Z_i\|^{2+\kappa} \right] < M$ *for some* $\kappa, M > 0$.

This assumption is analogous to Assumption 2.3 for the cross-sectional case, although the notation here is more involved, notably with the $\frac{1}{N}\sum_{i=1}^N(\cdot)$ terms accommodating i.n.i.d. observations across i. Condition (a) ensures that the number of observations contributing to the likelihood tends to infinity as the cross-sectional sample size increases, where ξ_N is the average probability of event Ξ_i given $\underline{x}_N$. Condition (b) introduces undersmoothing to eliminate asymptotic bias. Conditions (c) and (d) guarantee the validity of the Lindeberg-Feller CLT and uniform LLN for i.n.i.d. observations.

Under these assumptions, we establish the asymptotic normality of our estimator for panel data models. The proof builds on Wang and Tsai (2009), with significant extensions to accommodate both panel data structures and i.n.i.d. random variables. See also Hoadley (1971) for MLE asymptotics in the i.n.i.d. case. Here we extend it to tail analysis, where the number of tail observations grows at a slower rate than N.

Theorem 3.1 (Panel data: common parameters) *Let* $N_\Xi = \sum_{i=1}^N \mathbf{1}\{\Xi_i\}$. *Suppose Assumptions 3.1–3.3 hold. Then,*

$$\sqrt{N_\Xi}\Sigma_N^{1/2}\left(\hat{\theta}^* - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{d_\theta}\right).$$

Moreover, if we further have $f'_{i,X_{it}|Y_{it},Z_i}(x|y,z)$ is non-increasing for sufficiently large x, the extreme elasticity is given by

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

which can be consistently estimated by plugging in $\hat{\theta}^*$.

4 Extensions

4.1 Panel data with large T

Our previous panel data analysis has focused on short panels with fixed T. 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. When T is large, we are able to estimate these parameters, although the incidental parameter problem could lead to an asymptotic bias.

Based on equation (19) in Section 3.2, let $\tilde{A}_i = -\log A_i$, and we have that, as $x \to \infty$,

$$\mathbb{P}_i\left(Y_{it} = 1 | X_{it} = x, Z_i\right) \sim \frac{1}{1 + A_i \cdot x^{Z_i'\theta^*}} = \frac{1}{1 + \exp\left(-\tilde{A}_i + \log x \cdot Z_i'\theta^*\right)}. \tag{21}$$

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 satisfy

$$\hat{\theta}^* - \theta^* \overset{a}{\sim} \mathcal{N}\left(\frac{B}{N}, \frac{H^{-1}}{NT}\right).$$

as $N, T \to \infty$, where B denotes the asymptotic bias due to the incidental parameter. Analytic bias correction or jackknife can be applied to eliminate the bias for inference. See also Stammann, Heiss, and McFadden (2016). Once $\{\theta^*, \{\tilde{A}_i\}\}$ are consistently estimated, we can proceed to estimate the average partial effects (APEs) and provide unit-specific forecasts based on estimated $\mathbb{P}_i\left(Y_{it} = 1 | X_{it} = x, Z_i\right)$ as defined in equation (21).

4.2 Panels with time-varying covariates Z_{it}

In empirical studies, the 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 equation (17) becomes $\tilde{\alpha}_i^{(y)}\left(z_{it}\right) = z_{it}'\theta^{(y)} + \lambda_i$, and A_i in equation (19) becomes $A_{it} = \frac{\mathbb{P}_i(Y_{it}=0|Z_{it})}{\mathbb{P}_i(Y_{it}=1|Z_{it})} \frac{\tilde{\alpha}_i^{(0)}(Z_{it})}{\tilde{\alpha}_i^{(1)}(Z_{it})} \mathcal{L}_i^*(Z_{it})$.

Note that we cannot difference out A_{it} as in equation (19). 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}. Without loss of generality,

assuming $T = 2$ for simplicity, the estimator for θ^* is given by

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

where $k_h^{(d_z)}(z_{i1}, z_{i2}) = \frac{1}{h^{d_z}} \prod_{d=1}^{d_z} k\left(\frac{z_{i1,d} - z_{i2,d}}{h} \right)$ indicates a multidimensional kernel with $k(\cdot)$ being 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 dependent variables to capture potential persistence. We begin by examining the case with small T, and then briefly discuss the case with large T.

Without loss of generality, let us assume the first order Markov property: for each $i = 1, \cdots, N$,

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

given $\{\lambda_i, \mathcal{C}_i\}$, where $\mathcal{C}_i$ is a unit-specific assemble of functions in $F_{i,X_{it}|Y_{it},Y_{i,t-1},Z_i}$, similar to the tail approximation in Assumption 2.2. Likewise, $\mathbb{P}_i$ indicates the corresponding probability measure given $\{\lambda_i, \mathcal{C}_i\}$.

The first order Markov property implies the stationarity of the conditional joint distribution $Y_{it}, Y_{i,t-1}, X_{it} \mid Z_i$, which is essential to the following derivation. Recall that previously we applied Bayes' theorem by partitioning the data into two subsets based on $Y_{it} = 0$ and $Y_{it} = 1$. With dynamics, we must now further partition the data according to the transition dynamics. In the first other 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 equation (17), we introduce unit-specific tail thickness as

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

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

To ensure identification, our previous analysis for static panels is equivalent to normalizing $\theta^{(0)} = \mathbf{0}$ and estimating $\theta^* = \theta^{(1)}$. Now, for a similar reason, we normalize

$\theta^{(00)} = \mathbf{0}$. Given the presence of 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. We consider the following four events, which collectively encompass all transition patterns, allowing us to eliminate unit-specific terms as before:

$$E_1 : \mathbf{Y}_i = (0,0,1,1,0)', \quad E_2 : \mathbf{Y}_i = (0,1,1,0,0)',$$
$$E_3 : \mathbf{Y}_i = (1,1,0,0,1)', \quad E_4 : \mathbf{Y}_i = (1,0,0,1,1)'.$$

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, as $x_t \to \infty$ for $t = 1, \cdots, 5$,

$$\mathbb{P}_i\left(E_1 \middle| \cup_{e=1}^4 E_e, \{X_{it} = x_t\}_{t=1}^5, Z_i = z\right) \sim \frac{x_3^{-z'\theta^{(01)}} x_4^{-z'\theta^{(11)}} x_5^{-z'\theta^{(10)}}}{\left(\begin{array}{c} x_3^{-z'\theta^{(01)}} x_4^{-z'\theta^{(11)}} x_5^{-z'\theta^{(10)}} + x_2^{-z'\theta^{(01)}} x_3^{-z'\theta^{(11)}} x_4^{-z'\theta^{(10)}} \\ + x_2^{-z'\theta^{(11)}} x_3^{-z'\theta^{(10)}} x_5^{-z'\theta^{(01)}} + x_2^{-z'\theta^{(10)}} x_4^{-z'\theta^{(01)}} x_5^{-z'\theta^{(11)}} \end{array}\right)}.$$

Subsequently, the conditional MLE for $\left(\theta^{(01)}, \theta^{(10)}, \theta^{(11)}\right)$ can be constructed in a similar manner to the static panel case discussed 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 poses empirical challenges for the data, suggesting it 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 the 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 the estimation performance. This exercise provides intuitive insights into when and how the proposed estimator outperforms the alternatives. Experiment 2 investigates panel data with large N and large T,[13] and focuses on the pseudo out-of-sample forecasting performance, aligning more closely with the empirical example of bank loan charge-off rates.

[13]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.

5.1 Alternative estimators

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

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

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

where $k_h(\cdot)$ is a kernel function with bandwidth h. We employ a Gaussian kernel here, i.e., $k_h(\cdot) = \frac{1}{h}\phi(\frac{\cdot}{h})$, where $\phi(\cdot)$ is the standard normal pdf, 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 and N is the sample size. 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

$$\begin{aligned}
&\hat{\beta}(x) = \arg\max_{\beta} \sum_{i=1}^{N} k_h(X_i - x) \left[Y_i \log p_i(\beta, x) + (1 - Y_i) \log\left(1 - p_i(\beta, x)\right)\right], \\
&\hat{\mathbb{E}}[Y_i | X_i = x] = p_i(\hat{\beta}(x), x).
\end{aligned}$$

For panel data, the Logit estimator is given by

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

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 in finite samples. More specifically, suppose there is a sufficient statistic V_i which could be multi-dimensional, 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 the comparison of 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.[14] As discussed in Section 2.1, the degree of freedom of the t distribution converges to the tail index α asymptotically. Here we consider a range of α values from 0.5 to 2. As elaborated in Proposition 2.2, in the tail, $\alpha^{(0)} \to \alpha_X + \alpha_\varepsilon$ ranges from 1 to 4, $\alpha^{(1)} \to \alpha_X$ varies from 0.5 to 2, and the extreme elasticity $-\left|\alpha^{(1)} - \alpha^{(0)}\right| \to -\alpha_\varepsilon$ spans from -2 to -0.5.[15] From

[14] 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.

[15] In many economic datasets, the tail index α typically falls between 1 and 2. For example, in a review paper, Gabaix (2009) mentioned 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 Newman (2009) remarked that the tail index usually lies between 1 and 2 (note that the α in their notation corresponds to $\alpha - 1$ in ours).

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:	$X_i \sim \lvert t_{\alpha_X}\rvert$, $\alpha_X = 0.5, 1, 1.5, 2$
Error term:	$\varepsilon_i \sim \lvert t_{\alpha_\varepsilon}\rvert$, $\alpha_\varepsilon = 0.5, 1, 1.5, 2$
Sample Size:	$N = 10000$
# Repetitions:	$N_{sim} = 1000$

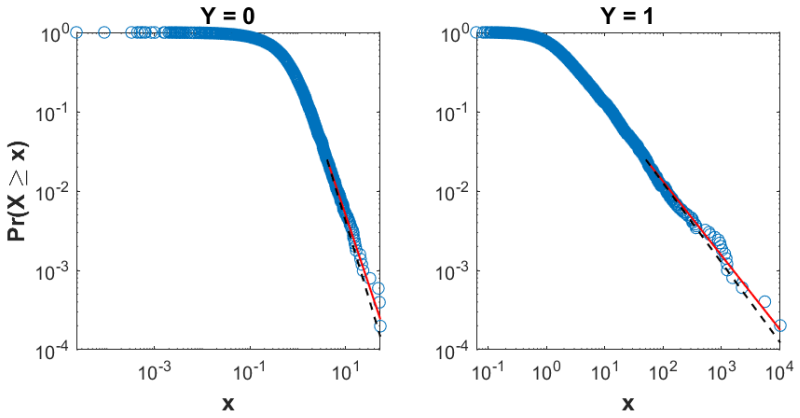
the Monte Carlo setup, we can clearly see that the structure of Bayes' theorem serves not as the data-generating process but as a useful tool to disentangle information from the tail. The sample size of $N = 10000$ is directly comparable with our empirical data sets on bank loan charge-off rates, which 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 classical Hill estimator and often performs well in finite samples. Then, we estimate the conditional probability $\pi(x)$ using the sample analog of equations (6) and (7). We also compare with alternative estimators described in Section 5.1. For both the tail estimator and the Logit estimator with tail observations, we set the cutoffs $\underline{x}_N^{(y)}$ at 97.5% of the distributions of X for $Y = 0$ and 1 separately.[16] In the main text, we plot the comparisons regarding estimated parameters, conditional probability, and extreme elasticity, for the specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$. Comprehensive tables covering all model specifications are provided in the Appendix for detailed reference: see Table 6 for parameter estimation, Table 7 for extreme elasticity, Table 8 for conditional probability, and Table 9 for partial effects. The main messages are similar across different values of tail indices.

Figure 1 depicts a log-log plot from one of the 1000 Monte Carlo simulations, illustrating that tail observations align with a downward sloping line, distinguishing them from the flatter non-tail region. This pattern suggests that the proposed tail estimator would be able to effectively capture the tail behavior. The estimated tail indices (red solid lines) closely match their true asymptotic values (black dashed lines). Figure 2 further shows

[16]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 downward patterns are strong with the slopes remaining stable across a range of cutoffs, which is common in many empirical data as well.

Figure 1: Log-log plot - Experiment 1



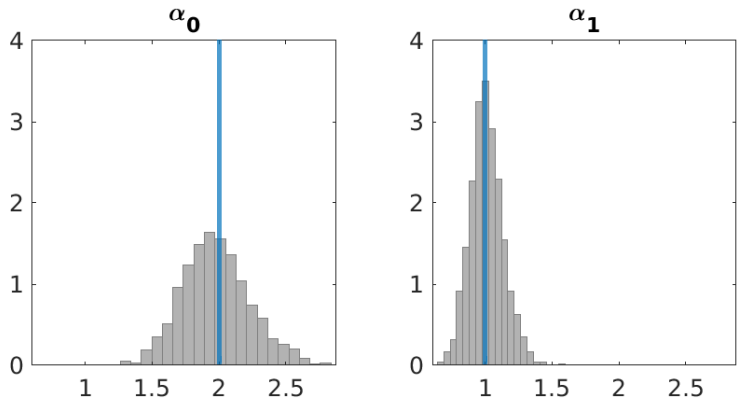
Notes: Based on one sample from specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$. In the tail, $\alpha^{(0)} \to 2$ and $\alpha^{(1)} \to 1$ are indicated by the black dash lines. The estimated $\hat{\alpha}^{(0)}$ and $\hat{\alpha}^{(1)}$ are indicated by the red solid lines.

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, $\alpha^{(1)}$ more precisely estimated than $\alpha^{(0)}$, as the former corresponds to more extreme covariates.

Figure 3 presents the estimated conditional probabilities for X at 90%, 95%, 97.5%, and 99% of its distribution. The true conditional probability $P(Y = 1|X = x)$ is given by the blue vertical lines. As described in the DGPs, these conditional probabilities increase with the evaluation point x. The proposed tail estimator dominates all other alternatives, centered around the true probabilities with low variance as well. The variance of the tail estimator decreases as the evaluation point x increases, reflecting reduced 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 tail 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"

Figure 2: Parameter estimation - Experiment 1



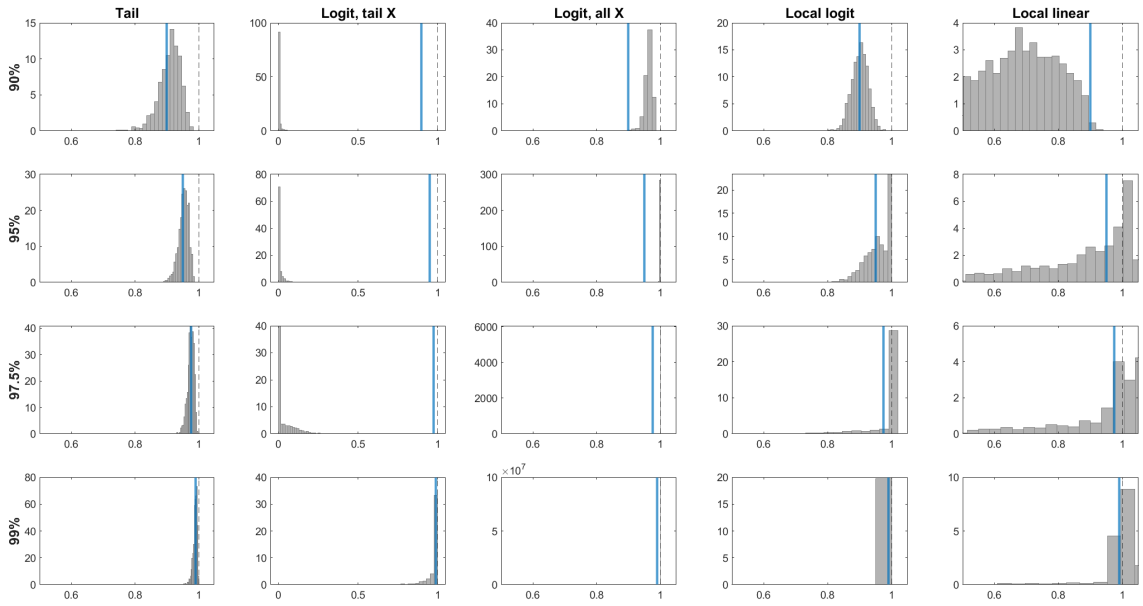
Notes: Specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$, so in the tail, $\alpha^{(0)} \to 2$ and $\alpha^{(1)} \to 1$, indicated by the blue vertical lines.

tends to exhibit a downward bias (except at the 99% percentile). The reason is that given the logistic model's misspecified assumption of a thinner tail, some tail evaluation points can appear relatively moderate compared with the tail observations (with cutoffs at 97.5% of the distributions of X for $Y = 0$ and 1 separately), 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 90% percentile, but it exhibits significant bias and variance at more extreme points. The local linear estimator is even more flexible, and thus yields even larger variance across all these evaluation points. Due to their poor estimation performance and relatively long computation times, we omit these nonparametric estimators from Experiment 2 and the empirical example below, except for the baseline case in Table 4.

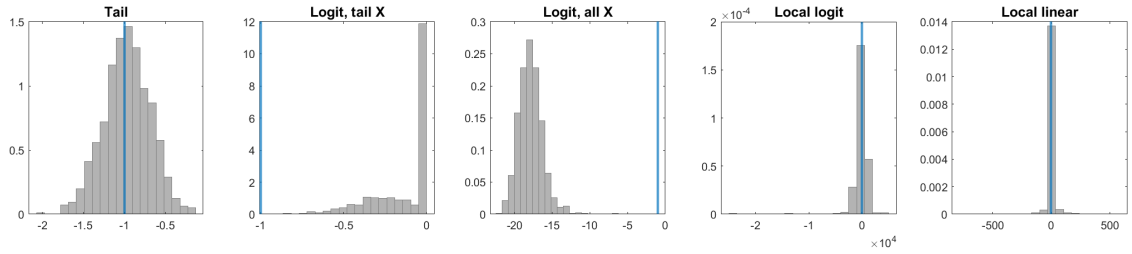
Figure 4 plots the estimated extreme elasticities evaluated at 97.5% of the distribution of X, and the key messages are similar—the Logit estimators induce excessive bias, the nonparametric estimators suffer from high variance, whereas the proposed tail estimator flexibly and efficiently capture the tail behavior and yields the most accurate estimates.

Figure 3: $\hat{P}(Y = 1|X = x)$ - Experiment 1



Notes: Specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$. Each row evaluates the probability at a specific percentile of the distribution of X (90%, 95%, 97.5%, and 99%). The true $P(Y = 1|X = x)$ is indicated by the blue vertical lines.

Figure 4: Extreme elasticity estimation - Experiment 1



Notes: Specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$, so the extreme elasticity $-\left|\alpha^{(1)} - \alpha^{(0)}\right| \approx -\alpha_\varepsilon = -1$, indicated by the blue vertical lines. Evaluated at 97.5% of the distribution of X.

Table 2: Monte Carlo design - Experiment 2

Model:	$Y_{it} = \mathbf{1}\left(X_{it} - \varepsilon_{it} \geq \text{med}_X - \text{med}\varepsilon\right)$
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:	$N_{sim} = 1000$

5.3 Experiment 2: panel data

Experiment 2 examines panel data with large N and large T, emphasizing the 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 modified from the cross-sectional case in Experiment 1 and described in Table 2. Now we incorporate the unit-specific tail thickness λ_i for X_{it}, where the underlying distribution of λ_i is bimodal. For the tail estimator, we resort to the bias corrections as in Fernández-Val and Weidner (2018) and Stammann, Heiss, and McFadden (2016). Further information can be found in Section 4.1. Additionally, details on the alternative estimators are outlined in Section 5.1. The cutoffs for both the tail estimator and "Logit, tail X" are the 90% of the X distributions in this experiment and the empirical analysis below, noting the tail estimator's robustness over a range of cutoffs as discussed in footnote 16.

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 $LPS = \frac{1}{N_f^\dagger} \sum_{i \in \mathcal{N}_f^\dagger} \log \hat{p}(y_{i,T+1}|D)$, 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, which meet the 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 LPS differences, I integrate the tests from Amisano and Giacomini (2007) for density forecasts and Qu, Timmermann, and Zhu (2023) for panel data.

In Table 3, each subpanel's first row presents extreme elasticity estimates from the

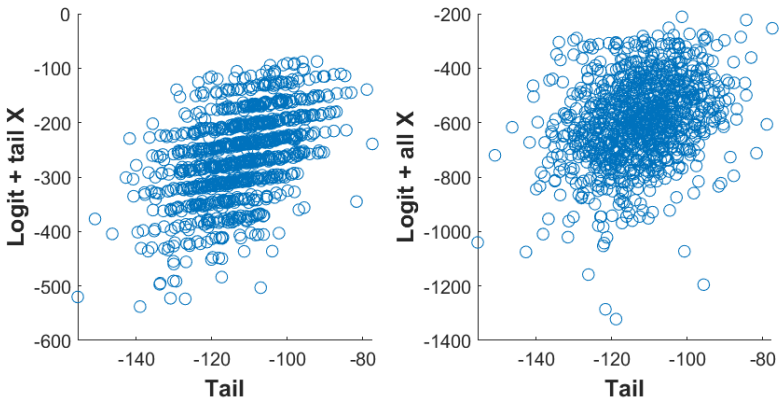
Table 3: Parameter estimation and forecast evaluation - Experiment 2

			$\alpha_X = 1$	$\alpha_X = 1.5$	$\alpha_X = 2$
$\alpha_\varepsilon = 1$	$\hat{\alpha}^*$		-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, with corresponding standard errors across repetitions in the parentheses. The forecasts are assessed by the LPS and a test integrating Amisano and Giacomini (2007) and Qu, Timmermann, and Zhu (2023). 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%.

tail estimator, which are close to the true values, $-\alpha_\varepsilon$. Subsequent rows compare forecast accuracy among estimators and show 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. This is also demonstrated in Figure 5, which displays scatter plots of the LPS from 1000 Monte Carlo repetitions in the setup with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$.

Figure 5: Log predictive score - Experiment 2



Notes: Specification with $\alpha_X = 1$ and $\alpha_\varepsilon = 1$. Each circle represents one Monte Carlo repetition. Note that the x- and y-scales are substantially different.

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 threshold. 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 predominantly 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 equation (21) 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 threshold 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. The extreme regressor X_{it} is given by 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 represents the average quarterly change in the local unemployment rate, accounting for the 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 Federal Reserve Bank of Chicago.[18] We define the local market at the county level, utilizing the annual Summary of Deposits from the Federal Deposit Insurance Corporation to determine the local market for each bank.[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] Finally, the county-level unemployment rates are obtained from the Bureau of Labor Statistics website. The original data are seasonally adjusted at a monthly frequency, and we aggregate them to a quarterly frequency by time averaging.

Our baseline sample focuses on the Residential Real Estate (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 forecast

[17] As the 90th percentiles of X_{it} are positive in our analyzed samples, $\log X_{it}$ is well-defined in the tail.

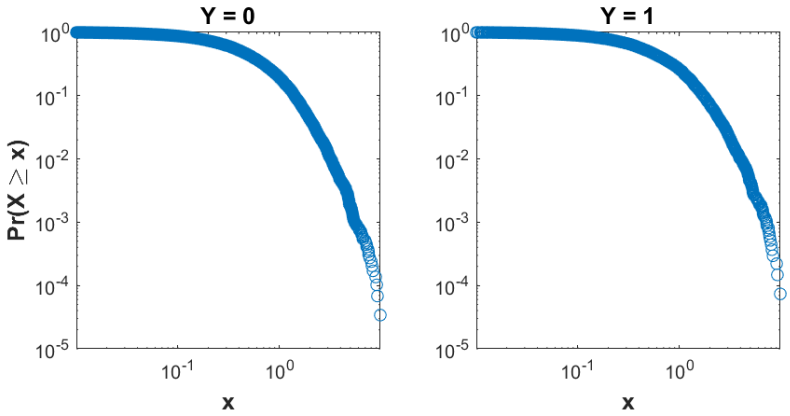
[18] Please refer to Appendix D in Liu, Moon, and Schorfheide (2023) for details on constructing loan charge-off rates from the raw data.

[19] We calculate the deposits received by each bank from every county and link the bank to the county from which it received the highest 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 one that is close to our empirical time periods. We also experimented with the county-level Zillow Home Value Index (ZHVI), and the results are qualitatively similar.

[21] The selection process for a subset of observations from the raw data in each sample is conducted as follows: First, we exclude banks with average non-missing domestic total assets above $1 billion. Second, we exclude banks whose target charge-off rates are missing for all time periods in the sample. In the baseline RRE sample, this process results in the elimination of 583 banks in the first step and an additional 162 banks in the second step.

Figure 6: Log-log plot - banking application, baseline sample



Notes: Baseline sample: RRE, forecasting period = 2009Q4, $T = 40$.

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 main results remain consistent across these setups. Based on the sample statistics in Table 10 for all samples, as well as the log-log plot in Figure 6 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.

6.2 Results

Table 4 compares forecasting performance across estimators. The estimators are similar to those in the Monte Carlo simulation 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 worse than the tail estimator, indicating the importance of carefully addressing the tail behavior. "Logit, all X" ranks third, indicating the presence of significant nonlinearity and possible distinct pattern in the tail compared to the middle range. Local

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

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

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 ***

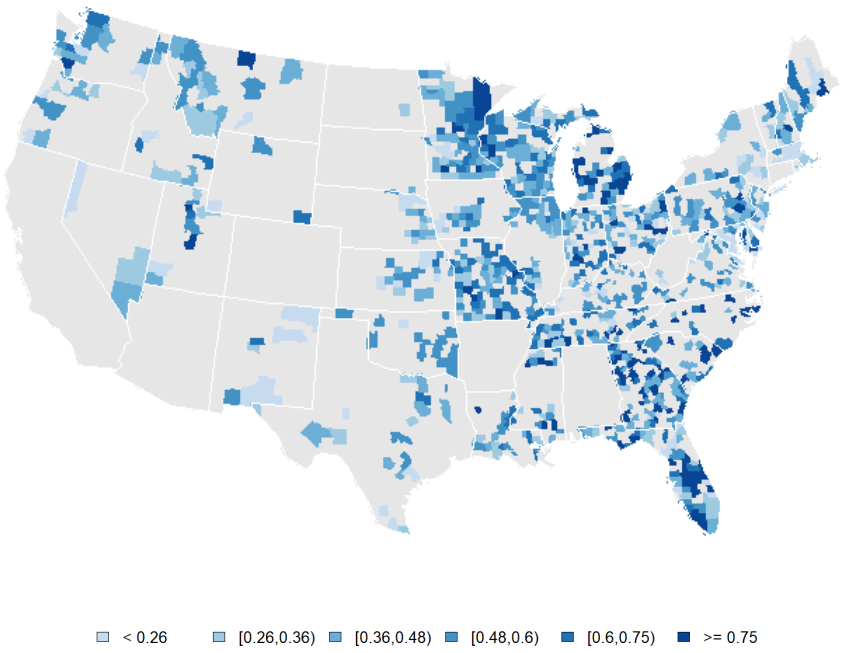
Notes: Forecasting period = 2009Q4, $T = 40$. The forecasts are assessed by the LPS and a test integrating Amisano and Giacomini (2007) and Qu et al. (2020). For the tail estimator, the table reports the exact values of LPS $\cdot N_f^\dagger$. For other estimators, the table reports their differences from the tail estimator. The tests compare other estimators with the tail estimator, with significance levels indicated by *: 10%, **: 5%, and ***: 1%.

Logit yields the least accurate forecasts, suggesting that while it captures nonlinearity, the nonparametric approach could be too noisy with the limited data available in the tail.

Table 11 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 intuitive expectations: higher unemployment increases could directly amplify the impact of a housing price drop on the risk associated with residential real estate loans, whereas changes in unemployment rates may not directly impact the risk profile of non-residential commercial real estate loans. 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, i.e., $Y_{i,T+1} = 1$, for RRE (CRE) loans. Finally, unobserved heterogeneity exhibits a greater dispersion than observed heterogeneity, as the sample variances of the estimate $\tilde{A}_i$ 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 11 and 12 in the Appendix also show that our results are robust with respect

Figure 7: Predictive probability of high risk, average by county, baseline sample



Notes: Baseline sample: RRE, forecasting period = 2009Q4, $T = 40$.

to the threshold $c = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1$,[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 7 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 Z_i, such as variations in local housing prices and local unemployment rates, as well as unobserved heterogeneity $\{\lambda_i, \mathcal{C}_i\}$, which could be captured by $\tilde{A}_i$, where larger values of $\tilde{A}_i$ are associated with higher risk: see equation (21).

Accordingly, in Table 5, we regress the estimated $\tilde{A}_i$ on bank characteristics, following Liu, Moon, and Schorfheide (2023). The "Initial" column uses bank characteristics at

[24] As the threshold 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.

the initial period 1999Q4, ensuring that these regressors are exogenous to subsequent dynamics. The "Average" column uses bank characteristics given by their time averages over the estimation period, incorporating more recent attributes of the banks. 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), or with more diversified earnings (measured by the share of non-interest income to total income) tend to have a greater capacity for assuming riskier RRE loans, reflected in a higher $\tilde{A}_i$. On the other hand, higher profitability (measured by return on assets) and operational efficiency (measured by overhead costs to assets ratio) correspond to a lower $\tilde{A}_i$ and thus reduced risk. 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.

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 maintaining flexibility in the relationship between the outcome and covariates outside of the tail region.

This 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, facilitating the estimation of unit-specific parameters and the prediction of unit-specific future outcomes. Furthermore, we also extend our method to dynamic panel data models with lagged outcomes.

The practical implications of this method potentially span 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 various geo-

Table 5: Heterogeneity and bank characteristics, baseline sample

	Initial	Average
Log assets	0.20***	0.19***
	(0.03)	(0.03)
Loan frac.	0.28***	0.68***
	(0.10)	(0.11)
Capital-assets	-0.65	-2.05***
	(0.56)	(0.60)
Loan-assets	0.83***	1.77***
	(0.18)	(0.19)
ALLL-loan	2.61	15.01***
	(3.18)	(3.45)
Diversification	1.00**	2.79***
	(0.44)	(0.50)
Ret. on assets	-30.53***	-36.51***
	(9.62)	(11.66)
OCA	-19.34**	-60.19***
	(9.05)	(12.30)
Intercept	-3.68***	-4.25***
	(0.37)	(0.38)
Observations	2257	2633
R^2	0.05	0.12
Adjusted R^2	0.05	0.12
F Statistic	15.49***	45.46***

Notes: Baseline sample: RRE, forecasting period = 2009Q4, $T = 40$. Regression of $\tilde{A}_i$ bank characteristics. In the "Initial" column, bank characteristics are given by their values at the initial period 1999Q4. In the "Average" column, bank characteristics are given by their time averages over the estimation sample. Significance levels are indicated by *: 10%, **: 5%, and ***: 1%.

graphic and even individual levels.

References

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.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2), 283–301.

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.

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.

Fosten, J., and R. Greenaway-McGrevy (2022): "Panel data nowcasting," *Econometric Reviews*, 41(7), 675–696.

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 and 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 (Methodological)*, 44(1), 37–42.

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

Hoadley, B. (1971): "Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case," *The Annals of mathematical statistics*, pp. 1977–1991.

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: journal of the Econometric Society*, 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.

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," *arXiv preprint arXiv:2105.12891*.

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: Journal of the Econometric Society*, pp. 239–270.

Qu, R., A. Timmermann, and Y. Zhu (2023): "Comparing forecasting performance in cross-sections," *Journal of econometrics*, 237(2), 105186.

Rasch, G. (1960): *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

——— (1961): "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 4, pp. 321–333.

Schorfheide, F., and D. Song (2021): "Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic," Discussion paper, National Bureau of Economic Research.

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.