A Grand Journey of Statistical Hierarchical Modelling

A Grand Journey of Statistical Hierarchical Modelling

a dissertation presented
by
Ju-Chen Justin Yang
to
The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts
April 2017

Thesis advisor: Professor Neil Shephard Ju-Chen Justin Yang

A Grand Journey of Statistical Hierarchical Modelling

Abstract

This thesis presents three research reports composed by the candidate and his collaborators on different perspectives and applications of statistical hierarchical modelling, which seeks to connect the observed quantities via the assumed and unobserved variables. The topics range from random graph modelling, Shrinkage estimation, to studying tick-by-tick financial data.

In Chapter 1, we discuss the modeling of unlabeled dense network, whose probabilistic property can be uniquely characterized by a sequence of uniformly distributed latent variables that mean the inter-connectivity of each node, and a bivariate function of interests, graphon, that will map each pair of these latent variables into the generative probability of each edge on the graph. A consistent estimation methodology for graphon is proposed.

For the case of simultaneously estimation, studied in Chapter 2, the introduction of an unknown prior distribution over all the parameters that we want to estimate naturally lead to the so-called shrinkage estimation, originally introduced in the frequentist sense. We then take an in-depth look into the theoretical property of these shrinkage estimators under the mean square error in the scenario where independent variables (or covariates) are available to use.

Finally, when studying the dynamics of high-frequency financial data in Chapter 3, assuming an non-observable Poisson random fields whose realizations are thought as pulses or innovations that indeed drive the seen movements of trade price processes, will naturally lead to a new family of theoretically coherent models for discrete-valued stochastic processes that could address different empirical features for market microstructure.

Thesis advisor: Professor Neil Shephard Ju-Chen Justin Yang

In this internet age, hierarchical modelling has become an effective methodology that could help to organize the tremendous amount of data in a logical manner, by pooling information together, reducing idiosyncratic noises, and finally achieving a better estimation (or prediction) that can still generalize to future data.

In addition to the viewpoint from the statistical stability, correctly identifying the pieces that we cannot see and, at the same time, explicitly considering these quantities into our modelling framework, will often lead to a more descriptive statistical model that can better represents different characteristics of the observed data.

Even though these research topics were introduced by the candidates and his collaborators from varied original considerations, it turns out that these seemly disconnected studies could all be seen as different representation and realization of the statistical hierarchical modeling. The candidate is eager to present this grand journey, during which he feels amazed, blessed, and breathtaking.

Contents

Page intentionally left blank.

To my Mother, smiling down in heaven.

Page intentionally left blank.

Acknowledgments

As if it just happened yesterday, I could still feel the excitement and the anxiety, deep inside of my heart, when I was informed my admission into Harvard University. It had always been my dream to study aboard for enlarging my vision and learning with the brilliant minds from all over the world. In February 2012, my dream came true, and a great journey in Harvard began.

The next five years in the Ph.D. program is an unforgettable period of my life while I experienced tremendous changes and lots of twists-and-turns. I stumbled between being a husband, finding my feet in a foreign country, making mistakes in the school, working hard to publish papers, and fighting for the careers, etc. Still, here I am.

Surviving through the sequel of these challenges, my mind has grown stronger much more than what I could have imagined on the first day I stepped into Harvard. To me, the real value I gain out of this Ph.D. journey is less about the knowledge the school has taught me but more about those countless moments to reset, adapt, and progress myself.

Such chance of upgrading myself into a better version is extraordinary and deeply appreciated. However, this process would have not completed or even started if it were not for the support and helps from many people. In the following, I could only address a few of them for the sake of space, but favors from people who I am unable to mention, for sure, will never be forgotten.

First I would like to give my thanks to Prof. Robert W. Chen (Department of Mathematics, University of Miami), Prof. Tzuu-Shuh Chiang (Institute of Mathematics, Academia), and Prof. Shuenn-Jyi Sheu (Department of Mathematics, National Central University), who kindly provided their recommendations during my graduate school application.

Second, I would like to thank Prof. Cheng-Der Fuh (Graduate Institute of Statistics, National

Central University), who not only recommended me into the graduate school but also prepared me the aptitude of conducting theoretical research during my time in National Central University. I could have not been able to start doing research in my G-1 year without his training.

Next, much appreciated is Prof. Samuel Kou. His strong advocate in the admission committee was probably the major reason that I could have the chance to live a new life in Harvard. Even after my admission, his continual support and sincere care for my well-being beyond academics indeed touched me. I'm grateful to have him as my G-1 advisor.

Furthermore, I will give my most thanks to Prof. Neil Shephard, the Chair of Department of Statistics at the time of this writing and my main advisor since G-2 year. Being his first Ph.D. advisee in our department, I'm happy to have worked with him throughout three undergraduate courses and a few other research projects beyond this thesis.

Prof. Shephard has been a well-knownly respected scholar in financial econometrics, but his charming, kindness, empathy, tolerance, and forgiving are what I respect the most. It is my great fortune to have his understanding on my choice of moving into the industry half-way through the G-4 year and to have him as my main advisor during most of my time in Harvard.

Beyond these, I thank many of the coworkers that once contributed their time into many insightful discussions that stimulated my intellectual interests. These include my classmate Xufei Wang, the visiting student Mikkel Bennedsen (Aarhus University), Prof. Pavlos Protopapas from Harvard IACS, and Prof. Luke Born (Simon Fraser University).

Particularly, I would love to thank Prof. Edoardo M. Airoldi, who pushed the submission of my joint research work with him and financially supported me a conference trip. The success of my first submission boosted my confidence toward completing this long Ph.D. journey. It is also my honor to have him in my thesis committees, alone with Prof. Shephard and Prof. Kou.

Finally, the greatest appreciation is devoted to my wife, Carrie, who not only leaves her maternal home, but also takes care of the housework whenever I need to concentrate on academic works, continuously cheers for me whenever I felt frustrated, and always stands by my side wherever I go–from Taiwan, Harvard, to New York. I cannot be luckier to have her love.

My place would have not been here without all the helps from these kind people. My Ph.D. degree and any future achievement based on this experience should give full credits to them. Even through my mother no longer stayed with me, I will continue my journey toward an even brighter future, and my next stage is going to be the whole world, making her even prouder.

Page intentionally left blank.

Stay hungry. Stay foolish.

Steve Jobs

1

Nonparametric Estimation and Testing of Exchangeable Graph Models

Exchangeable graph model (ExGM) is a nonparametric approach for modeling network data that subsumes a number of popular models. The key object that defines an ExGM is often referred to as a graphon, or graph kernel. Here, we make three contributions to refine the theory of graphon estimation. We determine a condition under which a unique canonical representation

This Chapter is advised by and coauthored with Professor Edoardo M. Airoldi with the help of Qiuyi Han.

for a graphon exists and is identifiable. We propose a 3-step procedure to estimate the canonical graphon of any ExGM that satisfies this condition. We then focus on a specific estimator, built using the proposed 3-step procedure, which combines probability matrix estimation using Universal Singular Value Thresholding (USVT) proposed by Chatterjee (2012) and empirical degree sorting using the observed adjacency matrix. We prove that this estimator is consistent in the sense of mean square error. As an application of the theory and the methodology we propose for estimating graphons , we illustrate how they can be used to develop hypothesis testing procedures for models of graphs.

1.1 Introduction

Network data is ubiquitous and research approaches to network modeling have been gaining momentum in the past few years Goldenberg et al. (2009). Applications span a diverse range of scientific areas, from biology and genetics, to social sciences and economics, to the information sciences. These applications raise challenging statistical questions about modeling, inference and computation for network data.

Here, we focus on exchangeable graph model (ExGM) and discuss circumstances under which the graphon that define it can be consistently estimated. A traditional way to formulate this estimation problem is to focus on the probability matrix, which generates the observed adjacency matrix in the sense of independent Bernoulli trials. Several recent papers focus on this direction (Airoldi et al. (2013), Bickel and Chen (2009), Chatterjee (Chatterjee), Choi et al. (2012)) , but one of the deficiencies for this formulation is that the resulting estimate always lacks the global structural information to the generating graphon.

To make this point clear, consider the case that, we have two observed networks which are conjectured to be generated from the same mechanism, but we don't know whether those nodes of one observed network match exactly to nodes of the other network. If we conduct probability matrix estimations–usually under the probability measure given some unknown latent variables which uniquely characterize the nodes we are working on–separately on these two networks, then how can we compare the similarity between these two estimations when we are uncertain about whether the two observed networks share exactly the same set of nodes, that is, when the two estimations might not be coming from the same probability measure?

Therefore, in this Chapter, we would like to adopt an alternative way to formulate the estimation problem for the generating graphon of an ExGM. Our goal in this work is to seek a fully functional form and a nonparametric estimation to the unknown graphon. A functional form of a graphon estimate means it is a function of the latent variables. The qualifier ”fully” above refers to the Aldous-Hoover-Kallenbger characterization theorem Aldous (1981), Hoover (1979), Kallenberg (1989), which uses both of the latent variables and a graphon function to fully specify an ExGM. Precisely, we estimate not only the probability matrix but also the node-specific latent variables and then combine them (in addition to some smoothing techniques) to get a functional estimate for the graphon. So far as we know, this is the first implementable* study focusing on such aspect of the graphon estimation, so hopefully this work could open a door toward more sophisticated statistical tools in this formulation.

We make three contributions in this Chapter. First, we clearly discuss the identifiability issue when pursuing a functional form estimation to the unknown graphon. In other words, before we have a functional form estimate, we need to uniquely define an estimand that is also in a functional form. We especially emphasize an identifiability condition that requires an ExGM to have an absolutely continuous degree proportions distribution, which was originally and vaguely pro-

*The functional form estimation of a graphon hasn't been accounted for by other existing approaches except for the works studied by Bickel et al. (2011) and Wolfe and Olhede (2013). However, their papers don't provide a functional estimation that could be implemented in practice. See Section 1.6.1 for more discussions.

posed by Bickel and Chen (2009) in another format but has been clarified and reformulated by us here. Under this condition, there is always a uniquely defined canonical representation for the generating graphons of an ExGM, which is the primary object we are interested in estimating in this work.

Second, for any ExGM satisfying the identifiability condition, we propose a 3-step procedure to construct a flexible set of nonparametric estimates to the canonical graphon. This procedure requires (i) a probability matrix estimation, (ii) a latent variables estimation by empirical degree sorting using another probability matrix estimation, and finally (iii) an optional smoothing step as needed. This three steps procedure allows any combination of two probability matrix estimations and one smoothing device, so researchers can flexibly design their own nonparametric estimates to the canonical graphon depending on their goal or willingness to accept additional assumptions.

Third, we consider a specific estimator, build with the 3-step procedure, by combining probability matrix estimation using Universal Singular Value Thresholding (USVT) Chatterjee (Chatterjee) and empirical degree sorting using the observed adjacency matrix. This combination, which we informally refer to as the USVT-A estimation, is proved to be consistent in the sense of mean square error, only requiring continuity, or the weaker piecewise continuity, on the true canonical graphon. Here are two advantages of our proposed USVT-A method which we need to especially emphasize: (i) It requires assumptions on the canonical graphon in a least degree, so researchers can almost always use our proposed method as the first step or preliminary analysis to their network data without anxiety. (ii) It is both easy to implement and quick to compute, so a large amount of replications of the estimation calculation is allowable; hence, a simulation based hypothesis testing to the network data becomes plausible.

The rest of this Chapter is organized as follows. In Section 1.2, we discuss in details the identifiability issue to propose a functional form estimate for the generating graphon. In Section 1.3, we demonstrate and explain our 3-step procedure to construct estimates. In Section 1.4, we focus on a specific choice of estimate under the 3-step procedure and provide its theoretical asymptotic consistency. In Section 1.5, We demonstrate the power of pursuing a functional form estimate in the context of classical hypothesis testing. We offer some discussions to related works and conclude in Section 1.6.

1.2 Identifiability of ExGM

Here we define some key notions. We then move on to the discussion of the identifiability of graphons and conclude with a special but flexible subclass of ExGM, which we will focus on in the remainder of this Chapter.

1.2.1 Basic setup

Let $U_1, ..., U_N$ be i.i.d. uniform random variables on the closed interval $[0,1]$, and let $W : [0,1]^2 \to [0,1]$ be an unknown symmetric measurable function. The observed data is an undirected simple graph described by an adjacency matrix A, which is a $N \times N$ symmetric random matrix with binary elements such that, for $\mathcal{U}_N \triangleq \sigma\left(U_1, ..., U_N\right)$,

$$A_{ii} = 0 \text{ for each } i \quad \text{and} \quad A_{ij} | \mathcal{U}_N \backsim \operatorname{Ber}\left(W\left(U_i, U_j\right)\right) \text{ for } i < j,$$

where A_{ij}'s are, conditionally on $\mathcal{U}_N$, independent to each other for $i < j$. The unknown symmetric parameter matrix

$$P_{ii} \triangleq 0 \text{ for each } i \quad \text{and} \quad P_{ij} \triangleq W\left(U_i, U_j\right) \text{ for } i < j$$

is then called the probability matrix.

The model specification that assigns a probability on an undirected graph represented by an infinitely large adjacency matrix containing the observed part A (which has size $N \times N$) is then called the exchangeable graph model (ExGM). W is called a graphon generating this ExGM, and $U_1, U_2, ..., U_N$ are called the latent variables for the observed graph. We will refer the probability distribution on the infinitely large adjacency matrix as $\mathbb{P}$ and simply call the ExGM as $\mathbb{P}$.

Our goal is to draw inferences about the unknown graphon W from the observed adjacency matrix A. Researchers typically formulate the estimation problem as follows:

Estimation problem 1 (P1). Build a probability matrix estimator $\hat{P}$ of P under the probability measure given the latent variables $U_1, ..., U_N$.

Even though this is the most common way to formulate the estimation problem, this approach often leads to an estimator $\hat{P}$ that is unable to describe the global structural information encoded by the generating graphon W, which then blocks us from doing inferences on some interesting and practical problems, like model similarity checking or prediction inferences. Here, we pursue an alternative formulation of the estimation problem as follows:

Estimation problem 2 (P2). Build a functional form or nonparametric estimator $\hat{W}(u, v)$ of $W(u, v)$ under the top probability measure without conditioning.

However, there is an unavoidable well-posedness issue before we further study Estimation problem 2. Due to the highly symmetry structure resulted by the exchangeability of ExGM, several graphons might generate the same[†] ExGM simultaneously, so Estimation problem 2 won't be well-posed unless we can assign a unique and identifiable representation among those graphons generating the same underlying ExGM. We discuss this issue next.

[†] We say two ExGM's $\mathbb{P}_1$ and $\mathbb{P}_2$ are the same if, for any binary and symmetric $N \times N$ matrix realization A and any $N \in \mathbb{N}$, $\mathbb{P}_1(A) = \mathbb{P}_2(A)$.

1.2.2 Identifiability of graphons

The discussion of the identifiability for Estimation problem 2 starts from a non-trivial statement, which seems to be true at a first glance, for any two graphons generating the same ExGM. It is often stated that, for any measure preserving mapping $\varphi : [0,1] \to [0,1]$,

$$W'(u,v) \triangleq W(\varphi(u), \varphi(v)) \tag{1.1}$$

for almost everywhere (a.e.)[‡] $(u,v) \in [0,1]^2$, generates the same ExGM as W. Conversely, for a given ExGM $\mathbb{P}$, is the relationship above the only uncertainty about W? In other words, suppose that both W and W' generate the same ExGM $\mathbb{P}$, does there exist a measure preserving mapping φ such that equation (1.1) holds?

The answer is negative, while the opposite wrong answer has been widely misused in many statistical literatures about ExGM (for example, Bickel and Chen (2009), p. 21069). An easy counterexample proposed by Diaconis and Janson (2008) is

$$W(u,v) = uv \quad \text{and} \quad W'(u,v) \triangleq (2u \mod 1)(2v \mod 1).$$

Then these two graphons will generate the same ExGM but there exists no such a measure preserving mapping φ satisfying equation (1.1). Actually, Theorem 7.1 in Diaconis and Janson (2008) said W and W' will generate the same ExGM if and only if[§] there exist two–rather than one–measure preserving mappings φ and φ' such that

$$W(\varphi(u), \varphi(v)) = W'(\varphi'(u), \varphi'(v))$$

[‡] For $[0,1]^d$ space, we always refer the term almost everywhere with respect to the complete Lebesgue measure on it.

[§] Another equivalent characterization for W and W' generating the same ExGM is

$$\delta_\square(W, W') = 0,$$

where $\delta_\square$ is the so-called cut-metric defined by Borgs et al. (2008).

for a.e. $(u,v) \in [0,1]^2$.

However, this doesn't mean that equation (1.1) should be fully abandoned, for this relationship can still hold among graphons generating the same ExGM and satisfying the following condition:

Twin-free condition. There is no such a pair (u_1, u_2) in $[0,1]$ such that $W(u_1, v) = W(u_2, v)$ for a.e. $v \in [0,1]$.

For any two twin-free graphons W_1 and W_2 generating the same ExGM, Borgs et al. (2010) proved that there is actually a measure preserving bijection $\varphi_{12} : [0,1] \to [0,1]$ such that

$$W_1(u,v) = W_2(\varphi_{12}(u), \varphi_{12}(v))$$

for a.e. $(u,v) \in [0,1]^2$. Thus, for those earlier literatures misusing equation (1.1) as the only uncertainty to graphons generating the same ExGM, they just need to rephrase their results by limiting their interests to a subclass of ExGM generated by a twin-free graphon, which we call twin-freely generated ExGM.

Unfortunately, as this Chapter is being written, there is no known result that states an appropriate way to choose a unique representation for graphons generating a twin-free ExGM. We would rather defer this identifiability issue for a twin-freely generated ExGM into a future study.

What we are going to pursue in the rest of this Chapter is to consider a relatively more restrictive subclass of ExGM, of which the original considerations came from Bickel and Chen (2009) . Their attempts to solve this identifiability issue was to claim that, for any ExGM $\mathbb{P}$ generated by a graphon W, one can find a measure preserving mapping φ such that, for $W_{\text{can}}^{\mathbb{P}} \triangleq W(\varphi(u), \varphi(v))$,

$$g_{\text{can}}^{\mathbb{P}}(u) \triangleq \int_0^1 W_{\text{can}}^{\mathbb{P}}(u,v)\,dv$$

is monotone non-decreasing for $u \in [0,1]$. They also argue that the so-called canonical form $W_{\text{can}}^{\mathbb{P}}$

of the graphon W is uniquely determined for a.e. $(u, v) \in [0,1]^2$.

Nevertheless, their arguments to show the uniqueness of $W_{\text{can}}^{\mathbb{P}}$ for a given ExGM $\mathbb{P}$ is incomplete[¶] without assuming the following condition:

Degree-identifiable condition. Let U be a uniform random variable on $[0,1]$. Then the degree proportion

$$g(U) \triangleq \int_0^1 W(U, v)\, dv$$

is an absolutely continuous random variable[‖] on $[0,1]$.

For any degree-identifiable graphon W generating the ExGM $\mathbb{P}$, a modification to the arguments provided by Bickel and Chen (2009) can show that there are φ, $W_{\text{can}}^{\mathbb{P}}$, and $g_{\text{can}}^{\mathbb{P}}$ defined as above such that $g_{\text{can}}^{\mathbb{P}}$ is strictly increasing on $[0,1]$, which was indeed the hidden assumption made in Bickel and Chen (2009) but stated clearly in their later work Bickel et al. (2011).

Therefore, our Estimation problem 2 will be well-posed if we only focus on a subclass of ExGM generated by a degree-identifiable graphon, which we call degree-identifiable ExGM, and if we treat the uniquely defined canonical graphon $W_{\text{can}}^{\mathbb{P}}$ associated with a degree-identifiable ExGM $\mathbb{P}$ as the major estimand of interest. The next Section will discuss an estimation procedure in this context.

Remark 1. *Starting from the next Section, we will simply write W and g to refer the canonical graphon of a degree-identifiable ExGM and its marginal integral.*

[¶] An easy example to check the incompleteness of their arguments is the following two graphons

$$W(u, v) \triangleq 1_{[0,1/2]^2}(u, v) + 1_{[1/2,1]^2}(u, v),$$
$$W'(u, v) \triangleq 1_{[0,1/2]\times[1/2,1]}(u, v) + 1_{[1/2,1]\times[0,1/2]}(u, v),$$

which give monotone non-decreasing $g(u) \equiv g'(u) \equiv 1/2$, generate a same ExGM, yet are different for a.e. $(u, v) \in [0,1]^2$. There is no canonical choice between W and W' in this example.

[‖] We should note that the random variable $g(U)$ here is uniquely determined by the ExGM $\mathbb{P}$ in the distribution sense.

Remark 2. *There are actually three equivalent characterizations for a degree-identifiable ExGM* $\mathbb{P}$*:*

- $g\left(U\right)$ *is an absolutely continuous random variable;*
- $g_{\text{can}}^{\mathbb{P}}$ *is strictly increasing on* $[0,1]$*;*
- *The cumulative distribution function (CDF) of* $g\left(U\right)$ *is absolutely continuous and hence is continuous.*

1.3 Three-Step Estimation of Degree-Identifiable ExGMs

In this Section, we will explain how, in a 3-step procedure, to construct a flexible class of functional form or nonparametric estimates for the canonical graphon generating a degree-identifiable ExGM. Then we will conclude with a special choice of nonparametric estimate.

The main idea behind the estimation procedure is to exploit the degree-identifiability feature of the canonical graphon and make use of empirical degree sorting to infer unknown latent variables. We now describe how to proceed this 3-step procedure in the following paragraphs.

Step 1: Probability matrix estimation. Conduct any P1 estimation $\hat{P}$ for the probability matrix P.

Step 2: Latent variables estimation. Construct an empirical CDF of degree proportions using another P1 estimation $\hat{P}'$, which may or may not be the same as $\hat{P}$, and then let $\hat{U}_i$'s be the estimators of the unknown latent variables U_i's defined as the values of the empirical CDF evaluating at the degree proportions of i-th node in $\hat{P}'$.

The rationale of doing this Step 2 is explained here. According to some simulation evidences, the empirical CDF $\hat{F}\left(x\right)$ of degree proportions seems to describe the CDF of $g\left(U\right)$, which we denote as $g^{-1}(x)$, quite accurately when the number of nodes N is large enough. On the other hand, the law of large numbers can somehow guarantee that the degree proportions in $\hat{P}'$ at i-th node, $\dfrac{1}{N}\sum_{j=1}^{N}\hat{P}'_{ij}$, will be a good approximation to $g(U_i)$ (assuming that given U_i, $\hat{P}'_{ij}$'s are roughly i.i.d. from the distribution $W(U_i, U)$). As for a degree-identifiable ExGM, which requires the canonical graphon marginal integral g to be strictly increasing, we must have $u = g^{-1}(g(u))$ for every $u \in [0,1]$, so we can trust the estimation of the latent variables $U_i = g^{-1}(g(U_i))$ by $\hat{U}_i \triangleq \hat{F}\left(\dfrac{1}{N}\sum_{j=1}^{N}\hat{P}'_{ij}\right)$.

Remark 3. *In the descriptions above, we temporarily assume that there is no over-lapping for degree proportions in $\hat{P}'$, i.e., $\left\{\dfrac{1}{N}\sum_{j=1}^{N}\hat{P}'_{ij}\right\}_{i=1}^{N}$ are distinct. We will solve this over-lapping issue later after we have a specific choice for $\hat{P}'$.*

Once we have conducted Step 1 and 2, we can start to construct a functional form estimate $\hat{W}(u,v)$. For now, we already have a set of three dimensional points

$$\left(\hat{U}_i, \hat{U}_j, \hat{W}\left(\hat{U}_i, \hat{U}_j\right)\right) \triangleq \left(\hat{U}_i, \hat{U}_j, \hat{P}_{ij}\right),$$

which we should treat as a noisy realization[**] of the unknown canonical graphon plane at $(U_i, U_j, W(U_i, U_j))$. To build a functional form estimation $\hat{W}(u,v)$ from those three dimensional points, we can either use a linear interpolation or a stepwise approximation as the pre-smoothed estimate. We majorly focus on the later one in this study, so the pre-smoothed estimate now takes the form of a step function

$$\hat{W}(u,v) \triangleq \sum_{1\le i,j\le N} \hat{P}_{ij} 1_{\left(\hat{U}_i - 1/N, \hat{U}_i\right] \times \left(\hat{U}_j - 1/N, \hat{U}_j\right]}(u,v).$$

Step 3: Smoothing. (This step is optional.) Apply any smoothing algorithm on the pre-smoothed estimate to get a smoothed estimate, which may or may not be in a form of step func-

[**] The noises here are coming into not only the z-direction but also the xy-directions.

tion.

Here are several notes related to this Step 3. First, it's a an optional step, and the choice of whether to include this step or not and how to conduct it depends on researchers' ultimate goal for inferences on network data and their willingness to accept those unavoidably additional assumptions on the canonical graphon. A detailed investigation of adding this third step in the estimation of canonical graphon is separately discussed in another paper of the third author Chan and Airoldi (2014), in which they alternatively construct a graphon estimator using our 3-step procedure by choosing the native adjacency matrix A itself for both Step 1 and 2 and the total variation minimization (TVM) smoothing Chan et al. (2011) for Step 3. We will illustrate our proposed graphon estimator build by the 3-step procedure in the following subsection.

1.3.1 USVT-A Estimation

Because both of the Step 1 and 2 above can take any kind of P1 estimator to proceed, we need to know how to explicitly specify $\hat{P}$ and $\hat{P}'$. In this Chapter, we especially limit our choice for the two P1 estimations to be either the method of Universal Singular Value Thresholding (USVT) proposed by Chatterjee (Chatterjee) or the adjacency matrix A itself. This limitation only reflects the authors' personal favor by the time of writing this Chapter, but it would be interesting to test on different combinations on Step 1 and 2 in the future study. We describe the USVT method in the following paragraph.

Universal Singular Value Thresholding (USVT) Let $\sum_{i=1}^{N} s_i u_i u_i^T$ be the singular value decomposition of the adjacency matrix A. Then the USVT estimation $\hat{P}$ of the probability matrix P is defined by (i) thresholding on the singular values to get $\hat{M} \triangleq \sum_{i \ni \left\{s_i \geq 1.01\sqrt{N}\right\}} s_i u_i u_i^T$ and (ii) capping those extreme values of elements in $\hat{M}$ that are either greater than 1 or smaller than

0 to get $\hat{P}_{ij} \triangleq \left(\left(\hat{M}_{ij}\right) \wedge 1\right) \vee 0$.

This spectral method proposed by Chatterjee (Chatterjee) is a handy and general way for dealing with Estimation problem 1 for the graphon. It requires only one observation network, can be easily implemented with a tremendous speed, and assume almost non-criterions on the underlying graphon except for its measurability. When the observation is large enough, USVT method provides a promising estimation result. Hence, USVT becomes a natural choice for $\hat{P}$ in the Step 1 of our 3-step procedure.

However, since it's still addressing on estimation problem in a P1 formulation, only the probability matrix $P_{ij} = W\left(U_i, U_j\right)$ can be written down–which depends on the unknown node specific latent variables U_i's. Thus, when two networks with possibly different sets of nodes (and hence possibly with different sets of latent variables) are considered, we cannot simply compare their own USVT estimations because of the uncertainty to their latent variables. We want to have a unifying framework, independent of those latent variables, that would be able to compare the generating mechanisms from multiple networks with possibly different sets of nodes, so this demand finally drives us to seek a canonical representation to a graphon and design a functional form estimation for it.

To achieve the goal stated above, we will need to choose a $\hat{P}'$ for Step 2 in the 3-step procedure to conduct an empirical degree sorting to estimate the latent variables and then get a presmoothed estimate $\hat{W}$. It seems to be natural to use USVT in this step again, but the comparative simulation study in the next Subsection suggests the use of the vanilla adjacency A for $\hat{P}'$.

For Step 3, the TVM smoothing[††] discussed by Chan et al. (2011) is tested but won't be used in our final proposal for the graphon estimation, because, as shown in the next Subsection, the error reduction from using their method seems to be mild. Thus, for this Chapter, our final pro-

[††] We want to emphasize here that this smoothing method still gives a step functional estimate, but other smoothing methods like spline are still open to be tested.

posal of graphon estimator is to use USVT for Step 1 and the adjacent matrix A for Step 2 (plus a careful definition to avoid degree over-lapping–see Theorem 3). We informally call it USVT-A estimation.

1.3.2 Comparative Simulation Study

In this Subsection, we demonstrate two simulations showing the performance of different combinations of graphon estimations constructed from the 3-step procedure. In each simulation, we calculate the root of the mean square error (RMSE) between the constructed estimator (using N ranging from 300 to 3000) and the true graphon, where only two cases are considered here: the quadratic graphon $W\left(u,v\right)=\left(u^2+v^2\right)/4$ and the logistic graphon $W\left(u,v\right)=\mathrm{logistic}\left(-5+5\left(u+v\right)\right)$, where $\mathrm{logistic}\left(x\right)\triangleq\left(1+\exp\left(-x\right)\right)^{-1}$. Our results are shown in Table 1 and 2.

In both of the two Tables, the rule of calling one combination is like ”Step 1 method”-”Step 2 method”-”Step 3 method”, while the last step is optional–for example, USVT-A-TVM method stands for using USVT in Step 1 for probability matrix estimation, using the plain adjacency matrix A in Step 2 for the empirical degree sorting, and finally using the TVM smoothing in Step 3. To have a clear contrast, we also include the worst combination A-A as a base line estimation.

From the two Tables, we see that, for Step 1, using USVT is clearly better than using the vanilla A; for Step 2, sorting according to USVT estimate gives approximately the same result as (sometimes worse than) sorting according to the plain A; for Step 3, TVM smoothing can be helpful and reduce some mean square errors. It’s interesting that A-A-TVM method gives a fairly as good performance as both USVT-USVT-TVM and USVT-A-TVM methods, so this also motivates the third author’s another work Chan and Airoldi (2014). They separately discuss the relevant theoretical property of A-A-TVM method and call it Sort-and-Smooth (SAS) algorithm.

Even so adding a third smoothing step is helpful in these two specific examples, we note that

N	300	900	1500
A-A	0.344657	0.359469	0.357767
USVT-USVT	0.035505	0.024235	0.017006
USVT-A	0.037397	0.024614	0.017132
A-A-TVM	0.02453	0.013044	0.009418
USVT-USVT-TVM	0.040357	0.01202	0.008492
USVT-A-TVM	0.040601	0.011967	0.008526
N	1800	2400	3000
A-A	0.351921	0.360604	0.358457
USVT-USVT	0.01695	0.013922	0.012097
USVT-A	0.017059	0.013995	0.012168
A-A-TV	0.010491	0.00895	0.007025
USVT-USVT-TV	0.009912	0.00813	0.006128
USVT-A-TV	0.009926	0.008101	0.006066

Table 1.1: RMSE Simulation for Quadratic Graphon

N	300	900	1500
A-A	0.38003	0.3812	0.383409
USVT-USVT	0.102721	0.065759	0.034483
USVT-A	0.106061	0.069602	0.034192
A-A-TVM	0.085428	0.034208	0.02423
USVT-USVT-TVM	0.084122	0.051107	0.023318
USVT-A-TVM	0.075824	0.043535	0.02324
N	1800	2400	3000
A-A	0.380116	0.379474	0.379676
USVT-USVT	0.03164	0.024988	0.019176
USVT-A	0.031406	0.024771	0.019082
A-A-TVM	0.023198	0.019551	0.014232
USVT-USVT-TVM	0.023003	0.019187	0.014117
USVT-A-TVM	0.022963	0.019178	0.014113

Table 1.2: RMSE Simulation of Logistic Graphon

Chan and Airoldi (2014) actually requires more smoothness assumption on the underlying canonical graphon W and hence limits the availability of their method. Furthermore, we note that USVT-A estimation has only slightly bigger RMSE than SAS method, but its decreasing rate on the RMSE as the number of nodes increasing is the same as the decreasing rate of SAS method. Therefore, we are going to pursue the USVT-A method in this Chapter, which has a fairly good performance when compared with SAS algorithm but owns more clear theoretical property with less constraints on the graphon. In the remainder of the Chapter, we focus on the USVT-A estimation in order to seek the least assumptions on W. Its theoretical property is discussed in the next Section.

1.4 Consistency

In this Section, we will talk about the theoretical consistency of the USVT-A estimation, which is defined through a combination of probability matrix estimation using USVT and the latent variables estimation using the empirical CDF of observed degree proportions in A. The main theoretical result of this Chapter is as follows:

Theorem 1 (USVT-A Consistency)**.** *Assume that W is the canonical graphon of a degree-identifiable ExGM. If W is continuous on $[0,1]^2$, then the $\hat{W}$ constructed by the USVT-A method is consistent for estimating W in the sense that*

$$\mathbb{E}\left(\int_0^1\int_0^1\left(\hat{W}\left(u,v\right)-W\left(u,v\right)\right)^2 dudv\right)\to 0.$$

Here are two cornerstones that make our main result hold, both of which correspond to the consistency of Step 1 and Step 2 in our proposed 3-step procedure in Section 1.3.

Theorem 2 (USVT Consistency Chatterjee (Chatterjee))**.** *Let $\hat{P}$ be the USVT estimation of* *probability matrix P. Then*

$$\mathbb{E}\left(\frac{1}{N^2}\sum_{i,j=1}^{N}\left|\hat{P}_{ij}-P_{ij}\right|^2\right)\to 0. \tag{1.2}$$

Here are some notations we are going to use throughout this Section. Let the observed degree proportions in A to be

$$D_i \triangleq \frac{1}{N}\sum_{j=1}^{N}A_{ij},$$

with their empirical CDF defined as

$$\hat{F}\left(x\right) \triangleq \frac{1}{N}\sum_{i=1}^{N}1_{\{D_i\leq x\}}.$$

The following Theorem describes the consistency of latent variables estimation via empirical degree sorting using the observed adjacency matrix A.

Theorem 3 (Degree Sorting Consistency)**.** *Let the empirical degree sorting estimate of the latent variables to be*

$$\hat{U}_i \triangleq \hat{F}\left(D_i\right), \tag{1.3}$$

but to avoid the over-lapping issues we will instead use

$$\tilde{U}_i \triangleq \hat{U}_i - \frac{\kappa_i - 1}{N} \tag{1.4}$$

in the proposed USVT-A estimation, where κ_i, given all of $\hat{U}_i$, is jointly distributed as follows: let $C_1,...,C_M$ be those unique values of D_i's, and, if $D_{i_1}=\cdots=D_{i_{k_m}}=C_m$, then $\kappa_{i_1},...,\kappa_{i_{k_m}}$ are a uniform resampling of the set $\{1,2,...,k_m\}$. Then we have, for each $i=1,...,N$, $\left|U_i-\hat{U}_i\right|\to 0$ and $\left|\tilde{U}_i-\hat{U}_i\right|\to 0$ in probability, and hence $\left|U_i-\tilde{U}_i\right|\to 0$ in probability.

A complete proof for Theorem 1 will make use of the two key consistency results described above and be found in Appendix.

1.5 Hypothesis Testing

In this Section, we illustrate how the proposed USVT-A estimator can help to develop a classical statistical inference procedure—specifically, hypothesis testing—in the burgeoning analysis of network data. There is ample room for improvement of the procedure we describe here.

Hypothesis testing is a powerful procedure with limited literature in network data analysis. Olding and Wolfe (2009) presents likelihood ratio test on three examples–(i) Erdős-Rényi graph, (ii) stochastic blockmodel, and (iii) the degree fixed test–in a ground-breaking favor. However, such a method lacks the flexibility to cope with more sophisticated models, such as exchangeable graph models.

There are mainly two reasons why it is difficult to extend classical hypothesis testing theory to network data. The first is that modeling network data often involves latent variables. In case of ExGM, the U_i's are especially hard to handle. The second reason is the high computational cost of fitting existing methods, so it is often untenable to get the sampling distribution of the test statistic under the null hypothesis from simulations. Instead, the proposed USVT-A estimation captures the structure of ExGM by design and is also so computationally efficient that a large number of Monte Carlo replications can be employed for obtaining the sampling distribution under the null hypothesis. Two illustrative examples are discussed in the following two Subsections.

1.5.1 Simple Null Hypothesis with Quadratic Graphon

Suppose that we observe network data represented by an adjacency matrix A, which is generated by a degree-identifiable ExGM with canonical graphon W. We want to test the two hypothesis: for $W_Q(u,v) \triangleq \frac{1}{4}(u^2+v^2)$,

$$H_0 : W(u,v) = W_Q(u,v) \text{ versus } H_a : W(u,v) \neq W_Q(u,v).$$

By Theorem 1, we will have the USVT-A estimate $\hat{W}$ is getting closer and closer in the sense of mean square errors to the true canonical graphon W when N is sufficiently large. Thus we can choose the test statistic to be

$$T \triangleq \sqrt{\int_0^1 \int_0^1 \left| W_Q\left(u,v\right) - \hat{W}\left(u,v\right) \right|^2 dudv} \triangleq \left\| W_Q - \hat{W} \right\|, \tag{1.5}$$

the L^2 distance between W_Q and $\hat{W}$. Although we can't analytically know the sampling distribution of T under the null hypothesis H_0, we can easily approximate it using a large amount of simulations because of the fast implementation of our USVT-A method. Using 5000 Monte Carlo samples for $N = 3000$, we get the histogram of T as in Figure 1. We can see that the sampling distribution of T under H_0 is right skewed. The mean and standard deviation of the Monte Carlo samples are 0.0115 and 5.656×10^{-4}, and the 95% quantile is 0.0126, so the rejection region is $T \geq 0.0126$.

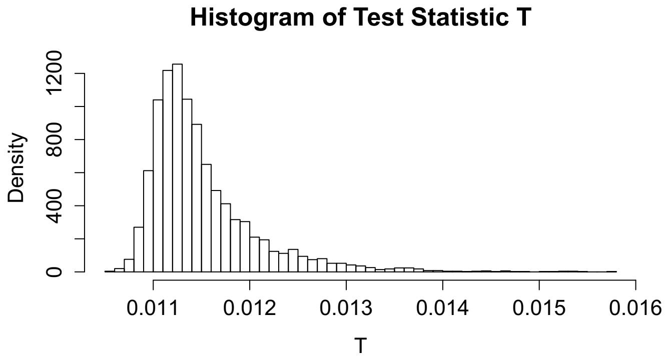


Figure 1.1: 5000 Monte Carlo draws of T under H_0 for quadratic graphon.

1.5.2 Composite Null Hypothesis with Logistic Graphon

In this Subsection, we will consider the testing of a composite hypothesis that whether a data generated by a degree-identifiable ExGM actually has a canonical graphon in a given parametric family.

Precisely, let $W_{\beta_0,\beta_1}(u,v) \triangleq \mathrm{logistic}(\beta_0 + \beta_1(u+v))$ to be the logistic graphon with given coefficients β_0 and β_1. Also, let $\mathcal{W}$ be a parametric family of canonical graphons (or equivalently a parametric subspace of degree-identifiable ExGMs) consisting of W_{β_0,β_1} for $\beta_0 \in [-5,5]$ and $\beta_1 \in (0,5]$. Here we must restrict β_1 to be positive because W_{β_0,β_1} won't be a canonical representation once $\beta_1 \leq 0$. On the other hand, we also assume that $\beta_0 \in [-5,5]$ because we need a finite but reasonably large range in order to (i) practically implement estimations and by the same time (ii) cover enough possibilities of β_0's which a traditional dense graph scenario could fit into.[‡‡]

Now, our hypothesis testing setting will be

$$H_0 : W(u,v) \in \mathcal{W} \text{ versus } H_a : W(u,v) \notin \mathcal{W},$$

so a reasonable test statistic for this example would be

$$T \triangleq \min_{\beta_0 \in [-5,5], \beta_1 \in (0,5]} \left\| W_{\beta_0,\beta_1} - \hat{W} \right\|. \tag{1.6}$$

Clearly, T represents a L^2 distance (in the space of all symmetric measurable $[0,1]$-valued functions) between $\hat{W}$ (which we thought to be very close to the true canonical graphon W as N being large enough) and $\mathcal{W}$, so the optimal candidate $W_{\beta_0^*,\beta_1^*} \in \mathcal{W}$ we can get by solving the optimization problem (1.6) is in fact an approximation to the projection of W onto $\mathcal{W}$. Therefore, under the null hypothesis and the asymptotic, $W(u,v) \in \mathcal{W}$ and hence $T = 0$.

Again, to approximate the sampling distribution of T under the null hypothesis H_0, we use 5000 Monte Carlo samples for $N = 3000$, where each sample is coming from (i) randomly drawing $\beta_0^* \in [-5,5]$ and $\beta_1^* \in (0,5]$, (ii) sampling the adjacency matrix A using the graphon $W_{\beta_0^*,\beta_1^*}$, and (iii) calculating T using the data A. Another histogram of T is shown in Figure 2. The mean

[‡‡] In this setting, when $\beta_1 \approx 0$, the range of W_{β_0,β_1} is about $[0.7\%, 99\%]$, which basically cover most of the case we will encounter in a dense graph scenario.

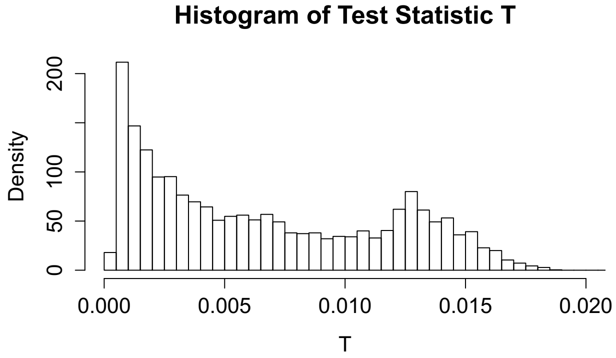


Figure 1.2: 5000 Monte Carlo draws of T under H_0 for logistic graphon.

and standard deviation of the Monte Carlo samples are 0.0072 and 0.015, and the 95% quantile is 0.0152, so the rejection region is $T \geq 0.0152$.

To conclude the two hypothesis testing examples, given an observed adjacency matrix A, we can calculate its corresponding $\hat{W}$ and T according to either equation (1.5) or (1.6), and then (i) reject the simple null hypothesis that the data is simply generated from a quadratic graphon if $T \geq 0.0152$ or (ii) reject the composite null hypothesis that the data is generated from a logistic family (with certain constraints) if $T \geq 0.0126$.

With a graphon estimation only in a P1 formulation, which implicitly uses a probability measure conditional on some latent variables, we can never write down a comparative null hypothesis like those described in the previous two Subsections–because of the lake of knowledge to those underlying latent variables. Thus, the needs of doing hypothesis testing suggests the necessity to go beyond the P1 formulation and consider our proposed P2 formulation of graphon estimation.

1.6 Discussion

1.6.1 Related work

In this Subsection, we review some common ways to estimate canonical graphon W of a degree-identifiable ExGM, contrasting them to show the strengths of our proposed USVT-A method.

The most prevailing way to estimate W in the literature is through the blockmodel approximation. Bickel and Chen (2009) first propose the use of blockmodel as a parametric approximation to W with the following disadvantages: (i) the high computation cost to do the community detection, (ii) the model misspecification of using a parametric form to approximate a nonparametric object, (iii) the inflexibility of using a fixed number of classes K, and (iv) the difficulty of choosing K. Even though the authors propose the estimation problem as though a P2 formulation (see Section 2.1), they don't provide a nonparametric estimator $\hat{W}$ and neither do they specify the issue of choosing K.

To bypass the disadvantages (iii) and (iv) listed above, Choi et al. (2012) consider a generalization of blockmodel with a growing number of classes $K = O\left(N^{1/2}\right)$, which doesn't need to choose the number of classes K and at the same time leads to a smaller model bias compared with Bickel and Chen (2009). However, the disadvantages (i) and (ii) still apply for the work of Choi et al. (2012). Besides, they address the estimation problem in terms of the P1 formulation, so neither are they able to write down (and prove the consistency of) a nonparametric estimate in a P2 formulation using this blockmodel with a growing number of classes.

On the other hand, Bickel et al. (2011) want to address the graphon estimation problem in a P2 formulation by proposing a moment estimation, which counts the motif frequencies in an observed graph. They theoretically characterize the unknown graphon by an abstract linear functional constructed by the moment estimates. Their approach looks principally convincing with theoretical beauty, but these results are almost impossible to implement. More precisely, it's difficult to explicitly solve the canonical graphon from the characteristic linear functional described above, because this will need an accurate and expensive computation to the eigenvalues and the eigenvectors of the characteristic functional.

The recent work by Wolfe and Olhede (2013) also claim to seek a nonparametric estimation to the graphon. Nevertheless, they don't clearly and uniquely define an estimand in a functional form. Furthermore, they measure the error between the underlying graphon and the estimation via the cut-metric in the theory of graph limits defined by Borgs, Chayes, Lovász, Sós, and Vesztergombi (2008), so their results in its nature cannot allow explicitly numerical simulations to check the performance of the estimation. Finally, their asymptotic theory requires very sophisticated assumptions than what we made in this study.

As a summary to the contrasts with the related works above, our proposed USVT-A method fully addresses a functional form or nonparametric estimation to the canonical graphon W and, at the same time, allows an easy and quick implementation with an exciting performance. Thus, our proposed method is more useful in the practical context.

1.6.2 Conclusions

In this Chapter we summarize some existing literatures about estimation problems for ExGM and dichotomize them into P1 and P2 formulations, one addressing only on the probability matrix estimation while the other one pursuing the fully functional form estimate for the underlying graphon.

Besides, we discuss the issue of identifiability, which must be faced before any attempt on addressing Estimation problem 2. We propose a subclass of ExGM, named degree-identifiable ExGM to allow a uniquely-defined canonical graphon to make the Estimation problem 2 formulation well-defined. Under the scope of degree-identifiable ExGM, we propose a 3-step procedure for constructing a flexible class of nonparametric estimates to the canonical graphon, while the future researchers can freely test on different combinations of (i) probability matrix estimation, (ii) latent variable estimation, and (iii) smoothing methods.

We especially propose a pre-smoothed estimate, called USVT-A method, theoretically prove its mean square error consistency with the only assumption being the continuity on the canonical graphon, and witnesses its extremely computational tractability. Some simulation results also confirm both its consistency and ease for implementation. For practical usage, if we have a strong belief that the canonical graphon W should be smooth, then a smoothing algorithm like total variation minimization method Chan et al. (2011) could be applied to get a further reduction of estimation errors; however, as shown in Subsection 1.3.2, the reduction of using total variation minimization method seems to be relatively insignificant, so other smoothing methods that give the estimate of W a functional form beyond step function with the potential to reduce more estimation errors are still highly intriguing and deserving of more future investigations.

Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning.

Albert Einstein

2

Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models

Shrinkage estimators have profound impacts in statistics and in scientific and engineering applications. In this Chapter, we consider shrinkage estimation in the presence of linear predictors. We formulate two heteroscedastic hierarchical regression models and study optimal shrinkage estimators in each model. A class of shrinkage estimators, both parametric and semiparametric, based on unbiased risk estimate (URE) is proposed and is shown to be (asymptotically) optimal

This Chapter is advised by and coauthored with Professor S.C. Samuel Kou.

under mean squared error loss in each model. Simulation study is conducted to compare the performance of the proposed methods with existing shrinkage estimators. We also apply the method to real data and obtain encouraging and interesting results.

2.1 Introduction

Shrinkage estimators, hierarchical models and empirical Bayes methods, dating back to the groundbreaking works of Stein (1956) and Robbins (1956), have profound impacts in statistics and in scientific and engineering applications. They provide effective tools to pool information from (scientifically) related populations for simultaneous inference—the data on each population alone often do not lead to the most effective estimation, but by pooling information from the related populations together (for example, by shrinking toward their consensus "center"), one could often obtain more accurate estimate for each individual population. Ever since the seminal works of Stein (1956) and James and Stein (1961), an impressive list of articles has been devoted to the study of shrinkage estimators in normal models, including Berger and Strawderman (1996), Brown (2008), Efron and Morris (1972, 1973, 1975), Green and Strawderman (1985), Jones (1991), Lindley (1962), Morris (1983), Rubin (1980), Stein (1962), among others.

In this Chapter, we consider shrinkage estimation in the presence of linear predictors. In particular, we study optimal shrinkage estimators for heteroscedastic data under linear models. Our study is motivated by three main considerations. First, in many practical problems, one often encounters heteroscedastic (unequal variance) data; for example, the sample sizes for different groups are not all equal. Second, in many statistical applications, in addition to the heteroscedastic response variable, one often has predictors. For example, the predictors could represent longitudinal patterns Fearn (1975), Hui and Berger (1983), Strenio et al. (1983), exam scores Rubin (1980), characteristics of hospital patients Normand et al. (1997), etc. Third, in applying shrinkage estimators to real data, it is quite natural to ask for the optimal way of shrinkage.

The (risk) optimality is not addressed by the conventional estimators, such as the empirical Bayes ones. One might wonder if such an optimal shrinkage estimator exists in the first place. We shall see shortly that in fact (asymptotically) optimal shrinkage estimators do exist and that the optimal estimators are not empirical Bayes ones but are characterized by an unbiased risk estimate (URE).

The study of optimal shrinkage estimators under the heteroscedastic normal model was first considered in Xie et al. (2012), where the (asymptotic) optimal shrinkage estimator was identified for both the parametric and semiparametric cases. Xie et al. (2015) extends the (asymptotic) optimal shrinkage estimators to exponential families and heteroscedastic location-scale families. The current Chapter can be viewed as an extension of the idea of optimal shrinkage estimators to heteroscedastic linear models.

We want to emphasize that this Chapter works on a theoretical setting somewhat different from Xie et al. (2015) but can still cover its main results. Our theoretical results show that the optimality of the proposed URE shrinkage estimators does not rely on normality nor on the tail behavior of the sampling distribution. What we require here are the symmetry and the existence of the fourth moment for the standardized variable.

This Chapter is organized as follows. We first formulate the heteroscedastic linear models in Section 2.2. Interestingly, there are two parallel ways to do so, and both are natural extensions of the heteroscedastic normal model. After reviewing the conventional empirical Bayes methods, we introduce the construction of our optimal shrinkage estimators for heteroscedastic linear models in Section 2.3. The optimal shrinkage estimators are based on an unbiased risk estimate (URE). We show in Section 2.4 that the URE shrinkage estimators are asymptotically optimal in risk. In Section 2.5 we extend the shrinkage estimation to a semiparametric family. Simulation studies are conducted in Section 2.6. We apply the URE shrinkage estimators in Section 2.7 to the baseball data set of Brown (2008) and observe quite interesting and encouraging results. We conclude in Section 2.8 with some discussion and extension. The appendix details the proofs and derivations for the theoretical results.

2.2 Heteroscedastic Hierarchical Linear Models

Consider the heteroscedastic estimation problem

$$Y_i|\boldsymbol{\theta} \overset{\text{indep.}}{\sim} \mathcal{N}\left(\theta_i, A_i\right), \qquad i = 1, ..., p, \tag{2.1}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ is the unknown mean vector, which is to be estimated, and the variances $A_i > 0$ are unequal, which are assumed to be known. In many statistical applications, in addition to the heteroscedastic $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$, one often has predictors $\boldsymbol{X}$. A natural question is to consider a heteroscedastic linear model that incorporates these covariates. Notation-wise, let $\{Y_i, \boldsymbol{X}_i\}_{i=1}^p$ denote the p independent statistical units, where Y_i is the response variable of the i-th unit, and $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ki})^T$ is a k-dimensional column vector that corresponds to the k covariates of the i-th unit. The $k \times p$ matrix

$$\boldsymbol{X} = [\boldsymbol{X}_1|\cdots|\boldsymbol{X}_p], \qquad \boldsymbol{X}_1, .., \boldsymbol{X}_p \in \mathbb{R}^k,$$

where $\boldsymbol{X}_i$ is the i-th column of $\boldsymbol{X}$, then contains the covariates for all the units. Throughout this Chapter we assume that $\boldsymbol{X}$ has full rank, i.e., $\text{rank}(\boldsymbol{X}) = k$.

To include the predictors, we note that, interestingly, there are two different ways to build up a heteroscedastic hierarchical linear model, which lead to different structure for shrinkage estimation.

Model I: Hierarchical linear model. On top of (2.1), the θ_i's are $\theta_i \overset{\text{indep.}}{\sim} \mathcal{N}\left(\boldsymbol{X}_i^T\boldsymbol{\beta}, \lambda\right)$, where $\boldsymbol{\beta}$ and λ are both unknown hyper-parameters. Model I has been suggested as early as Stein

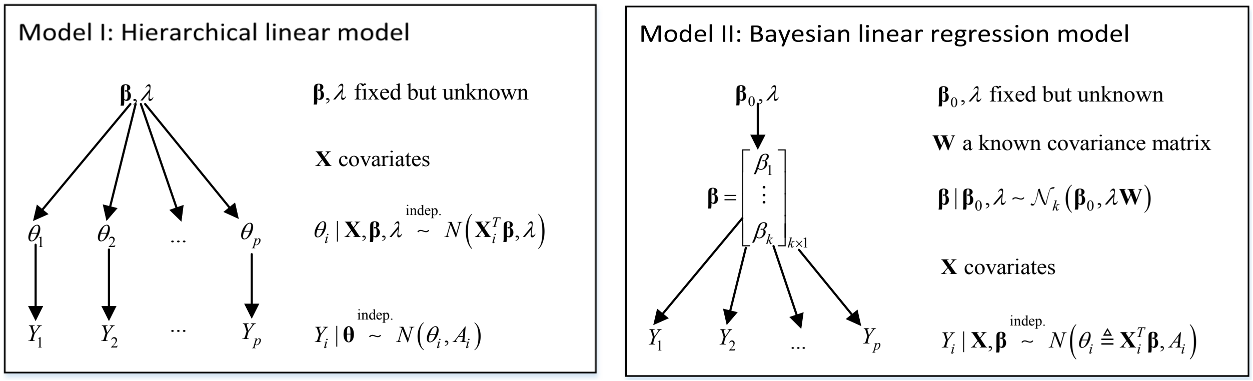

Model I: Hierarchical linear model
$\boldsymbol{\beta}, \lambda$
$\boldsymbol{\beta}, \lambda$ fixed but unknown
$\mathbf{X}$ covariates
$\theta_i \mid \mathbf{X}, \boldsymbol{\beta}, \lambda \overset{\text{indep.}}{\sim} N\left(\mathbf{X}_i^T \boldsymbol{\beta}, \lambda\right)$
θ_1 θ_2 ... θ_p
Y_1 Y_2 ... Y_p
$Y_i \mid \boldsymbol{\theta} \overset{\text{indep.}}{\sim} N\left(\theta_i, A_i\right)$
Model II: Bayesian linear regression model
$\boldsymbol{\beta}_0, \lambda$
$\boldsymbol{\beta}_0, \lambda$ fixed but unknown
$\mathbf{W}$ a known covariance matrix
$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{k\times 1}$
$\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \lambda \sim \mathcal{N}_k\left(\boldsymbol{\beta}_0, \lambda \mathbf{W}\right)$
$\mathbf{X}$ covariates
Y_1 Y_2 ... Y_p
$Y_i \mid \mathbf{X}, \boldsymbol{\beta} \overset{\text{indep.}}{\sim} N\left(\theta_i \triangleq \mathbf{X}_i^T \boldsymbol{\beta}, A_i\right)$

Figure 2.1: Graphical illustration of the two heteroscedastic hierarchical linear models.

(1966). See Morris (1983) and Morris and Lysy (2012) for more discussions. The special case of no covariates (i.e., $k = 1$ and $\boldsymbol{X} = [1|\cdots|1]$) is studied in depth in Xie et al. (2012).

Model II: Bayesian linear regression model. Together with (2.1), one assumes $\boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ following a conjugate prior distribution $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{\beta}_0, \lambda \boldsymbol{W}\right)$, where $\boldsymbol{W}$ is a known $k \times k$ positive definite matrix and $\boldsymbol{\beta}_0$ and λ are unknown hyper-parameters. Model II has been considered in Copas (1983), Lindley and Smith (1972), Raftery et al. (1997) among others; it includes ridge regression as a special case when $\boldsymbol{\beta}_0 = \mathbf{0}_k$ and $\boldsymbol{W} = \boldsymbol{I}_k$.

Figure 2.1 illustrates these two hierarchical linear models. Under Model I, the posterior mean of $\boldsymbol{\theta}$ is $\hat{\theta}_i^{\lambda,\boldsymbol{\beta}} = \lambda\left(\lambda + A_i\right)^{-1} Y_i + A_i\left(\lambda + A_i\right)^{-1} \boldsymbol{X}_i^T\boldsymbol{\beta}$ for $i = 1, ..., p$, so the shrinkage estimation is formed by directly shrinking the raw observation Y_i toward a linear combination of the k covariates $\boldsymbol{X}_i$. If we denote $\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$, and $\boldsymbol{\mu} = \left(\mu_1, ..., \mu_p\right)^T \in \mathcal{L}_{\text{row}}\left(\boldsymbol{X}\right)$, the row space of $\boldsymbol{X}$, then we can rewrite the posterior mean of $\boldsymbol{\theta}$ under Model I as

$$\hat{\theta}^{\lambda,\boldsymbol{\mu}} = \frac{\lambda}{\lambda + A_i} Y_i + \frac{A_i}{\lambda + A_i}\mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathcal{L}_{\text{row}}\left(\boldsymbol{X}\right). \tag{2.2}$$

Under Model II, the posterior mean of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0}, \quad \text{with } \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda\boldsymbol{W}(\lambda\boldsymbol{W} + \boldsymbol{V})^{-1}\hat{\boldsymbol{\beta}}^{\text{WLS}} + \boldsymbol{V}\left(\lambda\boldsymbol{W} + \boldsymbol{V}\right)^{-1}\boldsymbol{\beta}_0, \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}^{\text{WLS}} = \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}$ is the weighted least squares estimate of the regression coefficient, $\boldsymbol{A}$ is the diagonal matrix $\boldsymbol{A} = \text{diag}\left(A_1, ..., A_p\right)$, and $\boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$. Thus, the estimate for θ_i is linear in $\boldsymbol{X}_i$, and the "shrinkage" is achieved by shrinking the regression coefficient from the weighted least squares estimate $\hat{\boldsymbol{\beta}}^{\mathrm{WLS}}$ toward the prior coefficient $\boldsymbol{\beta}_0$.

As both Models I and II are natural generalizations of the heteroscedastic normal model (2.1), we want to investigate if there is an optimal choice of the hyper-parameters in each case. Specifically, we want to investigate the best empirical choice of the hyper-parameters in each case under the mean squared error loss

$$l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{p} \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 = \frac{1}{p} \sum_{i=1}^{p} \left(\theta_i - \hat{\theta}_i \right)^2 \tag{2.4}$$

with the associated risk of $\hat{\boldsymbol{\theta}}$ defined by

$$R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \right),$$

where the expectation is taken with respect to $\boldsymbol{Y}$ given $\boldsymbol{\theta}$.

Remark 4. *Even though we start from the Bayesian setting to motivate the form of shrinkage estimators, our discussion will be all based on the frequentist setting. Hence all probabilities and expectations throughout this Chapter are fixed at the unknown true $\boldsymbol{\theta}$, which is free in $\mathbb{R}^p$ for Model I and confined in $\mathcal{L}_{\mathrm{row}}(\boldsymbol{X})$ for Model II.*

Remark 5. *The diagonal assumption of $\boldsymbol{A}$ is quite important for Model I but not so for Model II, as in Model II we can always apply some linear transformations to obtain a diagonal covariance matrix. Without loss of generality, we will keep the diagonal assumption for $\boldsymbol{A}$ in Model II.*

For the ease of exposition, we will next overview the conventional empirical Bayes estimates in a general two-level hierarchical model, which includes both Models I and II:

$$\boldsymbol{Y}|\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{A}) \text{ and } \boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{B}), \tag{2.5}$$

where $\boldsymbol{B}$ is a non-negative definite symmetric matrix that is restricted in an allowable set $\mathcal{B}$, and $\boldsymbol{\mu}$ is in the row space $\mathcal{L}_{\mathrm{row}}(\boldsymbol{X})$ of $\boldsymbol{X}$.

Remark 6. *Under Model I, $\boldsymbol{\mu}$ and $\boldsymbol{B}$ take the form of $\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{\beta}$ and $\boldsymbol{B} \in \mathcal{B} = \{\lambda \boldsymbol{I}_p : \lambda > 0\}$, whereas under Model II, $\boldsymbol{\mu}$ and $\boldsymbol{B}$ take the form of $\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{\beta}_0$ and $\boldsymbol{B} \in \mathcal{B} = \left\{\lambda \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} : \lambda > 0\right\}$. It is interesting to observe that in Model I, $\boldsymbol{B}$ is of full rank, while in Model II, $\boldsymbol{B}$ is of rank k. As we shall see, this distinction will have interesting theoretical implications for the optimal shrinkage estimators.*

Lemma 4. *Under the two-level hierarchical model (2.5), the posterior distribution is*

$$\boldsymbol{\theta}|\boldsymbol{Y} \sim \mathcal{N}_p\left(\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}, \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\right),$$

and the marginal distribution of $\boldsymbol{Y}$ is $\boldsymbol{Y} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \boldsymbol{A}+\boldsymbol{B}\right)$.

For given values of $\boldsymbol{B}$ and $\boldsymbol{\mu}$, the posterior mean of the parameter $\boldsymbol{\theta}$ leads to the Bayes estimate

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}} = \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}. \tag{2.6}$$

To use the Bayes estimate in practice, one has to specify the hyper-parameters in $\boldsymbol{B}$ and $\boldsymbol{\mu}$. The conventional empirical Bayes method uses the marginal distribution of $\boldsymbol{Y}$ to estimate the hyper-parameters. For instance, the empirical Bayes maximum likelihood estimates (EBMLE) $\hat{\boldsymbol{B}}^{\mathrm{EBMLE}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{EBMLE}}$ are obtained by maximizing the marginal likelihood of $\boldsymbol{Y}$:

$$\left(\hat{\boldsymbol{B}}^{\mathrm{EBMLE}}, \hat{\boldsymbol{\mu}}^{\mathrm{EBMLE}}\right) = \underset{\substack{\boldsymbol{B}\in\mathcal{B} \\ \boldsymbol{\mu}\in\mathcal{L}_{\mathrm{row}}(\boldsymbol{X})}}{\operatorname{argmax}} \; -\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)^T\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right) - \log\left(\det\left(\boldsymbol{A}+\boldsymbol{B}\right)\right).$$

Alternatively, the empirical Bayes method-of-moment estimates (EBMOM) $\hat{\boldsymbol{B}}^{\mathrm{EBMOM}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{EBMOM}}$ are obtained by solving the following moment equations for $\boldsymbol{B} \in \mathcal{B}$ and $\boldsymbol{\mu} \in \mathcal{L}_{\mathrm{row}}\left(\boldsymbol{X}\right)$:

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{X}^T\left(\boldsymbol{X}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{Y}, \\ \boldsymbol{B} &= \left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)^T - \boldsymbol{A}.\end{aligned}$$

If no solutions of $\boldsymbol{B}$ can be found in $\mathcal{B}$, we then set $\hat{\boldsymbol{B}}^{\text{EBMOM}} = \mathbf{0}_{p\times p}$. Adjustment for the loss of k degrees of freedom from the estimation of $\boldsymbol{\mu}$ might be applicable for $\boldsymbol{B} = \lambda \boldsymbol{C}$ ($\boldsymbol{C} = \boldsymbol{I}_p$ for Model I and $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ for Model II): we can replace the second moment equation by

$$\lambda = \left(\frac{p}{p-k} \frac{\|\boldsymbol{Y}-\boldsymbol{\mu}\|^2}{\operatorname{tr}(\boldsymbol{C})} - \frac{\operatorname{tr}(\boldsymbol{A})}{\operatorname{tr}(\boldsymbol{C})} \right)^+ .$$

The corresponding empirical Bayes shrinkage estimator $\hat{\boldsymbol{\theta}}^{\text{EBMLE}}$ or $\hat{\boldsymbol{\theta}}^{\text{EBMOM}}$ is then formed by plugging ($\hat{\boldsymbol{B}}^{\text{EBMLE}}, \hat{\boldsymbol{\mu}}^{\text{EBMLE}}$) or ($\hat{\boldsymbol{B}}^{\text{EBMOM}}, \hat{\boldsymbol{\mu}}^{\text{EBMOM}}$) into equation (2.6).

2.3 URE Estimates

The formulation of the empirical Bayes estimates raises a natural question: which one is preferred $\hat{\boldsymbol{\theta}}^{\text{EBMLE}}$ or $\hat{\boldsymbol{\theta}}^{\text{EBMOM}}$? More generally, is there an optimal way to choose the hyper-parameters? It turns out that neither $\hat{\boldsymbol{\theta}}^{\text{EBMLE}}$ nor $\hat{\boldsymbol{\theta}}^{\text{EBMOM}}$ is optimal. The (asymptotically) optimal estimate, instead of relying on the marginal distribution of $\boldsymbol{Y}$, is characterized by an unbiased risk estimate (URE). The idea of forming a shrinkage estimate through URE for heteroscedastic models is first suggested in Xie et al. (2012). We shall see that in our context of hierarchical linear models (both Models I and II) the URE estimators that we are about to introduce have (asymptotically) optimal risk properties.

The basic idea behind URE estimators is the following. Ideally we want to find the hyper-parameters that give the smallest risk. However, since the risk function depends on the unknown $\boldsymbol{\theta}$, we cannot directly minimize the risk function in practice. If we can find a good estimate of the risk function instead, then minimizing this proxy of the risk will lead to a competitive estimator.

To formally introduce the URE estimators, we start from the observation that, under the mean squared error loss (2.4), the risk of the Bayes estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}$ for fixed $\boldsymbol{B}$ and $\boldsymbol{\mu}$ is

$$R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}) = \frac{1}{p} \left\| \boldsymbol{A} \left(\boldsymbol{A} + \boldsymbol{B} \right)^{-1} \left(\boldsymbol{\mu} - \boldsymbol{\theta} \right) \right\|^2 + \frac{1}{p} \mathrm{tr} \left(\boldsymbol{B} \left(\boldsymbol{A} + \boldsymbol{B} \right)^{-1} \boldsymbol{A} \left(\boldsymbol{A} + \boldsymbol{B} \right)^{-1} \boldsymbol{B} \right), \tag{2.7}$$

which can be easily shown using the bias-variance decomposition of the mean squared error. As the risk function involves the unknown $\boldsymbol{\theta}$, we cannot directly minimize it. However, an unbiased estimate of the risk is available:

$$\mathrm{URE} \left(\boldsymbol{B}, \boldsymbol{\mu} \right) = \frac{1}{p} \left\| \boldsymbol{A} \left(\boldsymbol{A} + \boldsymbol{B} \right)^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu} \right) \right\|^2 + \frac{1}{p} \mathrm{tr} \left(\boldsymbol{A} - 2 \boldsymbol{A} \left(\boldsymbol{A} + \boldsymbol{B} \right)^{-1} \boldsymbol{A} \right), \tag{2.8}$$

which again can be easily shown using the bias-variance decomposition of the mean squared error. Intuitively, if $\mathrm{URE} \left(\boldsymbol{B}, \boldsymbol{\mu} \right)$ is a good approximation of the actual risk, then we would expect the estimator obtained by minimizing the URE to have good properties. This leads to the URE estimator $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$, defined by

$$\hat{\boldsymbol{\theta}}^{\mathrm{URE}} = \hat{\boldsymbol{B}}^{\mathrm{URE}} (\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}})^{-1} \boldsymbol{Y} + \boldsymbol{A} (\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}})^{-1} \hat{\boldsymbol{\mu}}^{\mathrm{URE}}, \tag{2.9}$$

where

$$\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}} \right) = \underset{B \in \mathcal{B},\ \boldsymbol{\mu} \in \mathcal{L}_{\mathrm{row}}(\boldsymbol{X})}{\mathrm{argmin}} \mathrm{URE} \left(\boldsymbol{B}, \boldsymbol{\mu} \right).$$

It is worth noting that the value of $\boldsymbol{\mu}$ that minimizes (2.8) for a given $\boldsymbol{B}$ is neither the ordinary least square (OLS) nor the weighted least square (WLS) regression estimate, echoing similar observation as in Xie et al. (2012).

In the URE estimator (2.9), $\hat{\boldsymbol{B}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{URE}}$ are jointly determined by minimizing the URE. When the number of independent statistical units p is small or moderate, joint minimization of $\boldsymbol{B}$ and the vector $\boldsymbol{\mu}$, however, may be too ambitious. In this setting, it might be beneficial to set $\boldsymbol{\mu}$ by a predetermined rule and only optimize $\boldsymbol{B}$, as it might reduce the variability of the resulting estimate. In particular, we can consider shrinking toward a generalized least squares (GLS) regression estimate

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{Y},$$

where $\boldsymbol{M}$ is a prespecified symmetric positive definite matrix. This use of $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$ gives the shrinkage estimate $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}} = \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$, where one only needs to determine $\boldsymbol{B}$. We can construct another URE estimate for this purpose. Similar to the previous construction, we note that $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}$ has risk

$$\begin{aligned} R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}) &= \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\boldsymbol{\theta}\right\|^2 \\ &+ \frac{1}{p}\mathrm{tr}\left(\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\right)\boldsymbol{A}\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\right)^T\right). \end{aligned} \tag{2.10}$$

An unbiased risk estimate of it is

$$\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right) = \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right)\right\|^2 + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A} - 2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\boldsymbol{A}\right). \tag{2.11}$$

Both (2.10) and (2.11) can be easily proved by the bias-variance decomposition of mean squared error. Minimizing $\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right)$ over $\boldsymbol{B}$ gives the URE GLS shrinkage estimator (which shrinks toward $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$):

$$\hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{\boldsymbol{M}} = \hat{\boldsymbol{B}}^{\mathrm{URE}}_{\boldsymbol{M}}\left(\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}}_{\boldsymbol{M}}\right)^{-1}\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}}_{\boldsymbol{M}}\right)^{-1}\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}, \tag{2.12}$$

where

$$\hat{\boldsymbol{B}}^{\mathrm{URE}}_{\boldsymbol{M}} = \underset{\boldsymbol{B}\in\mathcal{B}}{\mathrm{argmin}}\,\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right).$$

Remark 7. *When $\boldsymbol{M} = \boldsymbol{I}_p$, clearly $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, the ordinary least squares regression estimate. When $\boldsymbol{M} = \boldsymbol{A}^{-1}$, then $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, the weighted least squares regression estimate.*

Remark 8. *Tan (2015) briefly discussed the URE minimization approach for Model I without the covariates in Xie et al. (2012) in relation to Jiang et al. (2011), where Model I is assumed but an unbiased estimate of the mean prediction error (rather than the mean squared error) is* *used to form a predictor (rather than an estimator).*

Remark 9. *In the homoscedastic case, (2.12) reduces to standard shrinkage toward a subspace $\mathcal{L}_{\text{row}}(\boldsymbol{X})$, as discussed, for instance, in Sclove et al. (1972) and Omen (1982).*

2.4 Theoretical Properties of URE Estimates

This section is devoted to the risk properties of the URE estimators. Our core theoretical result is to show that the risk estimate URE is not only unbiased for the risk but, more importantly, uniformly close to the actual loss. We therefore expect that minimizing URE would lead to an estimate with competitive risk properties.

2.4.1 Uniform Convergence of URE

To present our theoretical result, we first define $\mathcal{L}$ to be a subset of $\mathcal{L}_{\text{row}}(\boldsymbol{X})$:

$$\mathcal{L} = \{\boldsymbol{\mu} \in \mathcal{L}_{\text{row}}(\boldsymbol{X}) : \|\boldsymbol{\mu}\| \leq M p^{\kappa} \|\boldsymbol{Y}\|\},$$

where M is a large and fixed constant and $\kappa \in [0, 1/2)$ is a constant. Next, we introduce the following regularity conditions:

(A) $\sum_{i=1}^{p} A_i^2 = O(p)$; (B) $\sum_{i=1}^{p} A_i \theta_i^2 = O(p)$; (C) $\sum_{i=1}^{p} \theta_i^2 = O(p)$;

(D) $p^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T \to \boldsymbol{\Omega}_D$; (E) $p^{-1}\boldsymbol{X}\boldsymbol{X}^T \to \boldsymbol{\Omega}_{\boldsymbol{E}} > 0$;

(F) $p^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T \to \boldsymbol{\Omega}_F > 0$; (G) $p^{-1}\boldsymbol{X}\boldsymbol{A}^{-2}\boldsymbol{X}^T \to \boldsymbol{\Omega}_G$.

The theorem below shows that $\text{URE}(\boldsymbol{B}, \boldsymbol{\mu})$ not only unbiasedly estimates the risk but also is (asymptotically) uniformly close to the actual loss.

Theorem 5. *Assume conditions (A)-(E) for Model I or assume conditions (A) and (D)-(G) for*

Model II. In either case, we have

$$\sup_{\boldsymbol{B}\in\mathcal{B},\ \boldsymbol{\mu}\in\mathcal{L}} \left|\mathrm{URE}\left(\boldsymbol{B},\boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \to 0 \ \textit{in}\ L^1,\ \ \textit{as}\ p\to\infty.$$

We want to remark here that the set $\mathcal{L}$ gives the allowable range of $\boldsymbol{\mu}$: the norm of $\boldsymbol{\mu}$ is up to an $o\left(p^{1/2}\right)$ multiple of the norm of $\boldsymbol{Y}$. This choice of $\mathcal{L}$ does not lead to any difficulty in practice because, given a large enough constant M, it will cover the shrinkage locations of any sensible shrinkage estimator. We note that it is possible to define the range of sensible shrinkage locations in other ways (e.g., one might want to define it by ∞-norm in $\mathbb{R}^p$), but we find our setting more theoretically appealing and easy to work with. In particular, our assumption of the exponent $\kappa < 1/2$ is flexible enough to cover most interesting cases, including $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, the ordinary least squares regression estimate, and $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, the weighted least squares regression estimate (as in Remark 4) as shown in the following lemma.

Lemma 6. *(i) $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}} \in \mathcal{L}$. (ii) Assume* (A) *and* (A′) $\sum_{i=1}^{p} A_i^{-2-\delta} = O\left(p\right)$ *for some* $\delta > 0$*; then* $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}} \in \mathcal{L}$ *for* $\kappa = 4^{-1} + \left(4+2\delta\right)^{-1}$ *and a large enough* M*.*

Remark 10. *We want to mention here that Theorem 5 in the case of Model I covers Theorem 5.1 of Xie et al. (2012) (which is the special case of* $k = 1$ *and* $\boldsymbol{X} = [1|1|...|1]$*) because the restriction of* $|\mu| \leq \max_{1\leq i\leq p} |Y_i|$ *in Xie et al. (2012) is contained in* $\mathcal{L}$ *as*

$$\max_{1\leq i\leq p} |Y_i| = (\max_{1\leq i\leq p} Y_i^2)^{1/2} \leq (\sum_{i=1}^{p} Y_i^2)^{1/2} = \|\boldsymbol{Y}\|\,.$$

Furthermore, we do not require the stronger assumption of $\sum_{i=1}^{p} |\theta_i|^{2+\delta} = O\left(p\right)$ *for some* $\delta > 0$ *made in Xie et al. (2012). Note that in this case (*$k = 1$ *and* $\boldsymbol{X} = [1|1|...|1]$*) we do not even require conditions* (D) *and* (E)*, as condition* (A) *directly implies* $\mathrm{tr}(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T) = O\left(1\right)$*, the result we need in the proof of Theorem 5 for Model I.*

Remark 11. *In the proof of Theorem 5, the sampling distribution of* $\boldsymbol{Y}$ *is involved only through* *the moment calculations, such as* $\mathbb{E}(\mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^T)^2)$ *and* $\mathbb{E}(\|\boldsymbol{Y}\|^2)$*. It is therefore straightforward to generalize Theorem 5 to the case of*

$$Y_i = \theta_i + \sqrt{A_i} Z_i,$$

where Z_i *follows* any *distribution with mean* 0*, variance* 1*,* $\mathbb{E}\left(Z_i^3\right) = 0$*, and* $\mathbb{E}\left(Z_i^4\right) < \infty$*. This is noteworthy as our result also covers that of Xie et al. (2015) but the methodology we employ here does not require to control the tail behavior of* Z_i *as in Xie et al. (2012, 2015).*

2.4.2 Risk Optimality

In this section, we consider the risk properties of the URE estimators. We will show that, under the hierarchical linear models, the URE estimators have (asymptotically) optimal risk, whereas it is not necessarily so for other shrinkage estimators such as the empirical Bayes ones.

A direct consequence of the uniform convergence of URE is that the URE estimator has a loss/risk that is asymptotically no larger than that of any other shrinkage estimators. Furthermore, the URE estimator is asymptotically as good as the oracle loss estimator. To be precise, let $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ be the oracle loss (OL) estimator defined by plugging

$$\begin{aligned}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right) &= \underset{B\in\mathcal{B},\ \boldsymbol{\mu}\in\mathcal{L}}{\operatorname{argmin}}\ l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right) \\ &= \underset{B\in\mathcal{B},\ \boldsymbol{\mu}\in\mathcal{L}}{\operatorname{argmin}} \left\|\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu} - \boldsymbol{\theta}\right\|^2\end{aligned}$$

into (2.6). Of course, $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ is not really an estimator, since it depends on the unknown $\boldsymbol{\theta}$ (hence we use the notation $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ rather than $\hat{\boldsymbol{\theta}}^{\mathrm{OL}}$). Although not obtainable in practice, $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ lays down the theoretical limit that one can ever hope to reach. The next theorem shows that the URE estimator $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ is asymptotically as good as the oracle loss estimator, and, consequently, it is asymptotically at least as good as any other shrinkage estimator.

Theorem 7. *Assume the conditions of Theorem 5 and that $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathcal{L}$. Then*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,$$
$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right)\right) = 0.$$

Corollary 8. *Assume the conditions of Theorem 5 and that $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathcal{L}$. Then for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p} = \hat{\boldsymbol{B}}_p\left(\boldsymbol{A} + \hat{\boldsymbol{B}}_p\right)^{-1}\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A} + \hat{\boldsymbol{B}}_p\right)^{-1}\hat{\boldsymbol{\mu}}_p$ with $\hat{\boldsymbol{B}}_p \in \mathcal{B}$ and $\hat{\boldsymbol{\mu}}_p \in \mathcal{L}$, we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,$$
$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right)\right) \leq 0.$$

Corollary 8 tells us that the URE estimator in either Model I or II is asymptotically optimal: it has (asymptotically) the smallest loss and risk among all shrinkage estimators of the form (2.6).

2.4.3 Shrinkage toward the Generalized Least Squares Estimate

The risk optimality also holds when we consider the URE estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ that shrinks toward the GLS regression estimate $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{Y}$ as introduced in Section 2.3.

Theorem 9. *Assume the conditions of Theorem 5, $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} \in \mathcal{L}$, and*

$$p^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_1 > 0, \quad p^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_2, \quad p^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_3, \tag{2.13}$$

where only the first and third conditions above are assumed for Model I and only the first and the second are assumed for Model II. Then we have

$$\sup_{\boldsymbol{B}\in\mathcal{B}} \left|\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right)\right| \to 0 \text{ in } L^1 \text{ as } p \to \infty. \tag{2.14}$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}} = \hat{\boldsymbol{B}}_p\left(\boldsymbol{A} + \hat{\boldsymbol{B}}_p\right)^{-1}\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A} + \hat{\boldsymbol{B}}_p\right)^{-1}\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$ with $\hat{\boldsymbol{B}}_p \in$

$\mathcal{B}$, we always have

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{\boldsymbol{M}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) + \epsilon\right) = 0 \qquad \forall \epsilon > 0,$$
$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{\boldsymbol{M}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right)\right) \leq 0.$$

Remark 12. *For shrinking toward $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, where $\boldsymbol{M} = \boldsymbol{I}_p$, we know from Lemma 6 that $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$ is automatically in $\mathcal{L}$, so we only need one more condition $p^{-1}\boldsymbol{X}\boldsymbol{A}^2\boldsymbol{X}^T \to \boldsymbol{\Omega}_3$ for Model I. For shrinking toward $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, where $\boldsymbol{M} = \boldsymbol{A}^{-1}$, (2.13) is the same as the conditions (E) and (F) of Theorem 5, so additionally we only need to assume* (A$'$) *of Lemma 6 and (F) for Model I.*

2.5 Semiparametric URE Estimators

We have established the (asymptotic) optimality of the URE estimators $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{\boldsymbol{M}}$ in the previous section. One limitation of the result is that the class over which the URE estimators are optimal is specified by a parametric form: $\boldsymbol{B} = \lambda\boldsymbol{C}$ $(0 \leq \lambda \leq \infty)$ in equation (2.6), where $\boldsymbol{C} = \boldsymbol{I}_p$ for Model I and $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ for Model II. Aiming to provide a more flexible and, at the same time, efficient estimation procedure, we consider in this section a class of semiparametric shrinkage estimators. Our consideration is inspired by Xie et al. (2012).

2.5.1 Semiparametric URE Estimator under Model I

To motivate the semiparametric shrinkage estimators, let us first revisit the Bayes estimator $\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\mu}}$ under Model I, as given in (2.2). It is seen that the Bayes estimate of each mean parameter θ_i is obtained by shrinking Y_i toward the linear estimate $\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$, and that the amount of shrinkage is governed by A_i, the variance: the larger the variance, the stronger is the shrinkage. This feature makes intuitive sense.

With this observation in mind, we consider the following shrinkage estimators under Model I:

$$\hat{\theta}_i^{\boldsymbol{b},\boldsymbol{\mu}} = (1-b_i)\,Y_i + b_i\mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathcal{L}_{\text{row}}\left(\boldsymbol{X}\right),$$

where $\boldsymbol{b}$ satisfies the monotonic constraint

$$\text{MON}\left(\boldsymbol{A}\right): b_i \in [0,1]\,,\ b_i \leq b_j \text{ whenever } A_i \leq A_j.$$

$\text{MON}\left(\boldsymbol{A}\right)$ asks the estimator to shrink more for an observation with a larger variance. Since other than this intuitive requirement, we do not post any parametric restriction on b_i, this class of estimators is semiparametric in nature.

Following the optimality result for the parametric case, we want to investigate, for such a general estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}$ with $\boldsymbol{b} \in \text{MON}\left(\boldsymbol{A}\right)$ and $\boldsymbol{\mu} \in \mathcal{L}_{\text{row}}\left(\boldsymbol{X}\right)$, whether there exists an optimal choice of $\boldsymbol{b}$ and $\boldsymbol{\mu}$. In fact, we will see shortly that such an optimal choice exists, and this asymptotically optimal choice is again characterized by an unbiased risk estimate (URE). For a general estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}$ with fixed $\boldsymbol{b}$ and $\boldsymbol{\mu} \in \mathcal{L}_{\text{row}}\left(\boldsymbol{X}\right)$, an unbiased estimate of its risk $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}})$ is

$$\text{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\mu}\right) = \frac{1}{p}\left\|\text{diag}\left(\boldsymbol{b}\right)\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\right\|^2 + \frac{1}{p}\text{tr}\left(\boldsymbol{A} - 2\text{diag}\left(\boldsymbol{b}\right)\boldsymbol{A}\right),$$

which can be easily seen by taking $\boldsymbol{B} = \boldsymbol{A}(\text{diag}\left(\boldsymbol{b}\right)^{-1} - \boldsymbol{I}_p)$ in (2.8). Note that we use the superscript "*SP*" (semiparametric) to denote it. Minimizing over $\boldsymbol{b}$ and $\boldsymbol{\mu}$ leads to the semiparametric URE estimator $\hat{\boldsymbol{\theta}}_{SP}^{\text{URE}}$, defined by

$$\hat{\boldsymbol{\theta}}_{SP}^{\text{URE}} = (\boldsymbol{I}_p - \text{diag}(\hat{\boldsymbol{b}}_{SP}^{\text{URE}}))\boldsymbol{Y} + \text{diag}(\hat{\boldsymbol{b}}_{SP}^{\text{URE}})\hat{\boldsymbol{\mu}}_{SP}^{\text{URE}}, \tag{2.15}$$

where

$$\left(\hat{\boldsymbol{b}}_{SP}^{\text{URE}}, \hat{\boldsymbol{\mu}}_{SP}^{\text{URE}}\right) = \underset{\boldsymbol{b}\in\text{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathcal{L}_{\text{row}}(\boldsymbol{X})}{\text{argmin}} \text{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\mu}\right).$$

Theorem 10. *Assume conditions (A)-(E). Then under Model I we have*

$$\sup_{\boldsymbol{b}\in \mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathcal{L}} \left| \mathrm{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}\right)\right| \to 0 \textit{ in } L^1 \textit{ as } p\to\infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_{\boldsymbol{p}},\hat{\boldsymbol{\mu}}_p} = (\boldsymbol{I}_p - \mathrm{diag}(\hat{\boldsymbol{b}}_p))\boldsymbol{Y} + \mathrm{diag}(\hat{\boldsymbol{b}}_p)\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}\left(\boldsymbol{A}\right)$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathcal{L}$*, we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right) + \epsilon\right) = 0 \qquad \forall\epsilon>0,$$

$$\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right)\right) \leq 0.$$

The proof is the same as the proofs of Theorem 5 and Corollary 8 for the case of Model I except that we replace each term of $A_i/(\lambda + A_i)$ by b_i.

2.5.2 Semiparametric URE Estimator Under Model II

We saw in Section 2.2 that, under Model II, shrinkage is achieved by shrinking the regression coefficient from the weighted least squares estimate $\hat{\boldsymbol{\beta}}^{\mathrm{WLS}}$ toward the prior coefficient $\boldsymbol{\beta}_0$. This suggests us to formulate the semiparametric estimators through the regression coefficient. The Bayes estimate of the regression coefficient is

$$\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda\boldsymbol{W}(\lambda\boldsymbol{W}+\boldsymbol{V})^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{V}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{\beta}_0, \qquad \text{with } \boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$$

as shown in (2.3). Applying the spectral decomposition on $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}$ gives $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda} = \mathrm{diag}\left(d_1,...,d_k\right)$ with $d_1 \leq \cdots \leq d_k$. Using this decomposition, we can rewrite the regression coefficient as

$$\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda\boldsymbol{W}^{1/2}\boldsymbol{U}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{W}^{1/2}\boldsymbol{U}\boldsymbol{\Lambda}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0.$$

If we denote $\boldsymbol{Z} = \boldsymbol{U}^T \boldsymbol{W}^{1/2} \boldsymbol{X}$ as the transformed covariate matrix, the estimate $\hat{\boldsymbol{\theta}}^{\lambda, \boldsymbol{\beta}_0} = \boldsymbol{X}^T \hat{\boldsymbol{\beta}}^{\lambda, \boldsymbol{\beta}_0}$ of $\boldsymbol{\theta}$ can be rewritten as

$$\hat{\boldsymbol{\theta}}^{\lambda, \boldsymbol{\beta}_0} = \boldsymbol{Z}^T \left(\lambda \left(\lambda \boldsymbol{I}_k + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{\Lambda} \left(\lambda \boldsymbol{I}_k + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \boldsymbol{\beta}_0 \right).$$

Now we see that $\lambda \left(\lambda \boldsymbol{I}_k + \boldsymbol{\Lambda} \right)^{-1} = \mathrm{diag}(\lambda / \left(\lambda + d_i \right))$ plays the role as the shrinkage factor. The larger the value of d_i, the smaller $\lambda / \left(\lambda + d_i \right)$, i.e., the stronger the shrinkage toward $\boldsymbol{\beta}_0$. Thus, d_i can be viewed as the effective "variance" component for the i-th regression coefficient (under the transformation). This observation motivates us to consider semiparametric shrinkage estimators of the following form

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\boldsymbol{b}, \boldsymbol{\beta}_0} &= \boldsymbol{Z}^T \left(\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b} \right) \right) \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \mathrm{diag}\left(\boldsymbol{b} \right) \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \boldsymbol{\beta}_0 \right) \\ &= \boldsymbol{Z}^T \left(\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b} \right) \right) \boldsymbol{\Lambda} \boldsymbol{Z} \boldsymbol{A}^{-1} \boldsymbol{Y} + \mathrm{diag}\left(\boldsymbol{b} \right) \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \boldsymbol{\beta}_0 \right), \end{aligned} \tag{2.16}$$

where $\boldsymbol{b}$ satisfies the following monotonic constraint

$$\mathrm{MON}\left(\boldsymbol{D} \right): b_i \in [0, 1], \ b_i \le b_j \text{ whenever } d_i \le d_j.$$

This constraint captures the intuition that, the larger the effective variance, the stronger is the shrinkage.

For fixed $\boldsymbol{b}$ and $\boldsymbol{\beta}_0$, an unbiased estimate of the risk $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b}, \boldsymbol{\beta}_0})$ is

$$\begin{aligned} \mathrm{URE}^{SP}\left(\boldsymbol{b}, \boldsymbol{\beta}_0 \right) = \frac{1}{p} &\left\| \boldsymbol{Z}^T \left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b} \right) \right) \boldsymbol{\Lambda} \boldsymbol{Z} \boldsymbol{A}^{-1} \boldsymbol{Y} + \boldsymbol{Z}^T \mathrm{diag}\left(\boldsymbol{b} \right) \boldsymbol{U}^T \boldsymbol{W}^{-1/2} \boldsymbol{\beta}_0 - \boldsymbol{Y} \right\|^2 \\ &+ \frac{1}{p} \mathrm{tr}\left(2 \boldsymbol{Z}^T \left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b} \right) \right) \boldsymbol{\Lambda} \boldsymbol{Z} - \boldsymbol{A} \right), \end{aligned}$$

which can be shown using the bias-variance decomposition of the mean squared error. Minimizing it gives the URE estimate of $(\boldsymbol{b}, \boldsymbol{\beta}_0)$:

$$\left(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}}, \left(\hat{\boldsymbol{\beta}}_0\right)_{SP}^{\mathrm{URE}}\right) = \underset{\boldsymbol{b} \in \mathrm{MON}(\boldsymbol{D}),\ \boldsymbol{\beta}_0 \in \mathbb{R}^k}{\operatorname{argmin}} \mathrm{URE}^{SP}\left(\boldsymbol{b}, \boldsymbol{\beta}_0\right),$$

which upon plugging into (2.16) yields the semiparametric URE estimator $\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}$ under Model II.

Theorem 11. *Assume conditions (A), (D)-(G). Then under Model II we have*

$$\sup_{\boldsymbol{b} \in \mathrm{MON}(\boldsymbol{D}),\ \boldsymbol{X}^T\boldsymbol{\beta}_0 \in \mathcal{L}} \left| \mathrm{URE}^{SP}\left(\boldsymbol{b}, \boldsymbol{\beta}_0\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b}, \boldsymbol{\beta}_0}\right) \right| \to 0 \text{ in } L^1 \text{ as } p \to \infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_{\boldsymbol{p}}, \hat{\boldsymbol{\beta}}_{0,p}}$ obtained from (2.16) with $\hat{\mathbf{b}}_p \in \mathrm{MON}\left(\boldsymbol{D}\right)$ and $\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_0 \in \mathcal{L}$, we always have

$$\lim_{p \to \infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\beta}}_{0,p}}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p \to \infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\beta}}_{0,p}}\right)\right) \leq 0.$$

The proof of the theorem is essentially identical to those of Theorem 5 and Corollary 8 for the case of Model II except that we replace each $d_i/(\lambda + d_i)$ by b_i.

2.6 Simulation Study

In this section, we conduct simulations to study the performance of the URE estimators. For the sake of space, we will focus on Model I. The four URE estimators are the parametric $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ of equation (2.9), the parametric $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ of equation (2.12) that shrinks toward the OLS estimate $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$ (i.e., the matrix $\boldsymbol{M} = \boldsymbol{I}_p$), the semiparametric $\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}$ of equation (2.15), and the semiparametric $\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE,\,OLS}}$ that shrinks toward $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, which is formed similarly to $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ by replacing $A_i/(\lambda + A_i)$ with a sequence $\boldsymbol{b} \in \mathrm{MON}\left(\boldsymbol{A}\right)$. The competitors here are the two empirical Bayes estimators $\hat{\boldsymbol{\theta}}^{\mathrm{EBMLE}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{EBMOM}}$, and the positive part James-Stein estimator $\hat{\boldsymbol{\theta}}^{\mathrm{JS+}}$ as described in Brown (2008), Morris and Lysy (2012):

$$\hat{\theta}_i^{\text{JS+}} = \hat{\mu}_i^{\text{WLS}} + \left(1 - \frac{p-k-2}{\sum_{i=1}^{p} \left(Y_i - \hat{\mu}_i^{\text{WLS}}\right)^2 / A_i}\right)^{+} \left(Y_i - \hat{\mu}_i^{\text{WLS}}\right).$$

As a reference, we also compare these shrinkage estimators with $\tilde{\boldsymbol{\theta}}^{\text{OR}}$, the parametric oracle risk (OR) estimator, defined as plugging $\tilde{\lambda}^{\text{OR}} \boldsymbol{I}_p$ and $\tilde{\boldsymbol{\mu}}^{\text{OR}}$ into equation (2.6), where

$$\left(\tilde{\lambda}^{\text{OR}}, \tilde{\boldsymbol{\mu}}^{\text{OR}}\right) = \underset{0 \le \lambda \le \infty,\ \boldsymbol{\mu} \in \mathcal{L}_{\text{row}}(\boldsymbol{X})}{\text{argmin}} R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda, \boldsymbol{\mu}}\right)$$

and the expression of $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda, \boldsymbol{\mu}})$ is given in (2.7) with $\boldsymbol{B} = \lambda \boldsymbol{I}_p$. The oracle risk estimator $\tilde{\boldsymbol{\theta}}^{\text{OR}}$ cannot be used without the knowledge of $\boldsymbol{\theta}$, but it does provide a sensible lower bound of the risk achievable by any shrinkage estimator with the given parametric form.

For each simulation, we draw (A_i, θ_i) $(i = 1, 2, ..., p)$ independently from a distribution $\pi\left(A_i, \theta_i | \boldsymbol{X}_i, \boldsymbol{\beta}\right)$ and then draw Y_i given (A_i, θ_i). The shrinkage estimators are then applied to the generated data. This process is repeated 5000 times. The sample size p is chosen to vary from 20 to 500 with an increment of length 20. In the simulation, we fix a true but unknown $\boldsymbol{\beta} = (-1.5, 4, -3)^T$ and a known covariates $\boldsymbol{X}$, whose each element is randomly generated from $\text{Unif}\,(-10, 10)$. The risk performance of the different shrinkage estimators is given in Figure 2.2.

Example 1. The setting in this example is chosen in such a way that it reflects grouping in the data:

$$A_i \sim 0.5 \cdot 1_{\{A_i = 0.1\}} + 0.5 \cdot 1_{\{A_i = 0.5\}};$$

$$\theta_i | A_i \sim N\left(2 \cdot 1_{\{A_i = 0.1\}} + \boldsymbol{X}_i^T \boldsymbol{\beta}, 0.5^2\right);\ Y_i \sim N\left(\theta_i, A_i\right).$$

Here the normality for the sampling distribution of Y_i's is asserted. We can see that the four URE estimators perform much better than the two empirical Bayes ones and the James-Stein estimator. Also notice that both of the two (parametric and semiparametric) URE estimators that shrink towards $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$ is almost as good as the other two with general data-driven shrinkage location—largely due to the existence of covariate information. We note that this is quite different from the case of Xie et al. (2012), where without the covariate information the estimator that shrinks toward the grand mean of the data performs significantly worse than the URE estimator with general data-driven shrinkage location.

Example 2. In this example, we allow Y_i to depart from the normal distribution to illustrate that the performance of those URE estimators does not rely on the normality assumption:

$$A_i \sim \mathrm{Unif}\,(0.1, 1)\,;\;\; \theta_i = A_i + \boldsymbol{X}_i^T \boldsymbol{\beta};$$

$$Y_i \sim \mathrm{Unif}(\theta_i - \sqrt{3}A_i, \theta_i + \sqrt{3}A_i).$$

As expected, the four URE estimators perform better or at least as good as the empirical Bayes estimators. The EBMLE estimator performs the worst due to its sensitivity on the normality assumption. We notice that the EBMOM estimator in this example has comparable performance with the two parametric URE estimators, which makes sense as moment estimates are more robust to the sampling distribution. An interesting feature that we find in this example is that the positive part James-Stein estimator can beat the parametric oracle risk estimator and perform better than all the other shrinkage estimators for small or moderate p, even though the semiparametric URE estimators will eventually surpass the James-Stein estimator, as dictated by the asymptotic theory for large p. This feature of the James-Stein estimate is again quite different from the non-regression setting discussed in Xie et al. (2012), where the James-Stein estimate performs the worst throughout all of their examples. In both of our examples only the semiparametric URE estimators are robust to the different levels of heteroscedasticity.

We can conclude from these two simulation examples that the semiparametric URE estimators give competitive performance and are robust to the misspecification of the sampling distribution

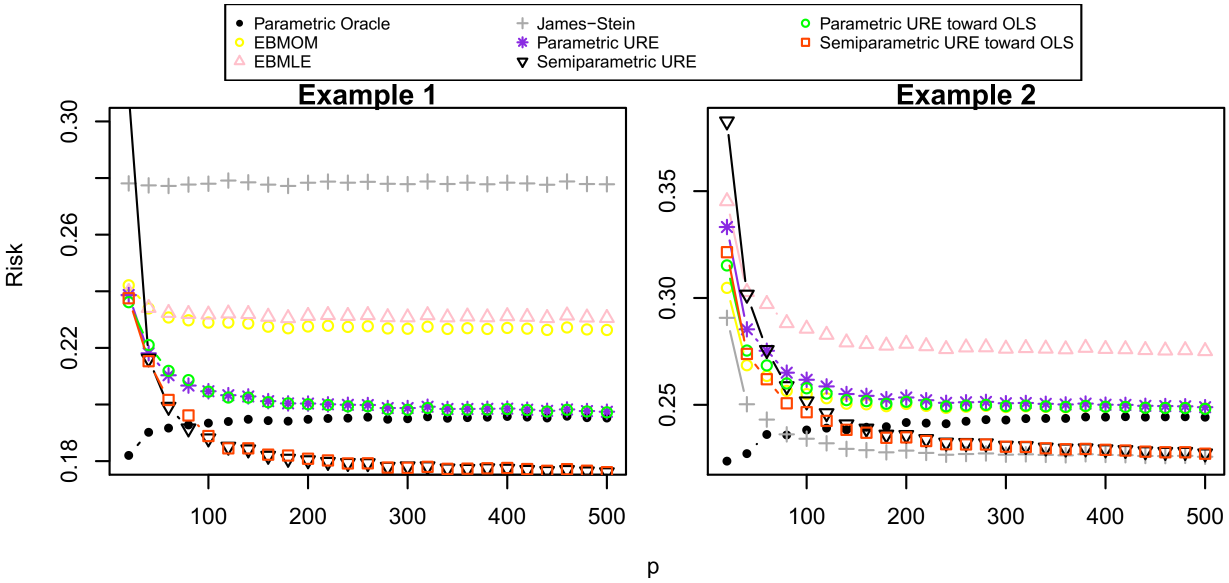


Figure 2.2: Comparison of the risks of different shrinkage estimators for the two simulation examples.

and the different levels of the heteroscedasticity. They thus could be useful tools in analyzing large-scale data for applied researchers.

2.7 Empirical Analysis

In this section, we study the baseball data set of Brown (2008). This data set consists of the batting records for all the Major League Baseball players in the 2005 season. As in Brown (2008) and Xie et al. (2012), we build a given shrinkage estimator based on the data in the first half season and use it to predict the second half season, which can then be checked against the true record of the second half season. For each player, let the number of at-bats be N and the successful number of batting be H, then we have $H_{ij} \sim Binomial(N_{ij}, p_j)$, where $i = 1, 2$ is the season indicator and $j = 1, \cdots, p$ is the player indicator. We use the following variance-stabilizing transformation Brown (2008) before applying the shrinkage estimators

$$Y_{ij} = \arcsin\sqrt{\frac{H_{ij} + 1/4}{N_{ij} + 1/2}},$$

which gives $Y_{ij} \dot{\sim} N(\theta_j, (4N_{ij})^{-1})$, $\theta_j = \arcsin\sqrt{p_j}$. We use

$$\mathrm{TSE}(\hat{\boldsymbol{\theta}}) = \sum_j (Y_{2j} - \hat{\theta}_j)^2 - \sum_j \frac{1}{4N_{2j}}.$$

as the error measurement for the prediction Brown (2008).

2.7.1 Shrinkage Estimation with Covariates

As indicated in Xie et al. (2012), there exists a significant positive correlation between the player's batting ability and his total number of at-bats. Intuitively, a better player will be called for batting more frequently; thus, the total number of at-bats will serve as the main covariate in our analysis. The other covariate in the data set is the categorical variable of a player being a pitcher or not.

Table 2.1 summarizes the result, where the shrinkage estimators are applied three times—to all the players, the pitchers only, and the non-pitchers only. We use all the covariate information (number of at-bats in the first half season and being a pitcher or not) in the first analysis, whereas in the second and the third analyses we only use the number of at-bats as the covariate. The values reported are ratios of the error of a given estimator to that of the benchmark naive estimator, which simply uses the first half season Y_{1j} to predict the second half Y_{2j}. Note that in Table 2.1, if no covariate is involved (i.e., when $\boldsymbol{X} = [1|\cdots|1]$), the OLS reduces to the grand mean of the training data as in Xie et al. (2012).

2.7.2 Discussion of the numerical result

There are several interesting observations from Table 2.1.

(i) A quick glimpse shows that including the covariate information improves the performance of essentially all shrinkage estimators. This suggests that in practice incorporating good covari-

	All		Pichers		Non-pichers	
p for estimation	567		81		486	
p for validation	499		64		435	
Covariates?	No	Yes	No	Yes	No	Yes
Naive	1	NA	1	NA	1	NA
Ordinary least squares (OLS)	0.852	0.242	0.127	0.115	0.378	0.333
Weighted least squares (WLS)	1.074	0.219	0.127	0.087	0.468	0.290
Parametric EBMOM	0.593	0.194	0.129	0.117	0.387	**0.256**
Parametric EBMLE	0.902	0.207	0.117	0.096	0.398	0.277
James-Stein	0.525	**0.184**	0.164	0.142	0.359	0.262
Parametric URE toward OLS	0.505	0.203	0.123	0.124	0.278	0.300
Parametric URE toward WLS	0.629	0.188	0.127	0.112	0.385	0.268
Parametric URE	0.422	0.215	0.123	0.130	0.282	0.310
Semiparametric URE toward OLS	0.409	0.197	0.081	0.097	0.261	0.299
Semiparametric URE toward WLS	0.499	**0.184**	0.098	**0.083**	0.336	**0.256**
Semiparametric URE	0.419	0.201	0.077	0.126	0.278	0.314

Table 2.1: Prediction errors of batting averages using different shrinkage estimators. Bold numbers highlight the best performance with covariate(s) in each case.

ates would significantly improve the estimation and prediction.

(ii) In general, shrinking towards WLS provides much better performance than shrinking toward OLS or a general data-driven location. This indicates the importance of a good choice of the shrinkage location in a practical problem. An improperly chosen shrinkage location might even negatively impact the performance. The reason that shrinking towards a general data-driven location is not as good as shrinking toward WLS is probably due to that the sample size is not large enough for the asymptotics to take effect.

(iii) Table 2.1 also shows the advantage of semiparametric URE estimates. For each fixed shrinkage location type (toward OLS, WLS, or general), the semiparametric URE estimator performs almost always better than their parametric counterparts. The only one exception is in the non-pitchers only case with the general data-driven location, but even there the performance difference is ignorable.

(iv) The best performance in all three cases (all the players, the pitchers only, and the non-pitchers only) comes from the semiparametric URE estimator that shrinks toward WLS.

(v) The James-Stein estimator with covariates performs quite well except in the pitchers only case, which is in sharp contrast with the performance of the James-Stein estimator without covariates. This again highlights the importance of covariate information. In the pitchers only case, the James-Stein performs the worst no matter one includes the covariates or not. This can be attributed to the fact that the covariate information (the total number of at-bats) is very weak for the pitchers only case; in the case of weak covariate information, how to properly estimate the shrinkage factors becomes the dominating issue, and the fact that the James-Stein estimator has only one uniform shrinkage factor makes it not competitive.

2.7.3 Shrinkage Factors

Figure 2.3 shows the shrinkage factors of all the shrinkage estimators with or without the covariates for the all-players case of Table 2.1. We see that the shrinkage factors are all reduced after including the covariates. This makes intuitive sense because the shrinkage location now contains the covariate information, and each shrinkage estimator uses this information by shrinking more toward it, resulting in smaller shrinkage factors.

2.8 Conclusion and Discussion

Inspired by the idea of unbiased risk estimate (URE) proposed in Xie et al. (2012), we extend the URE framework to multivariate heteroscedastic linear models, which are more realistic in practical applications, especially for regression data that exhibits heteroscedasticity. Several parallel URE shrinkage estimators in the regression case are proposed, and these URE shrinkage estimators are all asymptotically optimal in risk compared to other shrinkage estimators, including the classical empirical Bayes ones. We also propose semiparametric estimators and conduct simula-

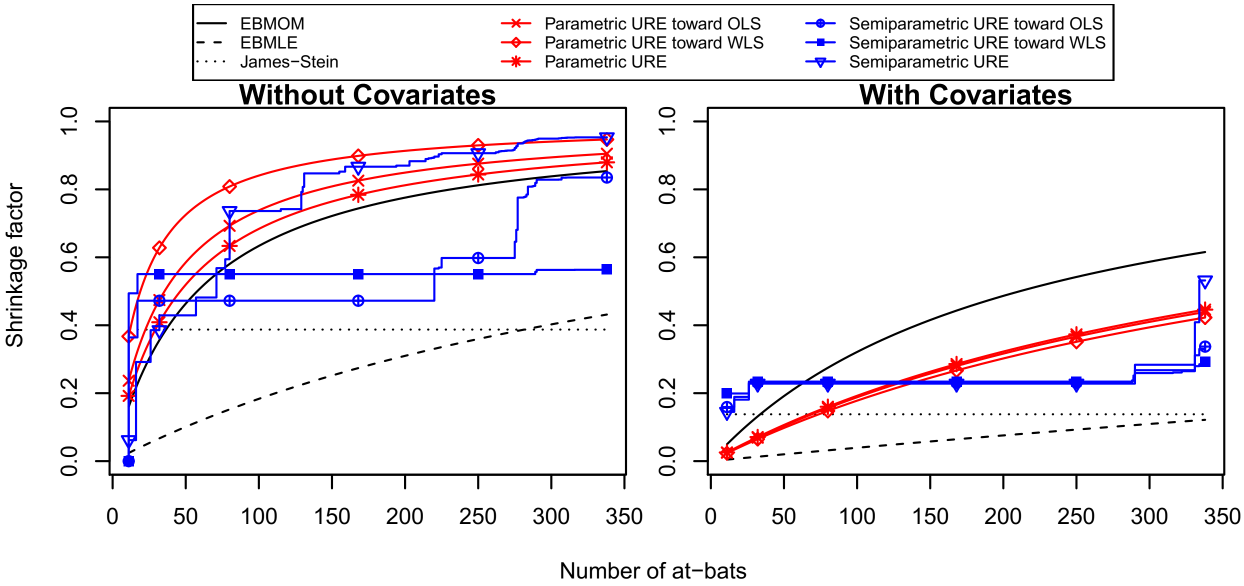


Figure 2.3: Plot of the shrinkage factors $\hat{\lambda}/\left(\hat{\lambda}+A_i\right)$ or $1-\hat{b}_i$ of all the shrinkage estimators for the case of all players.

tion to assess their performance under both normal and non-normal data. For data sets that exhibit a good linear relationship between the covariates and the response, a semiparametric URE estimator is expected to provide good estimation result, as we saw in the baseball data. It is also worth emphasizing that the risk optimality for the parametric and semiparametric URE estimators does not depend on the normality assumption of the sampling distribution of Y_i. Possible future work includes extending this URE minimization approach to simultaneous estimation in generalized linear models (GLMs) with canonical or more general link functions.

We conclude this Chapter by extending the main results to the case of weighted mean squared error loss.

Weighted mean squared error loss. One might want to consider the more general weighted mean squared error as the loss function:

$$l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}};\boldsymbol{\psi}\right)=\frac{1}{p}\sum_{i=1}^{p}\psi_i\left(\theta_i-\hat{\theta}_i\right)^2,$$

where $\psi_i>0$ are known weights such that $\sum_{i=1}^{p}\psi_i=p$. The framework proposed in this Chapter is straightforward to generalize to this case.

For Model II, we only need to study the equivalent problem by the following transformation

$$Y_i \to \sqrt{\psi_i} Y_i,\ \theta_i \to \sqrt{\psi_i}\theta_i,\ \boldsymbol{X}_i \to \sqrt{\psi_i}\boldsymbol{X}_i,\ A_i \to \psi_i A_i, \tag{2.17}$$

and restate the corresponding regularity conditions in Theorem 5 by the transformed data and parameters. We then reduce the weighted mean square error problem back to the same setting we study in this Chapter under the classical loss function (2.4).

Model I is more sophisticated than Model II to generalize. In addition to the transformation in equation (2.17), we also need $\lambda \to \psi_i \lambda$ in every term related to the individual unit i. Thus,

$$\sqrt{\psi_i}\theta_i | \boldsymbol{X}, \boldsymbol{\beta}, \lambda \overset{\text{indep.}}{\sim} N\left(\sqrt{\psi_i}\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda \psi_i\right),$$

so these transformed parameters $\sqrt{\psi_i}\theta_i$ are also heteroscedastic in the sense that they have different weights, while the setting we study before assumes all the weights on the θ_i are one. However, if we carefully examine the proof of Theorem 5 for the case of Model I, we can see that actually we do not much require the equal weights on the θ_i's. What is important in the proof is that the shrinkage factor for unit i is always of the form $A_i/\left(A_i + \lambda\right)$, which is invariant under the transformation $A_i \to \psi_i A_i$ and $\lambda \to \psi_i \lambda$. Thus, after reformulating the regularity conditions in Theorem 5 by the transformed data and parameters, we can still follow the same proof to conclude the risk optimality of URE estimators (parametric or semiparametric) even under the consideration of weighted mean squared error loss.

For completeness, here we state the most general result under the semiparametric setting for Model I. Let

$$\hat{\boldsymbol{\theta}}_{SP,\boldsymbol{\psi}}^{\text{URE}} = \left(\boldsymbol{I}_p - \text{diag}\left(\hat{\boldsymbol{b}}_{\boldsymbol{\psi}}^{\text{URE}}\right)\right)\boldsymbol{Y} + \text{diag}\left(\hat{\boldsymbol{b}}_{\boldsymbol{\psi}}^{\text{URE}}\right)\hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}}^{\text{URE}},$$

$$\mathrm{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right) = \frac{1}{p} \sum_{i=1}^{p} \psi_i \left(b_i^2 \left(Y_i - \mu_i\right)^2 + \left(1 - 2b_i\right) A_i \right),$$

$$\left(\hat{\boldsymbol{b}}_{\boldsymbol{\psi}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\psi}}^{\mathrm{URE}}\right) = \underset{\boldsymbol{b} \in \mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu} \in \mathcal{L}_{\mathrm{row}}(\boldsymbol{X})}{\operatorname{argmin}} \mathrm{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right).$$

Theorem 12. *Assume the following five conditions* (ψ-A) $\sum_{i=1}^{p} \psi_i^2 A_i^2 = O\left(p\right)$, ($\psi$-B) $\sum_{i=1}^{p} \psi_i^2 A_i \theta_i^2 = O\left(p\right)$, ($\psi$-C) $\sum_{i=1}^{p} \psi_i \theta_i^2 = O\left(p\right)$, ($\psi$-D) $p^{-1} \sum_{i=1}^{p} \psi_i^2 A_i \boldsymbol{X}_i \boldsymbol{X}_i^T$ *converges, and* (ψ-E) $p^{-1} \sum_{i=1}^{p} \psi_i \boldsymbol{X}_i \boldsymbol{X}_i^T \to \boldsymbol{\Omega}_{\boldsymbol{\psi}} > 0$. *Then we have*

$$\sup_{\boldsymbol{b} \in \mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu} \in \mathcal{L}_{\boldsymbol{\psi}}} \left| \mathrm{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b}, \boldsymbol{\mu}}; \boldsymbol{\psi}\right) \right| \underset{p \to \infty}{\to} 0 \text{ in } L^1,$$

where $\boldsymbol{\mu} \in \mathcal{L}_{\boldsymbol{\psi}}$ *if and only if* $\boldsymbol{\mu} \in \mathcal{L}_{\mathrm{row}}\left(\boldsymbol{X}\right)$ *and*

$$\sum_{i=1}^{p} \psi_i \mu_i^2 \leq M p^{\kappa} \sum_{i=1}^{p} \psi_i Y_i^2$$

for a large and fixed constant M *and a fixed exponent* $\kappa \in [0, 1/2)$. *As a corollary, for any estimator* $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p} = (\boldsymbol{I}_p - \mathrm{diag}(\hat{\boldsymbol{b}}_p))\boldsymbol{Y} + \mathrm{diag}(\hat{\boldsymbol{b}}_p)\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}\left(\boldsymbol{A}\right)$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathcal{L}_{\boldsymbol{\psi}}$, *we have*

$$\lim_{p \to \infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP, \boldsymbol{\psi}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p}\right) + \epsilon \right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p \to \infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP, \boldsymbol{\psi}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p}\right) \right) \leq 0.$$

You only have to do a very few things right

in your life so long as you don't do too many

things wrong.

Warren Buffett

3

Continuous Time Analysis of Fleeting Discrete Price Moves

A NOVEL MODEL OF FINANCIAL PRICES is proposed in this Chapter where: (i) prices are discrete; (ii) prices change in continuous time; (iii) a high proportion of price changes are reversed in a fraction of a second. Our model is analytically tractable and directly formulated in terms of the calendar time and price impact curve. The resulting càdlàg price process is a piecewise constant semimartingale with finite activity, finite variation and no Brownian motion component. We use

This Chapter is advised by and coauthored with Professor Neil Shephard.

moment-based estimations to fit four high frequency futures data sets and demonstrate the descriptive power of our proposed model. This model is able to describe the observed dynamics of price changes over three different orders of magnitude of time intervals.

3.1 Introduction

Extracting information from the order and trading flow in financial markets is important for trading at high and low frequencies, formulating policy and regulation and studying forensic finance. The distinctness about this area is the frequent focus on the very short term, usually over time intervals which may be much less than a second. At very short time scales, three essential aspects dominate: (i) prices are discrete, due to the tick structure of the market; (ii) prices change in continuous time; (iii) a high proportion of price changes are fleeting, reversed in a fraction of a second. However, the econometricians' cupboard is practically bare, for there are nearly no models or techniques that focus on all of the three features and put the role of the calendar time on center stage rather than the tick time.

In this paper we develop a novel continuous calendar time framework for prices out of a desire to capture these features in an analytically tractable but potentially semi-parametric manner. We will show that our model captures the serial dependence in price changes over three different time scales: 0.1 seconds, 1 seconds, 10 seconds and 1 minute.

Although our work is a distinctive move away from the existing literature, it will relate to a number of aspects that are often dealt with one at a time. Here we discuss some of this material.

Most of the econometric work on the modelling of high frequency financial data focuses on the times between trades and quote updates. This literature splits into two: the modelling of the conditional mean duration between events given past data and the modelling of the conditional intensity of trade arrivals given past data. It is reviewed by, for example, Engle (2000), Russell and Engle (2010) and Hautsch (2012). The former was initiated in Engle and Russell (1998) and contributions include Zhang et al. (2001), Hamilton and Jordá (2002) and Cipollini et al. (2009). The latter focuses around, for example, Russell (1999), Bowsher (2007) and Hautsch (2012), building on the stochastic analysis of Hawkes (1972).

There is much less econometric work on the discreteness of high frequency data. Papers that focus on discreteness include Rydberg and Shephard (2003), Russell and Engle (2006), Liesenfeld et al. (2006), Large (2011), Oomen (2005), Oomen (2006) and Griffin and Oomen (2008). Some of the early work on the impact of discreteness in practice includes Harris (1990), Gottlieb and Kalay (1985), Ball et al. (1985) and Ball (1988). A significant approach to deal with discreteness is to build continuous time models for prices on the positive half-line that are then rounded to induce discreteness, sometimes with extra additive measurement error. Examples include, for a variety of purposes, Hasbrouck (1999), Rosenbaum (2009), Delattre and Jacod (1997), Jacod (1996) and Li and Mykland (2014). Also note the statistical work by Kolassa and McCullagh (1990).

The most comparable literatures to our own include Bacry et al. (2013a), Bacry et al. (2013b), Fodra and Pham (2013a) and Fodra and Pham (2013b). See also Fauth and Tudor (2012). Bacry et al. model the evolution of price changes as the difference of two self-exciting and interacting simple counting processes. These multivariate Hawkes processes have intensities that react to previous moves, so an up move in the price will temporarily increase the intensity of a down move, creating the chance that the move will turn out to be fleeting. This elegant model only allows unit price moves, but could be extended, while the dynamics is tightly parameterized. Fodra and Pham directly assume an irreducible Markov chain structure on the sequence of price changes, which is less flexible as only the current price direction will impact the next jump direction.

Our paper has its intellectual roots in two papers. Barndorff-Nielsen et al. (2012) build Lévy processes (continuous time random walks) that are integer-valued. We are also inspired by the stationary integer-valued processes of Barndorff-Nielsen et al. (2014). Their processes are related to the up-stairs processes of Wolpert and Taqqu (2005) and the random measure processes of Wolpert and Brown (2011). Both of these processes are stationary. Barndorff-Nielsen et al. (2014) also bring out the relationship between their processes and M$/G/\infty$ queues (e.g. Lindley (1956), Reynolds (1968) and (Bartlett, 1978, Ch. 6.31)). They also connect these models to the so-called mixed moving average models of Surgailis et al. (1993). See also the work of Fuchs and Stelzer (2013). None of these papers can be used directly as a coherent model of high frequency data. Our paper fills this essential gap.

Our new approach will involve events arriving in continuous time, whose impacts on the prices may be fleeting and of variable size. The model is directly formulated in terms of the price impact of news. Each fleeting move is a temporary change in the price that has a random survival time until its impact disappears. The model allows a decomposition of the discrete price process into a continuous time random walk (due to permanent impacts) plus a temporary fleeting component (due to market microstructure noise). The resulting càdlàg price process will be a piecewise constant semimartingale with finite activity, finite variation and no Brownian motion component. It is also capable of generating negative autocorrelations for price changes that is consistent with the empirical observations. We have non-parametric freedom in choosing the level of dependence in the noises—which can even have long memory if this is needed in the data. Alternatively, the applied researcher can tightly parameterize the model if necessary.

In this paper our model is static: the parameters are time-invariant, not adapting to past data. This is an important deficiency, but a stochastic time-change can deal with most of these challenges. We will address them in a follow up paper. Our goal here is to set down a framework that is both empirically compelling and statistically scalable in the future work.

Finally, throughout our empirical work we have used trade prices. We could have used our model on the best bid or ask prices. This would have had the advantage that the best bid or ask are prices an investor can trade immediately, while trade prices are those which have been traded by someone in the past.

The outline of this paper is as follows. In Section 3.2 we set up the probability structure of our model and review a couple of building blocks from previous papers. In Section 3.3 we introduce the core of our contribution: defining our model for prices and providing an analysis of this process and the corresponding return sequence. In Section 3.4 we discuss the moment-based estimations for these models, while in Section 3.5 we apply these estimation methods to real data. Section 3.6 concludes. The Appendix has four sections. The first collects the proofs of the various theorems given in the main text of the paper as well as the details of some remarks. The second outlines how to compute probability mass functions of price changes using the inverse fast Fourier transform. The third details our data cleaning procedures. The fourth gives a nonparametric estimator of a part of our model.

3.2 Integer-Valued Stochastic Process in Continuous Time

Our model for (scaled, so they obey the tick structure) prices will be of the form

$$P_t = P_0 + \sum_{k=1}^{N_t} \varkappa_k, \tag{3.1}$$

where N_t is a counting process for the number of price moves and $\varkappa_k$ will be the corresponding integer-valued price move. Some of the price moves are permanent and some are fleeting, so (3.1) can be written as the sum of two (unobserved) components

$$P_t = P_0 + \sum_{k=1}^{N_t^{\mathrm{P}}} \varkappa_k^{\mathrm{P}} + \sum_{k=1}^{N_t^{\mathrm{T}}} \varkappa_k^{\mathrm{T}},$$

where N_t^{P} counts the number of moves due to permanent impacts, N_t^{T} due to temporary ones. The first component will be assumed to be a Lévy process, the second a stationary trawl process in continuous time. This links with the very large literature on the use of compound Poisson processes in financial econometrics, e.g. Press (1967). However, here we allow a fraction of the jumps to be fleeting, so the resulting counting process N_t is not simply a Poisson process.

In this Section, we will establish a very flexible stochastic framework for this type of model.

3.2.1 Poisson random measure

Our mathematical framework will revolve around (i) events arriving in continuous time, (ii) events whose impacts may be fleeting with a random survival time and (iii) events of variable size and sign. To generate these events, it is natural to base the underlying randomness on a three dimensional Poisson random measure N (see, e.g., Kingman (1993) for a review) with intensity measure

$$\mathbb{E}\left(N\left(\mathrm{d}y, \mathrm{d}x, \mathrm{d}s\right)\right) = \nu(\mathrm{d}y)\mathrm{d}x\mathrm{d}s.$$

Here s is time (with arrivals randomly scattered on $\mathbb{R}$), x is random source (uniformly scattered over $[0,1]$), which will drive the survival times of the fleeting events, and y marks integer size (with sign) of events. These names will become clearer later in Figure 3.1 and 3.2.

Price moves can be up or down, but zero is ruled out. Thus the size of the events will be assumed to have a Lévy measure $\nu(\mathrm{d}y)$ concentrated on $y \in \mathbb{Z}\backslash\left\{0\right\}$, the non-zero integers. With no confusions, we will sometimes abuse the notation $\nu\left(y\right)$ to denote the mass of the Lévy measure centered at y, so

$$\nu\left(\mathrm{d}y\right) = \sum_{y\in\mathbb{Z}\backslash\{0\}} \nu\left(y\right)\delta_{\{y\}}\left(\mathrm{d}y\right),$$

where $\delta_{\{y\}}\left(\mathrm{d}y\right)$ is the Dirac point mass measure centered at y. Throughout this paper, we assume that $\|\nu\| = \int_{-\infty}^{\infty}\nu\left(\mathrm{d}y\right) = \sum_{y\in\mathbb{Z}\backslash\{0\}}\nu\left(y\right) < \infty$.

Remark 13. *We will see in a moment that the mass $\nu(y)$ represents the intensity of events of size y, so in aggregate the Lévy measure ν will simultaneously controls the scope of all the possible jumping sizes in addition to their individual intensity.*

3.2.2 Lévy basis and Lévy process

Our model will be based on the resulting homogeneous Lévy basis on $[0,1] \times \mathbb{R} \longmapsto \mathbb{Z} \backslash \{0\}$, which records the size $y \in \mathbb{Z} \backslash \{0\}$ at each point in time $s \in \mathbb{R}$ and height $x \in [0,1]$. It is given by

$$L(\mathrm{d}x, \mathrm{d}s) = \int_{-\infty}^{\infty} y N(\mathrm{d}y, \mathrm{d}x, \mathrm{d}s), \qquad (x,s) \in [0,1] \times \mathbb{R},$$

and for any Borel measurable set $S \subseteq [0,1] \times \mathbb{R}$ we let

$$L(S) = \int_{[0,1]\times\mathbb{R}} 1_S(x,s) L(\mathrm{d}x, \mathrm{d}s),$$

where 1_S is the indicator function of S. To connect with the later discussion, a Lévy process generated from L can be defined as

$$L_t = L(R_t) = \int_0^t \int_0^1 L(\mathrm{d}x, \mathrm{d}s),$$

where $R_t = [0,1] \times (0,t]$ is a rectangle that grows with t, so L_t just counts up the points in the Lévy basis from time 0 to time t.

Example 1. *Suppose that $\nu(\mathrm{dy}) = \|\nu\| \left(0.5 \times \delta_{\{1\}}(\mathrm{dy}) + 0.5 \times \delta_{\{-1\}}(\mathrm{dy})\right)$. Then $\|\nu\|$ is the arrival rate of events in time, each with a random source for the survival time and having size ± 1 with equal probability. Figure 3.1 plots a Skellam Lévy basis L using $\|\nu\| = 7$, taking on $1, -1$ with black and white dots respectively. The lower panel shows the corresponding Skellam Lévy process, which is the difference of two independent Poisson processes with intensity $\|\nu\|/2$.*

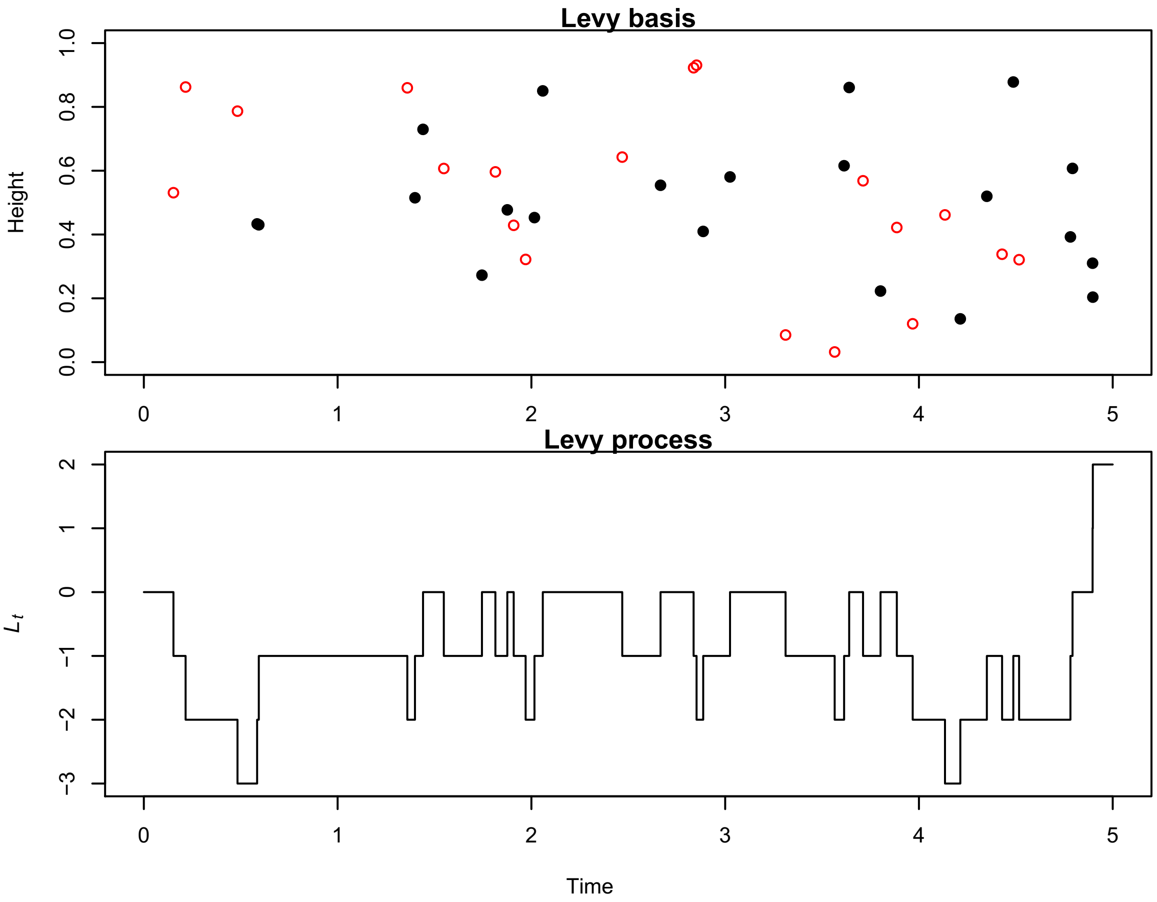


Figure 3.1: Top: Lévy basis $L(\mathrm{d}x, \mathrm{d}s)$, where the horizontal axis s is time and the vertical axis x is random source for the survival time (denoted height in this picture), which plays no rule in this construction of the Lévy process in the lower panel as all permanent arrivals survive forever. Black dots denote 1, white ones -1. Bottom: The corresponding Lévy process, which sums up all the effects in the Lévy basis (in the upper panel) from time 0 to time t, while the vertical axis here is the value of the Lévy process, which jumps up by 1 by the effect of black dots and down by 1 by white ones.

3.2.3 Stationary trawl process

To introduce fleeting moves, the random sources for the survival times in the Lévy basis will be exploited. We start from a fixed shape* $A \subseteq [b, 1] \times (-\infty, 0]$, where $b \in [0, 1]$ is called the permanence parameter. Throughout we assume that the area of A, $leb\,(A)$, is finite. Barndorff-Nielsen et al. (2014) call A a trawl for the case of $b = 0$, which is the core of their stationary integer-

*For technical reasons, we need to assume that the fixed set A is closed on the right and open on the left, that is, for every $x \in [b, 1]$, all the set $A \cap \{(x, s) : s \leq 0\}$ must be a union of half-closed intervals of the form $(a, b]$. This is enforced so the resulting jump process is càdlàg. Besides, we need to assume that the projection of A on the vertical axis has Lebesgue measure b so the parameter b is well-defined and statistically identifiable.

valued processes. Here we call A a squashed trawl, a minor variant on their idea.

Definition 1. *A squashed trawl A defined by a trawl function d is obtained from*

$$A = \{(x, s) : s \leq 0, b \leq x < d(s)\},$$

where $d : (-\infty, 0] \longmapsto [b, 1]$ is continuous and monotonically increasing ($d(s_1) \leq d(s_2)$ for all $s_1 \leq s_2 \leq 0$) and satisfies the following regularity conditions: $d(-\infty) = \lim_{s \to -\infty} d(s) = b$, $d(0) = 1$ and $\int_{-\infty}^{0} (d(s) - b)\, \mathrm{d}s < \infty$.

We now drag the set A through time without changing its shape

$$A_t = A + (0, t) = \{(x, s) : s \leq t, b \leq x < d(s - t)\}, \qquad t \geq 0.$$

Notice that $leb(A_t) = leb(A) < \infty$ for all $t \geq 0$. Then the stationary (trawl) process is defined as $L(A_t)$ for $t \geq 0$. In a moment this will be a component of our proposed price process.

Example 2. *The upper panel of Figure 3.2 illustrates A_t when $d(s) = 0.5 + (1 - 0.5)e^{2s}$. The middle panel of Figure 3.2 also shows $L(A_t)$ when L is a Skellam basis, which sums up all the effects (both positive and negative) captured by (or surviving in) the trawl. Dynamically, $L(A_t)$ will move up by 1 if the moving squashed trawl A_t either captures one positive event that has height above b or releases a negative one; conversely, it will move down by 1 if vice versa. Also notice that $L(A_0)$ might not be necessarily zero.*

Throughout we use $\kappa_j(X)$ as a generic notation for the j-th cumulant of an arbitrary random variable X. Recall that $L_1 = L(R_1) = \int_0^1 \int_0^1 L(\mathrm{d}x, \mathrm{d}s)$. In the following Proposition, we rephrase the key properties of the stationary process $L(A_t)$ mentioned in Barndorff-Nielsen et al. (2014) under the squashed trawl variant.

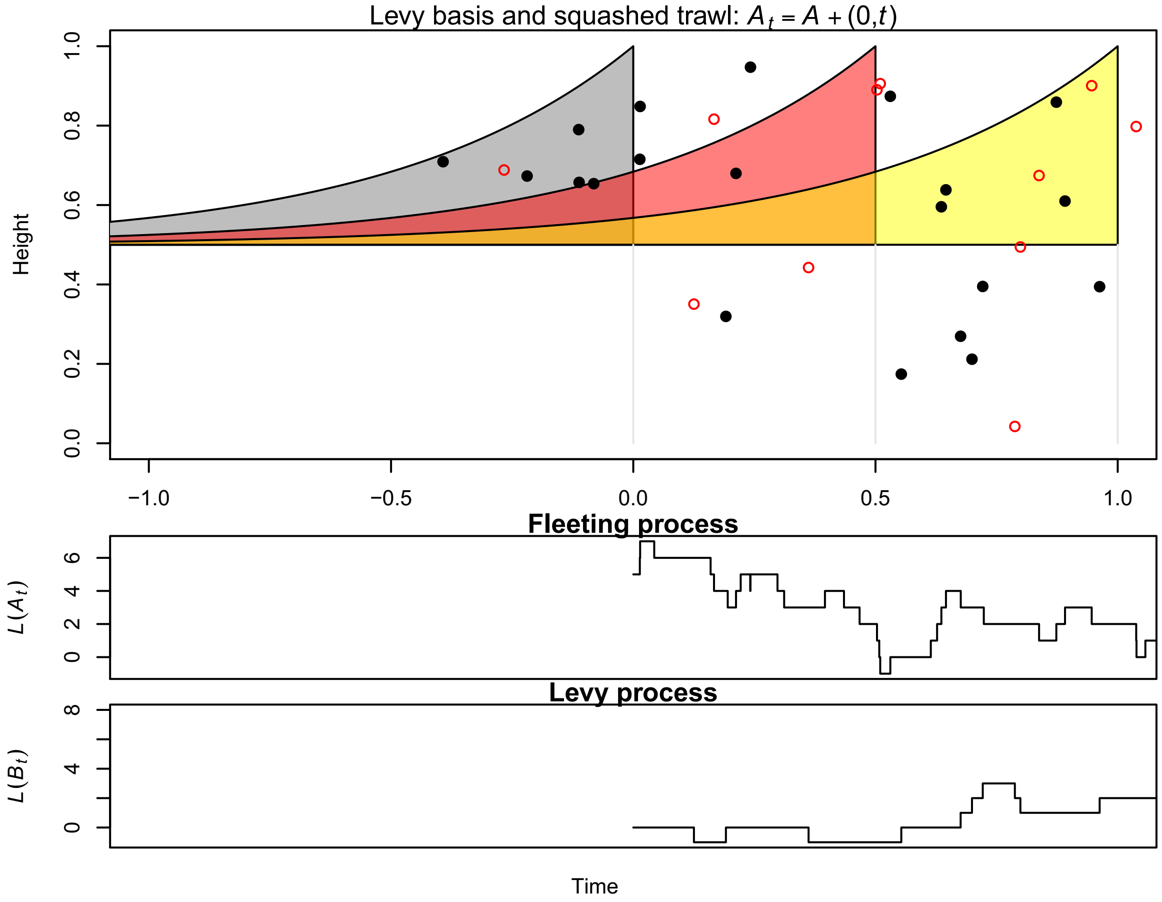


Figure 3.2: A moving squashed trawl A_t is joined by the Lévy basis $L(\mathrm{d}x, \mathrm{d}s)$, where the horizontal axis s is time and the vertical axis x is height. The shaded area is an example of the trawl A generated by the trawl function d, while we also show the outlines of A_t when $t = 1/2$ and $t = 1$. Also shown below is the implied stationary process $L(A_t)$ and the Lévy process $L(B_t)$ for $t \geq 0$, where $B_t = [0, b) \times (0, t]$ is a rectangle that grows with t and covers the region below b.

Proposition 13. *If $leb\,(A) < \infty$, then $L\,(A_t)$ is well-defined and* strictly stationary*. If $\kappa_2\,(L_1) < \infty$ as well, then it is* covariance stationary *and for $t > s$*

$$\mathrm{Cov}\,(L\,(A_t)\,, L\,(A_s)) = leb\,(A_{t-s} \cap A)\,\kappa_2\,(L_1)\,,\ \ \mathrm{Cor}\,(L\,(A_t)\,, L\,(A_s)) = \frac{leb\,(A_{t-s} \cap A)}{leb\,(A)}.$$

Furthermore, for any $t \geq 0$,

$$leb\,(A_t \cap A) = \int_{-\infty}^{-t} (d\,(s) - b)\,\mathrm{d}s \tag{3.2}$$

is monotonically decreasing as t increases.

3.3 Integer-Valued Price Process with Fleeting Moves

3.3.1 Definition

We now turn to the main contribution of this paper. Our integer-valued price process is

$$P_t = V_0 + L(C_t) = V_0 + L\left(A_t\right) + L\left(B_t\right), \qquad t \geq 0,$$

where we recall that $A_t = A + (0, t)$ and

$$B_t = [0, b) \times (0, t],\ C_t = A_t \cup B_t.$$

Here V_0 is a non-negative integer; L is a Lévy basis; $L\left(A_t\right)$ is a stationary integer-valued process that controls the fleeting movements of the price; $V_0 + L\left(B_t\right)$ is an integer-valued Lévy process (initiating at V_0, which is aggregated from the permanent arrivals in the past) that represents a non-stationary component of the price process. Recall that $L\left(B_t\right) = \int_0^t \int_0^b L\left(\mathrm{d}x, \mathrm{d}s\right)$, $t \geq 0$.

Example 3 (Continued from Example 2)**.** *The lower panel of Figure 3.2 shows the corresponding Skellam Lévy process $L\left(B_t\right)$. Notice that there are no permanent events in the negative time because they have been taken into account in V_0. Over short time scales it is hard to tell the difference between these two processes $L\left(A_t\right)$ and $L\left(B_t\right)$, but over long time scales they are starkly different. For any event arrival, if the random height x—not size y—is above b, then this effect stays in A_t temporarily and hence is fleeting; if the height is below b, then this effect is always in B_t and hence permanent.*

This càdlàg price process has finite activity (i.e. finite number of jumps in any finite interval of time, due to the Lévy basis being of finite activity), is piecewise-constant (i.e. jumps only when there are arrivals or departures) and consequently has finite variation. Thus the model is in keeping with the empirical data.

Remark 14. *The integer-valued price process P_t is a semimartingale with respect to its natural filtration. The details can be found in Appendix A.3. Here we especially point out that a semimartingale model that allows the fleeting behavior is atypical in the literature of market microstructure.*

Remark 15. *In this model some price moves have permanent impact. Others are fleeting, being reversed rapidly. The lifetime of an arrival event is determined by the trawl function. Assume that the trawl function d is strictly increasing and hence invertible. Then we can think of $G(s) = 1 - d(-s)$ (with $G(\infty) = 1$) as the cumulative distribution function of the lifetime for $s \geq 0$. Thus, for $U \backsim U(0,1)$, the standard uniform distribution, $G^{-1}(U)$ means the lifetime of an arrival event with random height U. When $U \leq b$, then $G^{-1}(U) = \infty$, meaning it is permanent. For $U > b$ then the event will last $G^{-1}(U) < \infty$, meaning it is fleeting.*

Remark 16. *If a new piece of news arrives at time t, it impacts the price through the arrival of a new point in the Lévy basis. For concreteness of exposition here, suppose it has unit impact. Then the expected impact of this individual event at time $t+s$ is $d(-s)$, where $s \geq 0$. Hence the trawl function directly describes the* price impact curve *of news arrivals. It is tempting to label d the price impact function, but we continue with the trawl nomenclature. The permanent impact of the unit news is thus b.*

3.3.2 Distribution of price changes

The following Theorem characterizes the distribution of price changes over a time length t.

Theorem 14. *Let $A \backslash B$ be set subtraction (all elements of A except those that are also in B). Then*

$$P_t - P_0 = L(C_t) - L(C_0) = L(C_t \backslash C_0) - L(C_0 \backslash C_t),$$

where $L\left(C_t \backslash C_0\right)$ is independent of $L(C_0 \backslash C_t)$. Consequently the logarithmic characteristic function of returns is

$$C\left(\theta \ddagger P_t - P_0\right) = btC\left(\theta \ddagger L_1\right) + leb(A_t \backslash A)\left(C\left(\theta \ddagger L_1\right) + C\left(-\theta \ddagger L_1\right)\right), \quad \textit{where}$$

$$C\left(\theta \ddagger L_1\right) = \log \mathbb{E}\left(e^{\mathrm{i}\theta L_1}\right), \qquad \mathrm{i} = \sqrt{-1}, \qquad L_1 = \int_0^1 \int_0^1 L(\mathrm{d}x, \mathrm{d}s).$$

Furthermore, if the j-th cumulant of L_1 exists, then

$$\kappa_j(P_t - P_0) = bt\kappa_j(L_1), \qquad j = 1, 3, 5, ...,$$

$$\kappa_j(P_t - P_0) = \left(bt + 2leb\left(A_t \backslash A\right)\right)\kappa_j(L_1), \qquad j = 2, 4, 6, ...$$

Remark 17. *Notice that $C_t \backslash C_0$ has the physical interpretation of arrivals during the time period 0 to t for both positive and negative effects; $C_0 \backslash C_t$ are departures instead. Further, the equalities*

$$leb(A_t \backslash A) = leb(A) - leb(A_t \cap A) = leb(A \backslash A_t) = \int_{-t}^{0} \left(d\left(s\right) - b\right) \mathrm{d}s \tag{3.3}$$

are often helpful in calculations.

Remark 18. *The probability mass function of $P_t - P_0$ can be computed using the characteristic function and the inverse fast Fourier transform. The details can be found in the Appendix B.1.*

Remark 19. *Even though our model is written down for the study of high frequency data, it can easily connect back to those diffusion based models that are commonly used to study data at less high frequency. Theorem 14 further implies that the fleeting price process becomes a Brownian motion at lower frequency. Precisely, if $\kappa_2\left(L_1\right) < \infty$ and $X_t^{(c)} = \left(\kappa_2\left(L_1\right)c\right)^{-1/2}\left(P_{ct} - P_0 - bct\kappa_1\left(L_1\right)\right)$, then $X_{\cdot}^{(c)} \overset{\mathcal{L}}{\to} W_{\cdot}$ as $c \to \infty$, where $W_{\cdot}$ is a Wiener process or a standard Brownian motion.*

Let $\Delta P_t = P_t - P_{t-}$ be the instantaneous jump (or return) of the price process at time t. By the instantaneous jumping distribution, we mean the probability of $\Delta P_t = y$ given that $\Delta P_t \neq 0$ for $y \in \mathbb{Z} \backslash \left\{0\right\}$. In the following we give a closed-form expression for this distribution.

Theorem 15. *The instantaneous jumping distribution is*

$$\mathbb{P}\left(\Delta P_t = y | \Delta P_t \neq 0\right) = \frac{\nu\left(y\right) + \nu\left(-y\right)\left(1-b\right)}{\left(2-b\right)\left\|\nu\right\|}. \tag{3.4}$$

Notice that the trawl function d in the fleeting component has no impact on the instantaneous jumping distribution: what is important is b, which controls the amount of potential departures among all the arrival jumps. Besides, the left-hand side of equation (3.4) can be easily estimated from the data, so we might in turn estimate ν and b by simple moment-matching. To calibrate the trawl, Theorem 14 imply the easy use of sample cumulants across different t to infer the shape of $leb\left(A_t \backslash A\right)$ and hence d. We will see these in Section 3.4 later.

3.3.3 Autocorrelation structure of price changes

Theorem 16 captures the linear dependence in the price changes.

Theorem 16. *Assume that $\kappa_2(L_1) < \infty$. Then the price changes have the autocorrelation structure, for some sampling interval $\delta > 0$ and $k = 1, 2, ...$*

$$\begin{aligned}
\gamma_k &= \text{Cov}\left(\left(P_{(k+1)\delta} - P_{k\delta}\right), \left(P_\delta - P_0\right)\right) \\
&= \left(leb(A_{(k+1)\delta} \backslash A) - 2leb(A_{k\delta} \backslash A) + leb(A_{(k-1)\delta} \backslash A)\right)\kappa_2(L_1), \\
\rho_k &= \text{Cor}\left(\left(P_{(k+1)\delta} - P_{k\delta}\right), \left(P_\delta - P_0\right)\right) \\
&= \frac{leb(A_{(k+1)\delta} \backslash A) - 2leb(A_{k\delta} \backslash A) + leb(A_{(k-1)\delta} \backslash A)}{b\delta + 2leb(A_\delta \backslash A)}.
\end{aligned}$$

Corollary 17. *$\rho_k \leq 0$ for all $k = 1, 2, ...$ This inequality becomes strict when d is strictly increasing (i.e. $d\left(s_1\right) < d\left(s_2\right)$ for all $s_1 < s_2 \leq 0$).*

Remark 20. *For a pure Lévy process ($b = 1$), $leb\left(A_t \backslash A\right) = 0$ for all t, so clearly $\rho_k = 0$ for all*

$k = 1, 2, \ldots$, as expected. On the other hand, equation (3.3) implies

$$\lim_{\delta \to 0} \frac{leb\left(A_{l\delta} \backslash A\right)}{\delta} = (1-b)\, l, \ \lim_{\delta \to \infty} leb\left(A_{l\delta} \backslash A\right) = leb\left(A\right), \qquad l = 1, 2, \ldots,$$

so it is easy to see that, for any fixed $k = 1, 2, \ldots$,

$$\lim_{\delta \to 0} \rho_k = \lim_{\delta \to \infty} \rho_k = 0.$$

Thus, Corollary 17 implies that ρ_k is not a monotonic function of the sampling interval δ. This matches with the empirical data, which we will see later in Figure 3.7.

3.3.4 Power variation

Quadratic variation plays a central role in stochastic analysis and modern finance (e.g. Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002)). For any $r \geq 0$, we define the r-th power Lévy basis as

$$\Sigma(\mathrm{d}x, \mathrm{d}s; r) = \int_{-\infty}^{\infty} |y|^r \, N(\mathrm{d}y, \mathrm{d}x, \mathrm{d}s)$$

with mean measure

$$\mu(\mathrm{d}x, \mathrm{d}s; r) = \mathrm{d}x\mathrm{d}s \int_{-\infty}^{\infty} |y|^r \, \nu(\mathrm{d}y),$$

assuming that $\int_{-\infty}^{\infty} |y|^r \, \nu(\mathrm{d}y) < \infty$. Theorem 18 relates Σ to

$$\{P\}_t^{[r]} = \lim_{\delta \to 0} \sum_{k=1}^{t/\delta} \left| P_{k\delta} - P_{(k-1)\delta} \right|^r = \sum_{0 < s \leq t} |\Delta P_s|^r \,,$$

the r-th (unnormalized) power variation, which was formalized in finance by Barndorff-Nielsen and Shephard (2004). The special case of $r = 2$ yields the quadratic variation. Notice that in our model we can compute $\{P\}_t^{[r]}$ exactly, just using the price path. It is finite for all $r \geq 0$ with probability one. This contrasts with the vast majority of work in econometrics that would take $\{P\}_t^{[r]}$ as infinity due to the impact of market microstructure.

Theorem 18. *For any $r \geq 0$, the r-th power variation is*

$$\{P\}_t^{[r]} = \Sigma(B_t; r) + Z_t^{[r]}, \qquad B_t = [0, b) \times (0, t],$$

$$Z_t^{[r]} = \Sigma(H_t; r) + \Sigma(G_t; r), \qquad H_t = [b, 1] \times (0, t],\ G_t = (H_t \cup A) \setminus A_t.$$

Furthermore, their expectations are

$$\mathbb{E}\left(\{P\}_t^{[r]}\right) = (2 - b)\, t \int_{-\infty}^{\infty} |y|^r\, \nu(\mathrm{d}y). \tag{3.5}$$

Remark 21. *Like (3.4), (3.5) does not feature the trawl function, as each arrival is joined by a departure. Hence it is always robust to the details of d. Further,*

$$\mathbb{E}\left(\{P\}_t^{[r]}\right) = \mathbb{E}\left(\{P\}_t^{[0]}\right) \int_{-\infty}^{\infty} |y|^r\, \frac{\nu\,(\mathrm{d}y)}{\|\nu\|}.$$

Notice that $\{P\}_t^{[0]}$ counts the total number of jumps of the process P up to time t, so throughout we call it the counting process of price moves. It will also play an important role in Section 3.4 for the construction of our moment-based estimate for the model parameters.

We think of the random $Z_t^{[r]}$, which is finite with probability one, as the component of power variation due to fleeting moves in prices, for

$$\{P\}_t^{[r]} - \{L\,(B_t)\}_t^{[r]} = Z_t^{[r]}$$

is the asymptotic stochastic bias of the power variation.

High frequency econometricians would typically think of terms like $Z_t^{[2]}$ as the driver of the bias in realized variance due to market microstructure effects (e.g. Hansen and Lunde (2006), Zhang (2006), Jacod et al. (2009), Mykland and Zhang (2012) and Barndorff-Nielsen et al. (2008)), but it is typically infinite in their studies while here and empirically it is finite with probability one[†].

We recall from Theorem 14 that

$$\mathbb{E}(P_t - P_0) = bt\kappa_1(L_1), \qquad \mathrm{Var}(P_t - P_0) = \{bt + 2leb\,(A_t \backslash A)\}\,\kappa_2\,(L_1)\,.$$

We now think about returns over the time interval $[0, T]$, so the realized variance is

$$RV^{(n)} = \sum_{k=1}^{n} (P_{k\delta_n} - P_{(k-1)\delta_n})^2, \qquad \delta_n = \frac{T}{n}.$$

Proposition 19. *Assume that $\kappa_2(L_1) < \infty$. Then*

$$\mathbb{E}\left(RV^{(n)}\right) = \left(b + 2\frac{leb\,(A_{\delta_n} \backslash A)}{\delta_n}\right) T\kappa_2\,(L_1) + b^2 T\delta_n \kappa_1^2\,(L_1)\,.$$

We can set the context of Proposition 19 by discussing the two extremes $n = 1$ and $n \to \infty$ for a large T. For $n = 1$, as $T \to \infty$,

$$\begin{aligned}\mathbb{E}\left(RV^{(1)}\right) &= \left(b + 2\frac{leb\,(A_T \backslash A)}{T}\right) T\kappa_2\,(L_1) + b^2 T^2 \kappa_1^2\,(L_1)\\ &\approx bT\kappa_2\,(L_1) + b^2 T^2 \kappa_1^2\,(L_1)\\ &= \kappa_2\,(L\,(B_T)) + (\kappa_1\,(L\,(B_T)))^2 = \mathbb{E}\left(L\,(B_T)^2\right),\end{aligned}$$

where the second line uses $leb\,(A_T \backslash A) \approx leb\,(A)$. For $n \to \infty$ and a fixed T,

$$\lim_{n\to\infty} \mathbb{E}\left(RV^{(n)}\right) = (2 - b)\,T\kappa_2\,(L_1)\,.$$

[†] Econometricians use a variety of models for market microstructure noise. Typically the noise appears each time a trade happens, e.g. in Zhou (1996) the noises are i.i.d. with a zero mean. Hence we can think of these types of models as purely statistical measurement error models. In more recent times, the i.i.d. assumption has been generalized to allow some levels of temporal dependence and volatility clustering, but all in tick time instead of the calendar time. All of these models of noise have the power variations being infinity. There is another set of papers that think of prices as being a rounded version of a semimartingale. This is closer to our paper, but here the level of dependence in price moves is entirely dependent on the size of the ticks in comparison to the volatility of the semimartingale. This is insufficiently flexible to fit the data. Another set of papers round a semimartingale with additive measurement noise, but again this has infinite power variation, which does not coincide with the empirical observations.

Therefore, in this model the realized variance and the volatility of price changes are highly distorted by the fleeting component. A variance signature plot ($RV^{(T/\delta)}$ against δ) for our model will start out high around $(2-b)\,T\kappa_2\,(L_1)$ (the expected quadratic variation of the price process) for large n (dense sampling) and tend downwards to approximately $bT\kappa_2\,(L_1)$ (the expected quadratic variation of the Lévy process component, assuming that $\kappa_1\,(L_1)$ being very small). A minor variant of this type of plots, which we will discuss in Remark 24, can be found in Figure 3.8 later in our empirical work.

3.3.5 Generalized compound representation

As the price process is of finite activity, it can be usefully written as a generalized compound process, driven by the counting process of price moves. Here we detail this. First recall that $G(s) = 1 - d\,(-s)$ (with $G\,(\infty) = 1$) denotes the cumulative distribution function of the lifetime for $s \geq 0$.

$L(A_0)$ is built out of N^{A*} initial surviving events, who arrive at times $\tau_1^{A*} < ... < \tau_{N^*}^{A*} \leq 0$ and jump with sizes $\varkappa_1^{A*}, ..., \varkappa_{N^*}^{A*}$. Each arrival has a lifetime $G^{-1}(U_1^{A*}), ..., G^{-1}(U_{N^{A*}}^{A*})$, where $\tau_j^{A*} + G^{-1}(U_j^{A*}) > 0$ and $U_j^{A*} \overset{\text{i.i.d.}}{\sim} U(b,1)$. Thus we can write $L(A_0) = \sum_{j=1}^{N^{A*}} \varkappa_j^{A*} 1_{\tau_j^{A*}+G^{-1}(U_j^{A*})>0}$. When $\varkappa_j^{A*} = 1$ for all j, this representation has a close connection to a M/G/∞ queue (i.e. Markov arrivals, with a fixed service time distribution G, but with an infinite number of servers).

As time progresses some events die and the initial values thin down to $\sum_{j=1}^{N^{A*}} \varkappa_j^{A*} 1_{\tau_j^{A*}+G^{-1}(U_j^{A*})>t}$ while new ones are born $\sum_{j=1}^{N_t^A} \varkappa_j^{A} 1_{\tau_j^{A}+G^{-1}(U_j^{A})>t}$, where N_t^A is the number of births from time 0 to time t with heights greater than b. The corresponding τ_j^A's and $\varkappa_j^A$'s are the arrival times of these events and size of the moves. Thus the stationary process is

$$L\,(A_t) = \sum_{j=1}^{N^{A*}} \varkappa_j^{A*} 1_{\tau_j^{A*}+G^{-1}(U_j^{A*})>t} + \sum_{j=1}^{N_t^A} \varkappa_j^{A} 1_{\tau_j^{A}+G^{-1}(U_j^{A})>t}, \qquad t \geq 0.$$

The corresponding impact of the permanent changes is a compound Poisson process $L(B_t) = \sum_{j=1}^{N_t^B} \varkappa_j^B$, where N_t^B counts the number of permanent arrivals up to time t and τ_j^B's and $\varkappa_j^B$'s are the corresponding arrival times and jump sizes. We also write τ_k to be any one of the jumping times from resulted chronologically from both the arrivals and departures; similarly for $\varkappa_k$. Then $N_t = \#\{k : \tau_k \leq t\}$ counts the total number of jumps of the price process up to time t.

All these imply that

$$P_t = V_0 + \sum_{j=1}^{N^{A*}} \varkappa_j^{A*} 1_{\tau_j^{A*}+G^{-1}(U_j^{A*})>t} + \sum_{j=1}^{N_t^A} \varkappa_j^A 1_{\tau_j^A+G^{-1}(U_j^A)>t} + \sum_{j=1}^{N_t^B} \varkappa_j^B = P_0 + \sum_{k=1}^{N_t} \varkappa_k,$$

which is called a generalized compound representation.

3.3.6 Parameterized trawl function

To fit this type of model using data, it is sometimes helpful to index the trawl function by a small number of parameters. Throughout we work within the following framework.

Definition 2. *A superposition trawl function has*

$$d(s) = b + (1-b)\int_0^\infty e^{\lambda s}\pi(\mathrm{d}\lambda), \qquad s \leq 0, \tag{3.6}$$

where π is an arbitrary probability measure on $(0,\infty)$. We constrain the superposition class to where $\int_0^\infty \lambda^{-1}\pi(\mathrm{d}\lambda) < \infty$.

Whatever the probability measure π the resulting d always exists since $0 \leq \int_0^\infty e^{\lambda s}\pi(\mathrm{d}\lambda) \leq \int_0^\infty \pi(\mathrm{d}\lambda) = 1$, as $s \leq 0$. The constraint $\int_0^\infty \lambda^{-1}\pi(\mathrm{d}\lambda) < \infty$ is needed to ensure that the area of A is finite, for this area is

$$\begin{aligned} leb(A) &= \int_{-\infty}^0 \int_b^{d(s)} \mathrm{d}x\mathrm{d}s = \int_{-\infty}^0 (d(s) - b)\,\mathrm{d}s = (1-b)\int_{-\infty}^0 \int_0^\infty e^{\lambda s}\pi(\mathrm{d}\lambda)\,\mathrm{d}s \\ &= (1-b)\int_0^\infty \int_{-\infty}^0 e^{\lambda s}\mathrm{d}s\pi(\mathrm{d}\lambda) = (1-b)\int_0^\infty \frac{1}{\lambda}\pi(\mathrm{d}\lambda). \end{aligned} \tag{3.7}$$

Using equation (3.2) the superposition framework (3.6) has

$$leb(A_t \cap A) = (1-b)\int_0^\infty \frac{e^{-t\lambda}}{\lambda}\pi\left(\mathrm{d}\lambda\right), \qquad t \geq 0,$$

so, combining it with equation (3.7), we have

$$\int_0^\infty \mathrm{Cor}(L\left(A_t\right), L(A_0))\mathrm{d}t = \frac{\int_0^\infty \lambda^{-2}\pi\left(\mathrm{d}\lambda\right)}{\int_0^\infty \lambda^{-1}\pi\left(\mathrm{d}\lambda\right)}.$$

Thus, the superposition trawl has long memory if and only if $\int_0^\infty \lambda^{-2}\pi\left(\mathrm{d}\lambda\right) = \infty$.

In the following we focus only on choices of specific π. These special cases have been analyzed in Barndorff-Nielsen et al. (2014), so here we only state them to establish notation for our applied work.

Example 4. *When π has a single atom of support at $\lambda > 0$, this is the exponential trawl*

$$d(s) = b + (1-b)\exp(\lambda s), \qquad s \leq 0, \tag{3.8}$$
$$leb\left(A\right) = \frac{1-b}{\lambda},\ leb\left(A_t \cap A\right) = \frac{1-b}{\lambda}e^{-\lambda t}.$$

Trivially it only allows short memory as $\int_0^\infty \bar{\lambda}^{-2}\pi\left(\mathrm{d}\bar{\lambda}\right) = \lambda^{-2} < \infty$ whenever $\lambda > 0$.

Example 5. *When*

$$\pi\left(\mathrm{d}\lambda\right) = \frac{\alpha^H}{\Gamma\left(H\right)}\lambda^{H-1}e^{-\lambda\alpha}\mathrm{d}\lambda, \qquad \alpha > 0,\ H > 1,$$

we produce the superposition gamma (sup-Γ) trawl

$$d(s) = b + (1-b)\left(1 - \frac{s}{\alpha}\right)^{-H}, \qquad s \leq 0, \tag{3.9}$$
$$leb\left(A\right) = (1-b)\frac{\alpha}{H-1},\ leb(A_t \cap A) = \frac{(1-b)\alpha}{H-1}\left(1 + \frac{t}{\alpha}\right)^{1-H}, \qquad t \geq 0.$$

It has long memory when $H \in (1, 2]$ and short memory when $H > 2$ as

$$\int_0^\infty \lambda^{-2}\pi\left(\mathrm{d}\lambda\right) = \frac{\Gamma\left(H-2\right)}{\Gamma\left(H\right)} < \infty \textit{ if and only if } H > 2.$$

Example 6. *When*

$$\pi\left(\mathrm{d}\lambda\right)=\frac{\left(\gamma/\delta\right)^{\nu}}{2K_{\nu}\left(\gamma\delta\right)}\lambda^{\nu-1}e^{-\left(\gamma^{2}\lambda+\delta^{2}\lambda^{-1}\right)/2}\mathrm{d}\lambda,\qquad\gamma,\delta>0,\ \nu\in\mathbb{R},$$

we produce the superposition generalized inverse Gaussian (sup-GIG) trawl

$$\begin{aligned}d\left(s\right)&=b+\left(1-b\right)\left(1-\frac{2s}{\gamma^{2}}\right)^{-\nu/2}\frac{K_{\nu}\left(\gamma\delta\sqrt{1-2s/\gamma^{2}}\right)}{K_{\nu}\left(\gamma\delta\right)},\qquad s\leq0\qquad(3.10)\\ \mathit{leb}\left(A\right)&=\left(1-b\right)\frac{\gamma}{\delta}\frac{K_{\nu-1}\left(\gamma\delta\right)}{K_{\nu}\left(\gamma\delta\right)},\\ \mathit{leb}(A_{t}\cap A)&=\left(1-b\right)\frac{\gamma}{\delta}\frac{\left(1+2t/\gamma^{2}\right)^{(1-\nu)/2}K_{\nu-1}\left(\gamma\delta\sqrt{1+2t/\gamma^{2}}\right)}{K_{\nu}\left(\gamma\delta\right)},\qquad t\geq0,\end{aligned}$$

where $K_{\nu}\left(x\right)$ is the modified Bessel function of the 2nd kind. It always has short memory as

$$\int_{0}^{\infty}\lambda^{-2}\pi\left(\mathrm{d}\lambda\right)=\frac{\left(\gamma/\delta\right)^{\nu}}{2K_{\nu}\left(\gamma\delta\right)}\frac{2K_{\nu-2}\left(\gamma\delta\right)}{\left(\gamma/\delta\right)^{\nu-2}}=\left(\frac{\gamma}{\delta}\right)^{2}\frac{K_{\nu-2}\left(\gamma\delta\right)}{K_{\nu}\left(\gamma\delta\right)}<\infty\ \textit{for all}\ \gamma,\delta>0,\ \nu\in\mathbb{R}.$$

However, it can also degenerate to the long memory sup-Γ *trawl by letting $\gamma=\sqrt{2\alpha}$, $\nu=H$ and $\delta\rightarrow0$. When $\gamma\rightarrow0$, $\pi\left(\mathrm{d}\lambda\right)$ becomes an inverse gamma distribution with scale parameter $\delta^{2}/2$ and shape parameter $-\nu$, so correspondingly we produce the superposition inverse gamma (sup-*Γ^{-1}*) trawl. This is an important case, for inverse gamma densities have polynomial decay in their tails so will generate short but substantial memory, which has the same pattern as the empirical data. We will see this clearly in Section 3.5.*

3.4 Moment-Based Inference

Here we discuss the inference technique based on matching moments using a path of prices P_t, $t\in[0,T]$. Due to (i) the stationarity of the price changes $P_\delta-P_0\overset{d}{\backsim}P_{t+\delta}-P_t$ for any t,δ and (ii) the high frequency nature of the data, moment-based estimates are plausible. The inference can basically split in two pieces: the inference of the Lévy measure ν and the inference on b and d.

3.4.1 Inference of Lévy measure

Due to the high frequency nature of the data, the instantaneous jumping distribution of the sample is close to the true value. Similarly, the sample power variation $\{P\}_t^{[r]}$ for any $r \geq 0$, when treated as a linear function of time t, has a slope that is also close to the truth. We can then use these facts to estimate the Lévy measure ν in terms of b.

Let us write the sample instantaneous jumping distribution as $\hat{\alpha}_y$, where $\sum_{y\in\mathbb{Z}\backslash\{0\}} \hat{\alpha}_y = 1$; also, estimate the slope of the r-th sample power variation against t by

$$\hat{\beta}_r = \frac{\{P\}_T^{[r]}}{T} = \frac{1}{T} \sum_{0<t\leq T} \left|\Delta P_t\right|^r .$$

Then by matching moments to equations (3.4) and (3.5), we should have

$$(2-b) \sum_{y\in\mathbb{Z}\backslash\{0\}} |y|^r \, \nu\left(y\right) = \hat{\beta}_r, \qquad r \geq 0, \tag{3.11}$$

$$\nu\left(y\right) + \nu\left(-y\right)\left(1-b\right) = \left(2-b\right)\hat{\alpha}_y \left\|\nu\right\|, \qquad y \in \mathbb{Z}\backslash\left\{0\right\}. \tag{3.12}$$

Using (3.11) with the case of $r = 0$, we have $\|\nu\| = \sum_{y\in\mathbb{Z}\backslash\{0\}} \nu\left(y\right) = \hat{\beta}_0 / \left(2-b\right)$ and hence

$$\begin{aligned} &\nu\left(y\right) + \nu\left(-y\right)\left(1-b\right) = \hat{\alpha}_y \hat{\beta}_0, \\ &\nu\left(-y\right) + \nu\left(y\right)\left(1-b\right) = \hat{\alpha}_{-y} \hat{\beta}_0, \qquad y \in \mathbb{N}. \end{aligned}$$

Solving these two equations gives us

$$\widehat{\nu\left(y\right)} = \frac{\hat{\alpha}_y - \left(1-b\right)\hat{\alpha}_{-y}}{\left(2-b\right) b} \hat{\beta}_0, \qquad y \in \mathbb{Z}\backslash\left\{0\right\}. \tag{3.13}$$

Remark 22. *This does not guarantee that $\widehat{\nu\left(y\right)} \geq 0$, so empirically we will truncate negative $\widehat{\nu\left(y\right)}$ by zero and at the same time tune the value of the corresponding $\widehat{\nu\left(-y\right)}$ such that*

$$\widehat{\nu\left(y\right)} + \widehat{\nu\left(-y\right)} = \frac{\hat{\alpha}_y - \left(1-b\right)\hat{\alpha}_{-y} + \hat{\alpha}_{-y} - \left(1-b\right)\hat{\alpha}_y}{\left(2-b\right) b} \hat{\beta}_0 = \frac{\hat{\alpha}_y + \hat{\alpha}_{-y}}{\left(2-b\right)} \hat{\beta}_0$$

remains unchanged. The advantage of this modification allows the conservation of all the (non-negative) moments of the estimated Lévy measure $\hat{\nu}$:

$$\sum_{y\in\mathbb{Z}\backslash\{0\}} |y|^r \widehat{\nu(y)} = \sum_{y=1}^{\infty} |y|^r \left(\widehat{\nu(y)} + \widehat{\nu(-y)}\right).$$

However, it comes with the price that the estimates for all of the odd *cumulants of $P_t - P_0$ are altered, but practically this will be neglectable as the truncation is only needed for larger y and the corresponding intensity $\nu(y)$ is usually quite small.*

To completely avoid the negative estimates, one might parameterize the Lévy measure as in Barndorff-Nielsen et al. (2014), but here we prefer to stay with the non-parametric estimates.

Remark 23. *We should note that (3.13) has included all the information we can access from equations (3.11) and (3.12), so we cannot rely on equations (3.11) and (3.12) to solve b and the Lévy measure ν at the same time. The details can be found in the Appendix A.3.*

3.4.2 Inference of permanence and trawl function

We will need to employ additional moment equations to estimate the trawl function d as well as b. The easiest way to do this is through Theorem 14. In particular, we will use the sample variance of $\left\{P_{k\delta} - P_{(k-1)\delta}\right\}_{k=1}^{T/\delta}$ to estimate

$$\mathrm{Var}\left(P_\delta - P_0\right) = \left(b\delta + 2leb\left(A_\delta\backslash A_0\right)\right)\kappa_2\left(L_1\right) = \left(b\delta + 2leb\left(A_\delta\backslash A_0\right)\right)\sum_{y\in\mathbb{Z}\backslash\{0\}} y^2\nu(y).$$

Denote the sample variance with the sampling interval δ as $\widehat{\sigma_\delta^2}$. Then by (3.13) and matching moments, we should have

$$\widehat{\sigma_\delta^2} = \left(\frac{b\delta + 2leb\left(A_\delta\backslash A_0\right)}{2-b}\right)\sum_{y\in\mathbb{Z}\backslash\{0\}} y^2\hat{\alpha}_y\hat{\beta}_0. \tag{3.14}$$

Appendix B.3 shows how to non-parametrically estimate the trawl function d using $\widehat{\sigma_\delta^2}$, but here we only demonstrate the inference for a parameterized trawl.

Suppose for now that the trawl function d is parameterized by ϕ, for example, $\phi = \lambda$ in the exponential trawl (3.8), $\phi = (\alpha, H)^T$ in the sup-Γ trawl (3.9) and $\phi = (\gamma, \delta, \nu)^T$ in the sup-GIG trawl (3.10). A simple way to estimate b and ϕ simultaneously is through a non-linear least square fitting to equation (3.14) divided by δ across different δ. The reason to work on $\widehat{\sigma_\delta^2}/\delta$ instead of $\widehat{\sigma_\delta^2}$ is to amplify the effect of empirical market microstructure for small δ, so the non-linear least square estimation of b and ϕ will not be overly dominated by the linear part of the variogram.

Remark 24. *By definition of the sample variance and the realized variance, as $T \to \infty$,*

$$\widehat{\sigma_\delta^2} \approx \frac{1}{T/\delta} \sum_{k=1}^{T/\delta} \left(P_{k\delta} - P_{(k-1)\delta}\right)^2 - \left(\frac{P_T - P_0}{T/\delta}\right)^2,$$

$$\frac{\widehat{\sigma_\delta^2}}{\delta} \approx \frac{1}{T} RV^{(T/\delta)} - \delta \frac{(P_T - P_0)^2}{T^2} \approx \frac{1}{T} RV^{(T/\delta)},$$

where we throw out the second-order term in the final approximation. Thus, essentially what we try to fit is the variance signature plot ($RV^{(T/\delta)}$ against δ). From now on, we also call the plot $\widehat{\sigma_\delta^2}/\delta$ against δ a variance signature plot.

Example 7. *To check the effectiveness of this moment estimator, we conduct a Monte Carlo simulation study on the price process model parameterized by the exponential trawl (3.8). Throughout the rest of this paper, all the numerical values are reported under the time unit being a second. Then we set $\lambda_{\text{true}} = 0.681$ and a non-symmetric Skellam basis with Lévy measure*

$$\nu\left(\mathrm{d}y\right) = \nu^+ \delta_{\{1\}}\left(\mathrm{d}y\right) + \nu^- \delta_{\{-1\}}\left(\mathrm{d}y\right),$$

where $\nu_{\text{true}}^+ = 0.0138$, $\nu_{\text{true}}^- = 0.0131$ and $b_{\text{true}} = 0.396$. All the $10,000$ Monte Carlo simulated paths are drawn with $V_0 = 7,486$ (ticks) during the time interval 72.03 to $75,600$ (seconds), where

75, 600 *means the closing time of the market, 21:00. All the settings here are taken from the empirical TNC1006 data set on March 22, 2010, which we will study in next Section.*

The non-linear least square fitting for (3.14) is conducted for δ's ranging from 0.1 *seconds to* 60 *seconds with* 60 *equally spaced grid points on its log-scale. We then repeat the moment-based estimates for $\theta = (b, \nu^+, \nu^-, \lambda)^T$ and derive histograms of these estimates in Figure 3.3. The estimates from the proposed methodology (using equations (3.13) and (3.14)) correctly center around the true values; also notice that this method is particularly accurate for estimating ν^+ and ν^-.*

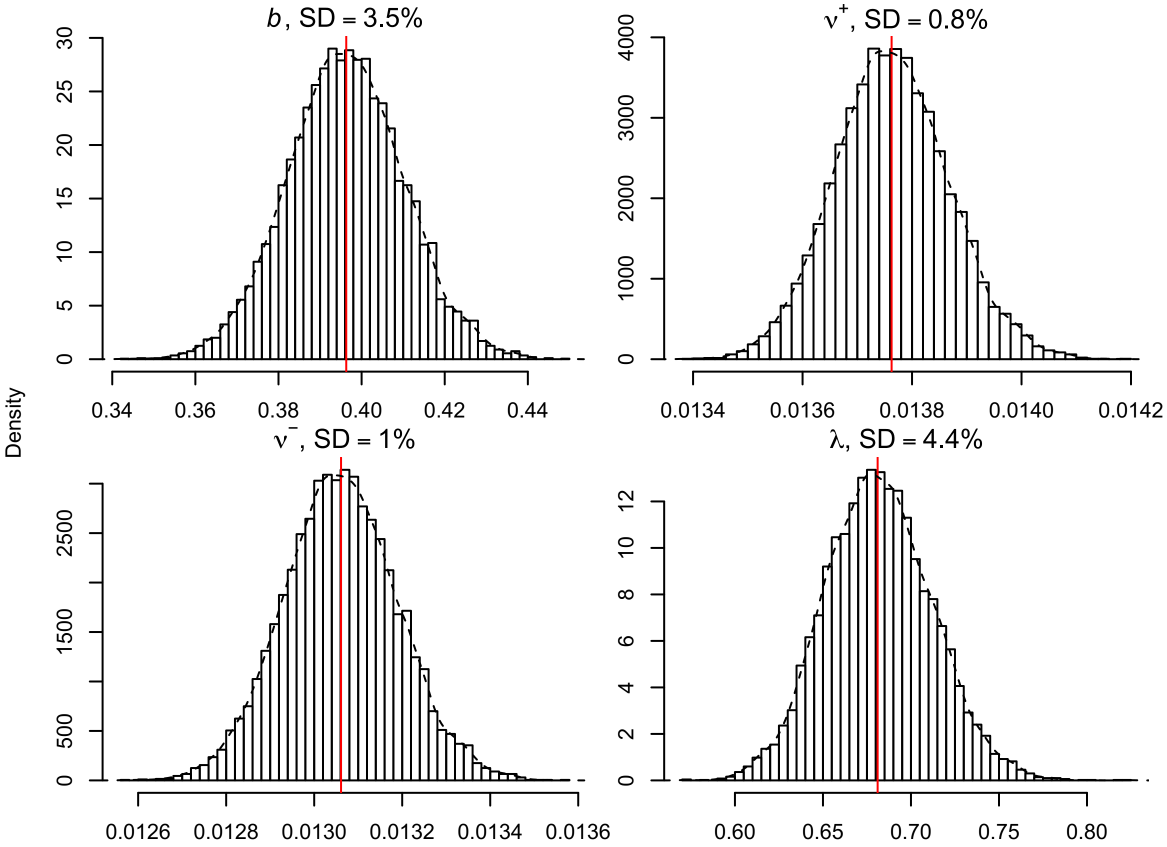


Figure 3.3: $10,000$ Monte Carlo simulation of moment estimations for the price process with exponential trawl $d(s) = b + (1-b)\exp(\lambda s)$ and the Skellam basis $\nu(\mathrm{d}y) = \nu^+\delta_{\{1\}}(\mathrm{d}y) + \nu^-\delta_{\{-1\}}(\mathrm{d}y)$. The vertical lines in each of the histograms mean the true value. The Monte Carlo standard deviations are reported on the scale of the true values.

Remark 25. *Except the moment-based estimations, we can also conduct maximum likelihood estimation for our proposed model, which requires sophisticated techniques to filter out the Lévy* *process $L(B_t)$. We are currently exploring particle methods toward this direction.*

3.5 Empirical Analysis for Futures Data

In this Section, we employ these moment-based estimators for empirical analysis. Covering two days of trading activities on two different assets, four data sets are studied here: (i) the Ten-Year US Treasury Note futures contract delivered in June 2010 (TNC1006) during March 22, 2010; (ii) the International Monetary Market (IMM) Euro-Dollar Foreign Exchange (FX) futures contract delivered in June 2010 (EUC1006) during March 22, 2010; (iii) TNC1006 during May 7, 2010; (iv) EUC1006 during May 7, 2010. These data sets come from the same database that is used by Barndorff-Nielsen et al. (2012). The first trading day is randomly chosen, while the second trading day is not only the release of US non-farm payroll numbers but was also experiencing the European sovereign debt crisis. These data sets are derived from data feeds at the Chicago Mercantile Exchange (CME). They have been preprocessed using the procedures described in Appendix B.2. From now on, we will no longer mention the delivery date of each data set and the year 2010.

3.5.1 Data features

All of these four data sets use all the trades from 00:00 to 21:00, shown in Figure 3.4. With such large time scales, each of the trace plots look like a continuous time diffusion process. However, if we focus these data sets to much smaller time scales (within one hour for TNC and within two minutes for EUC), shown in Figure 3.5, the discreteness becomes important. In particular, we can see several multiple-tick jumps in the two EUC data sets shown in Figure 3.5.

Table 3.1 summarizes some basic features of these four data sets. Both contracts have more activities during May 7 than during March 22 and the standard deviations of the jump size for

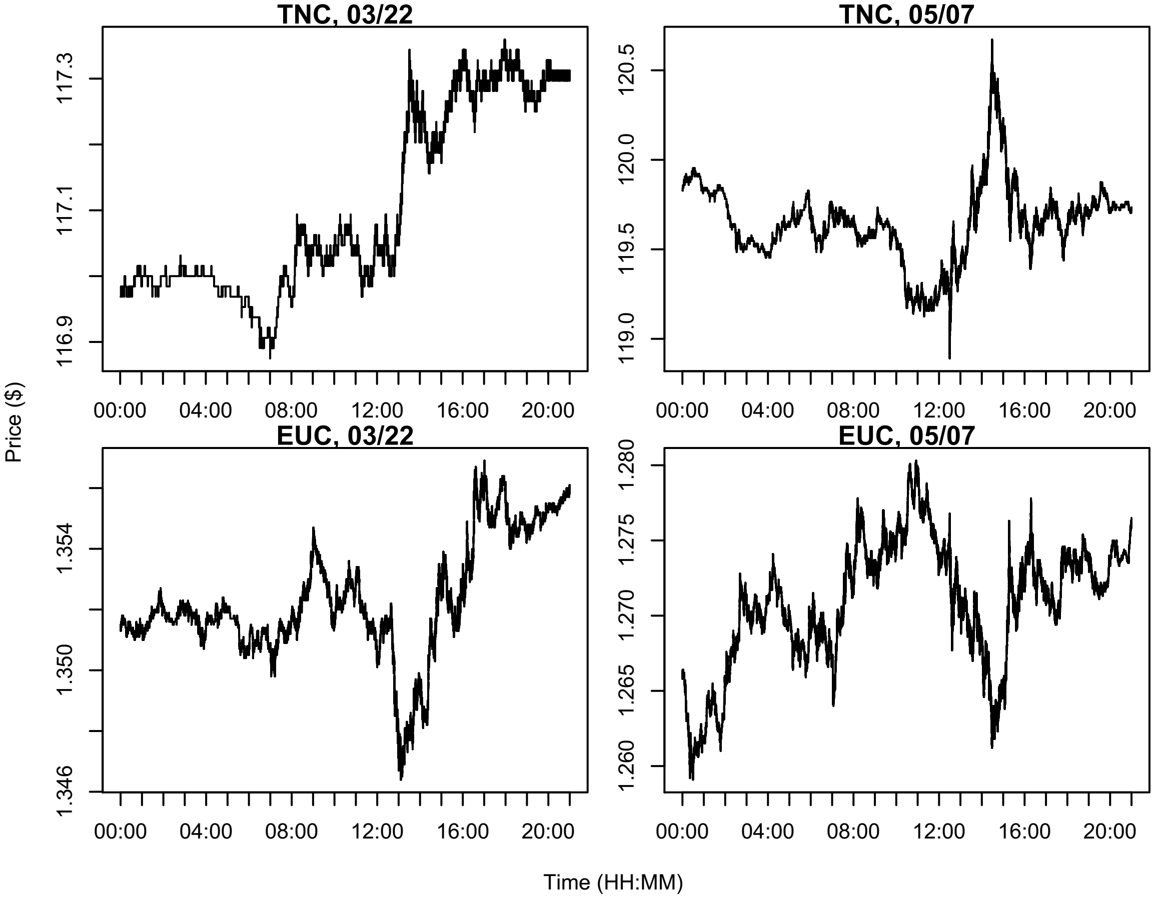


Figure 3.4: The complete trace plots for the four data sets during 00:00 to 21:00. The x-axis is the calendar time (HH:MM), while the y-axis is the price ($).

all the four data sets are close to 1 even though the range of all possible jump sizes might differ a lot.

We also plot the empirical instantaneous jumping distribution (on the log-scale) for the four data sets in Figure 3.6. Those estimated probabilities will be used as $\hat{\alpha}_y$ for the moment estimate defined in the previous Section. Generally, the jumps of EUC have more variability than the TNC. Furthermore, we can see that even for the same contract, say TNC, the jumping characteristic is completely different from a random chosen day (March 22) to a day with a major economic event (May 7). In a normal day like March 22, the TNC trading has depths so large that it always jumps by one tick, but the situation changes enormously for a highly active day like May 7, by this time the TNC trading behaves just like other multiple-tick markets.

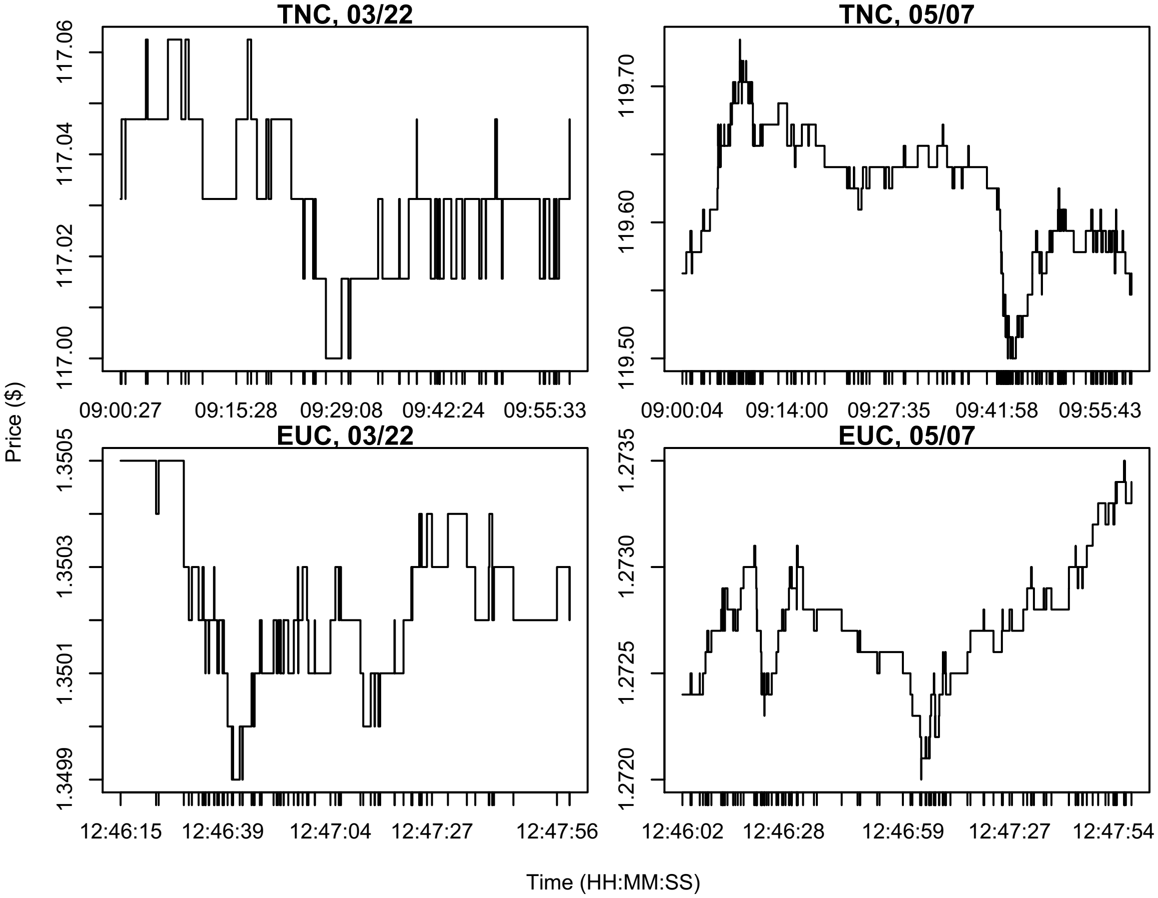


Figure 3.5: The trace plots for two TNC data sets during 09:00 to 10:00 and for two EUC data sets during 12:46 to 12:48. The x-axis is the calendar time (HH:MM:SS), while the y-axis is the price ($).

Remark 26. *One more implication from Figure 3.6 is that $\kappa_1(L_1)$ is, of course, a small number. To see this, we note that*

$$\widehat{\kappa_1(L_1)} = \sum_{y\in\mathbb{Z}\backslash\{0\}} y\widehat{\nu(y)} = \frac{\sum_{y\in\mathbb{Z}\backslash\{0\}} y\hat{\alpha}_y - (1-b)\sum_{y\in\mathbb{Z}\backslash\{0\}} y\hat{\alpha}_{-y}}{(2-b)\,b}\hat{\beta}_0 = \frac{\sum_{y\in\mathbb{Z}\backslash\{0\}} y\hat{\alpha}_y}{b}\hat{\beta}_0.$$

Hence, the more symmetric the Figure 3.6, the smaller the estimate of $\kappa_1(L_1)$.

Finally, we show the correlograms of the four data sets in Figure 3.7, using three orders of magnitude of sampling intervals δ: 0.1 second, 1 second, 10 seconds and 1 minute. For each data set, we will use a single set of parameters in our price model to fit all of the correlograms with different δ. In general, these autocorrelations are significantly negative and increasing as k increases, while if δ gets very large the autocorrelations will fall to roughly zero. Of course there is

Contract, Day	Tick Size ($)	Num. of Price Changes	Size of Price Changes (Tick)			
			Avg.	SD.	Min.	Max.
TNC, 03/22	1/64	$3,249$	0.00646	1.000	-1	1
EUC, 03/22	0.0001	$13,943$	0.00337	1.012	-2	3
TNC, 05/07	1/64	$12,849$	-0.00047	1.035	-13	15
EUC, 05/07	0.0001	$55,379$	0.00190	1.077	-13	15

Table 3.1: Summary statistics of the four futures data sets.

strong evidence that the empirical data cannot be well-described by a pure Lévy process, which always gives zero autocorrelations for returns. Our model is capable of describing these autocorrelation features (Theorem 16 and Corollary 17). The next Subsection conducts moment-based estimations for these empirical data sets.

3.5.2 Parameter estimation

We use the methodology described before on the four data sets with the three different trawls (3.8), (3.9) and (3.10). The estimation[‡] results are shown in Table 3.2 on page 84, where

$$\nu^+ = \sum_{y=1}^{\infty} \nu(y) \text{ and } \nu^- = \sum_{y=1}^{\infty} \nu(-y)$$

are the positive and negative jump intensities respectively. We observe in the Table that the estimation of ν^+ and ν^- are relatively robust across different choices of trawls. The estimate of H in Table 3.2 clearly suggests the insufficiency of using a sup-Γ trawl for the empirical data. Furthermore, even though we fit a more general sup-GIG trawl with three parameters, the four empirical data sets can almost be described by the sup-Γ^{-1} trawl with only two parameters (the case of $\gamma \to 0$ for sup-GIG trawl mentioned in Section 3.3.6). This phenomenon might be attributed to the fact that inverse gamma distributions decay exponentially near the origin but polynomially near infinity, allowing it to capture these very different time scales.

[‡] To especially emphasize the fitting of market microstructure effects, the sample variance is calculated on an equally distant grid on the log-scale of δ whose range is shown in Figure 3.9.

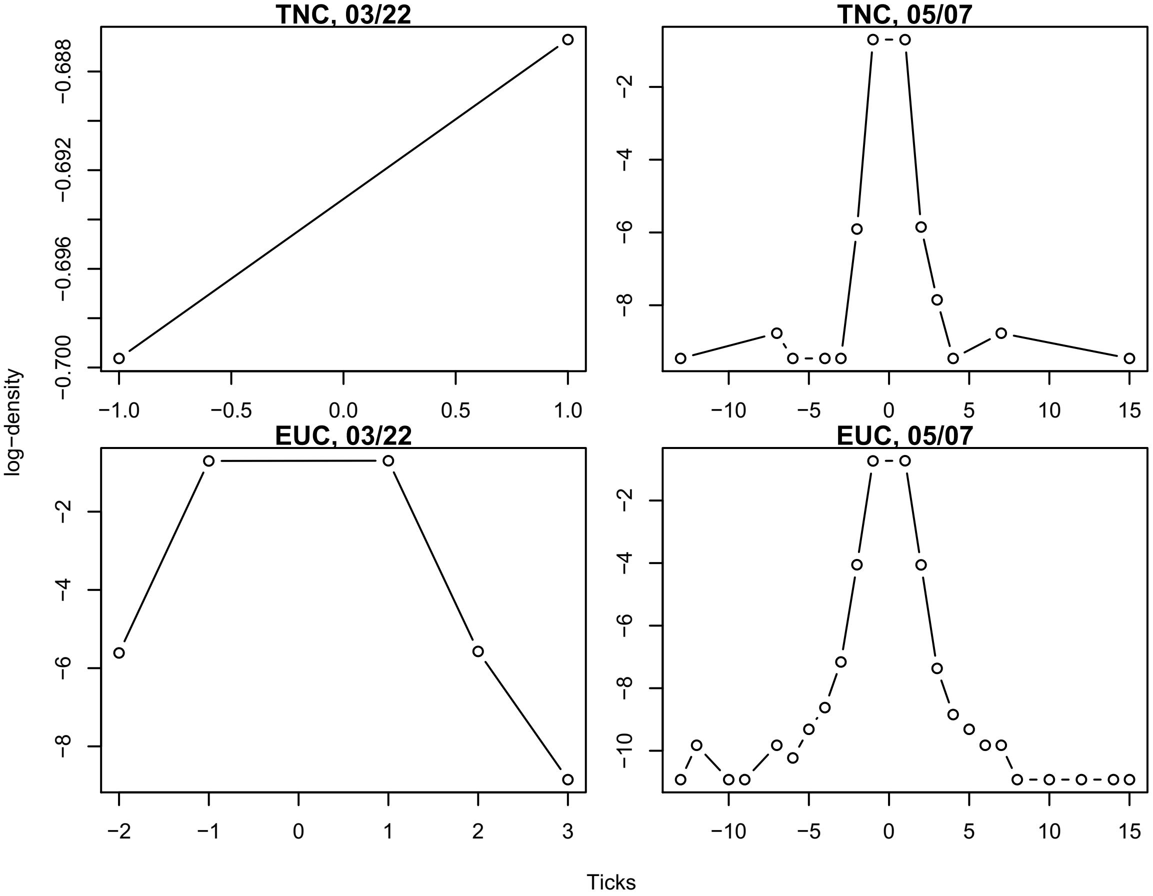


Figure 3.6: The log-histograms for the empirical instantaneous jumping distributions of the four data sets. The x-axis for each plot is the size of the jump, while the y-axis denotes the estimated probability value in a log-scale.

Remark 27. *In the same Table, we also provide the standard error (SE) estimates for these moment-based estimations using the model-based bootstrap, i.e., a vanilla Monte Carlo simulation with plugged-in parameters.*

Using these estimated parameters, we first show the variance signature plots of $\widehat{\sigma_\delta^2}/\delta$ against δ along with the corresponding theoretical curves (3.14) for each trawl in Figure 3.8 and 3.9, where the second of these graphs uses a log-scale for δ. In each of the plots, we put not only $\lim_{\delta\to 0} \widehat{\sigma_\delta^2}/\delta = \left(\partial_\delta \widehat{\sigma_\delta^2}\right)(0) = \sum_{y\in\mathbb{Z}\setminus\{0\}} y^2 \hat{\alpha}_y \hat{\beta}_0$ at the corresponding location of $\delta = 0$ but also a reference horizontal line from a pure Lévy process model ($b = 1$), which is calculated from the slope of a linear fitting line in the variogram of $\widehat{\sigma_\delta^2}$ against δ.

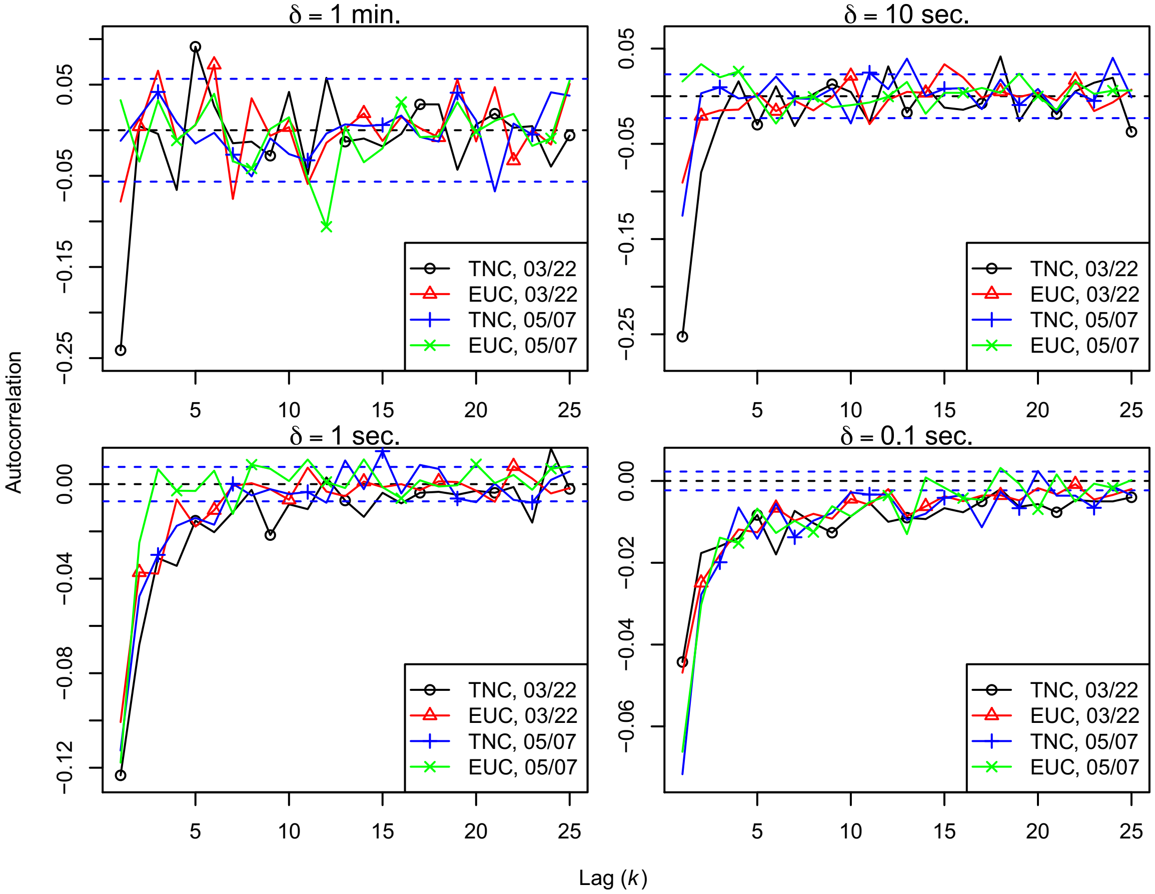


Figure 3.7: The correlograms with different sampling intervals $\delta = 0.1, 1, 10, 60$ (seconds) for the four data sets. The x-axis for each plot is the lag k, while the y-axis denotes the value of empirical autocorrelation. The dashed lines are located at $\pm 2/\sqrt{T/\delta}$.

These fittings to the variance signature plots show good results—here we particularly notice that using a sup-GIG trawl gives a very good fit; while the other two simpler trawls fail to fit the region with a smaller δ. This point becomes apparent when we check Figure 3.9.

To further examine our model fitting, we also show the log-histograms for the return distribution with different δ along with the theoretical curves (by applying the inverse Fourier transform on Theorem 14) in Figure 3.10. For a larger δ the sup-GIG trawl do a better job than the other two trawls (not shown in Figure 3.10) while for a smaller δ the difference among the three trawls is limited. As an overall comment, our model seems to underestimate the tail part of each of the empirical jumping distributions.

Trawl	Para	TNC, 03/22		EUC, 03/22		TNC, 05/07		EUC, 05/07	
		Est.	SE	Est.	SE	Est.	SE	Est.	SE
Exp	b	0.396	0.014	0.654	0.008	0.574	0.015	0.694	0.007
	ν^+	0.014	0.000	0.069	0.000	0.059	0.001	0.282	0.001
	ν^-	0.013	0.000	0.068	0.000	0.060	0.001	0.279	0.001
	λ	0.681	0.030	2.470	0.083	3.888	0.218	4.033	0.133
sup-Γ	b	0.283	0.021	0.604	0.012	0.525	0.016	0.649	0.010
	ν^+	0.013	0.000	0.067	0.001	0.057	0.001	0.272	0.002
	ν^-	0.012	0.000	0.066	0.001	0.058	0.001	0.270	0.002
	α	1.146	0.191	0.311	0.037	0.187	0.038	0.192	0.023
	H	1.000	0.125	1.000	0.104	1.000	0.139	1.000	0.102
sup-GIG	b	0.186	0.028	0.528	0.034	0.440	0.029	0.648	0.011
	ν^+	0.013	0.000	0.063	0.001	0.054	0.001	0.272	0.002
	ν^-	0.011	0.000	0.062	0.001	0.055	0.001	0.269	0.002
	γ	0.000	0.066	0.000	0.030	0.003	0.028	0.000	0.064
	δ	0.453	0.049	0.604	0.085	0.583	0.099	1.525	0.209
	ν	-0.604	0.078	-0.453	0.067	−0.332	0.077	−0.741	0.170

Table 3.2: Moment-based estimations under different trawls for the four data sets. Also shown are the standard error (SE) estimates for the moment estimator to each parameter using the model-based bootstrap, where the number of bootstrapped paths we draw is 10,000.

We now demonstrate the correlograms for the returns with different δ along with the theoretical curves in Figure 3.11. For a larger δ, the empirical returns look almost uncorrelated (insignificant from being 0) except for TNC on March 22, but the sup-GIG trawl still captures this anomaly at the first lag. As δ becomes smaller, those negative correlations become more significant; even though the exponential trawl and the sup-Γ trawl (not shown in Figure 3.11) can depict the shape of the autocorrelation, only sup-GIG trawl can fit the first few lags.

As a summary, the sup-GIG trawl (or essentially the sup-Γ^{-1} trawl) performs better than the other two trawls in every aspects. These empirical analyses demonstrate the descriptive power of our proposed model for the futures data.

Remark 28. *We now criticize the insufficient part of our proposed model. A plot (not shown) of the counting process of price moves for our four data sets will clearly show a non-linear increasing pattern that disobeys the linearity described by equation (3.5). This non-linear pattern can be*

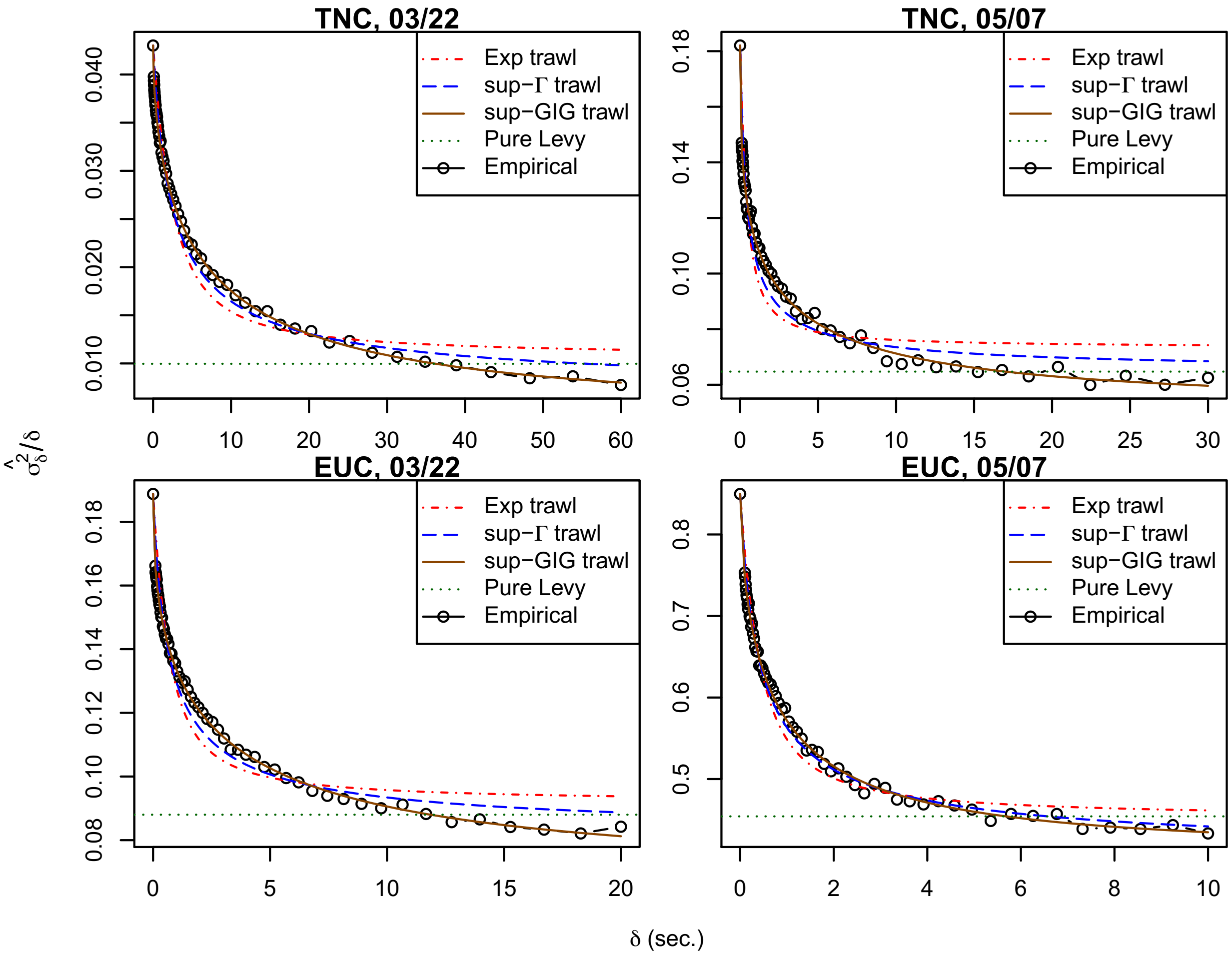


Figure 3.8: The variance signature plots for the four data sets along with the fitting curves from different trawls. The x-axis for each plot is δ (seconds), while the y-axis denotes the value of the sample variance of returns divided by δ.

attributed to the well-known diurnal time-varying levels of trading activity. For the same contract, its two counting process plots look alike (after rescaling) across different trading days.

We are currently exploring methods that can adjust the model to deal with these effects, hoping to report on them shortly. It will involve the use of two independent stochastic time changes for the positive events and the negative events. A special case on the Skellam Lévy process using this ideas has been addressed in Kerss et al. (2014).

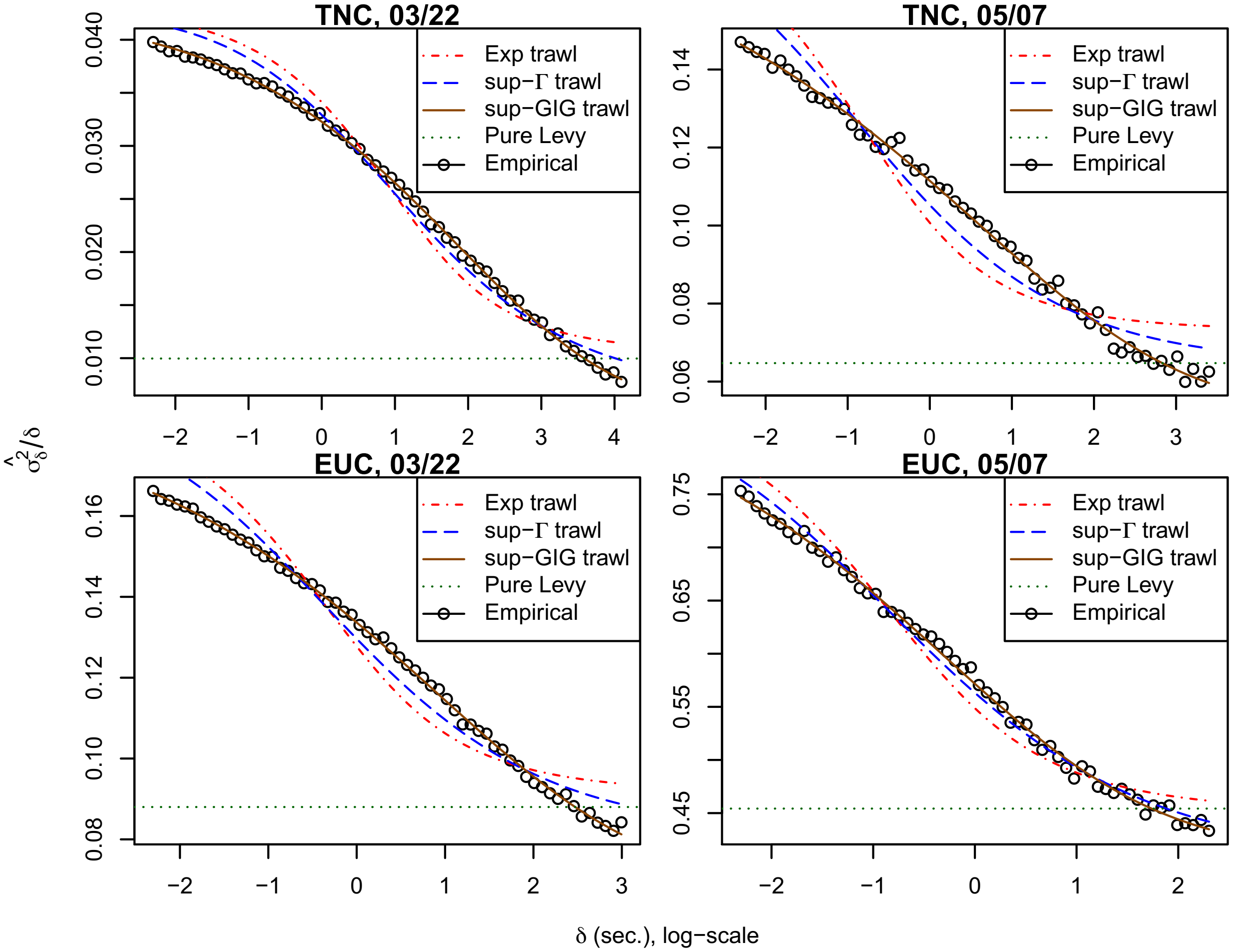
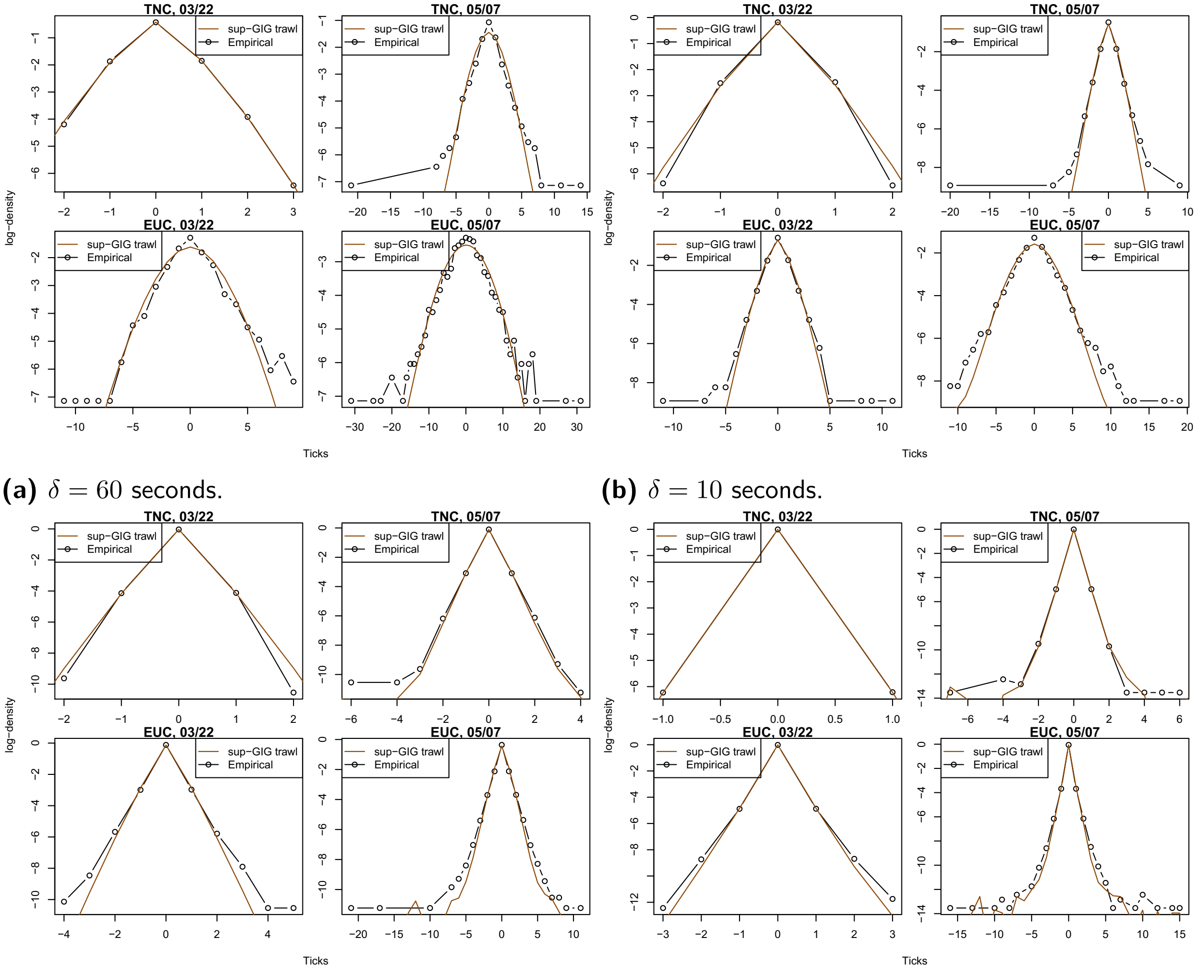


Figure 3.9: The variance signature plots for the four data sets along with the fitting curves from different trawls in the scale of $\log \delta$.

3.6 Conclusion

We propose a novel and simple model that can adequately capture some of the important features of high frequency financial data. It is able to deal with the dependence in price changes measured over three different orders of magnitude of time intervals. The model is directly formulated in terms of the price impact curve (or trawl function). It has a càdlàg price process that is a piecewise constant semimartingale with finite activity, finite variation and no Brownian motion component.

However, we need to emphasize that, the proposed model in this paper is just an initial step.



(a) $\delta = 60$ seconds.

(b) $\delta = 10$ seconds.

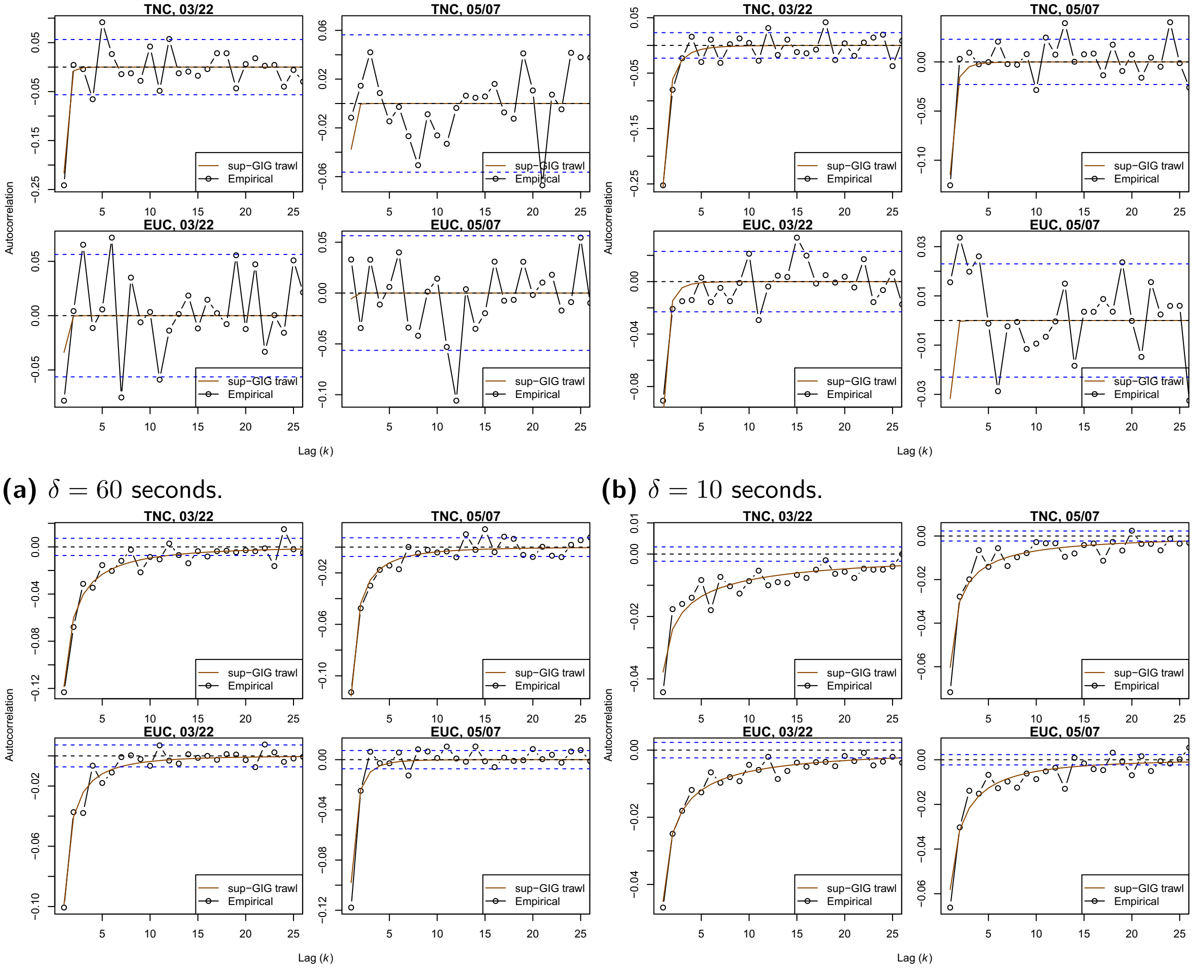
(c) $\delta = 1$ second.

(d) $\delta = 0.1$ seconds.

Figure 3.10: The log-histograms for the returns of the four data sets over several sampling intervals along with the theoretical curves from sup-GIG trawl.

Even though we emphasize the discreteness and the fleetingness in the movements of the price process, we have been assuming a simple structure so far with no time-varying features. We will shortly report on how to generalize this model to the more realistic case using a stochastic time-change.

Our model provides a good description to the empirical data, while we majorly focus on the trade prices, which is not always immediately tradable. For market practitioners who sit either on the buy side or the sell side, they might consider to apply the proposed model on either the ask price or bid price, so our model is much more widely applicable than the cases we report



(a) $\delta = 60$ seconds.

(b) $\delta = 10$ seconds.

(c) $\delta = 1$ second.

(d) $\delta = 0.1$ seconds.

Figure 3.11: The correlograms for the returns of the four data sets over several sampling intervals along with the theoretical curves from sup-GIG trawl. The dashed lines are located at $\pm 2/\sqrt{T/\delta}$.

here.

A

Proofs and Derivations

A.1 Chapter 1

Proof of Theorem 3. To start with the proof, we first provide a Lemma which plays an essential rule in the proofs of both the Glivenko-Cantelli Theorem and our Theorem 3.

Lemma 20. *Suppose F_n and F are (nonrandom) distribution functions on $\mathbb{R}$ such that*

$$F_n\left(x\right) \to F\left(x\right) \text{ for all } x \in \mathbb{R}.$$

If F is continuous, then we have

$$F_n\left(x-\right) \to F\left(x\right) \text{ for all } x \in \mathbb{R}, \tag{A.1}$$

and, moreover,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0. \tag{A.2}$$

Proof. We first show equation (A.1). Suppose x is a continuity point of F. Since $F_n(x-) \leq F_n(x)$,

$$\limsup_{n\to\infty} F_n(x-) \leq \limsup_{n\to\infty} F_n(x) = F(x).$$

For any $y < x$, we have

$$F_n(y) \leq F_n(x-),$$

which implies that

$$F(y) = \liminf_{n\to\infty} F_n(y) \leq \liminf_{n\to\infty} F_n(x-).$$

Since

$$\lim_{y\to x-} F(y) = F(x),$$

the desired result follows.

We now show equation (A.2). Let $\epsilon > 0$ be given and consider a partition of the real line into finitely many pieces of the form $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that, for $0 \leq j \leq k-1$,

$$F(t_{j+1}) - F(t_j) \leq \frac{\epsilon}{2}.$$

For any $x \in \mathbb{R}$, there exists j such that $t_j \leq x < t_{j+1}$. For such j,

$$F_n(t_j) \leq F_n(x) \leq F_n(t_{j+1}-),\ F(t_j) \leq F(x) \leq F(t_{j+1}),$$

which implies that

$$F_n(t_j) - F(t_{j+1}) \leq F_n(x) - F(x) \leq F_n(t_{j+1}-) - F(t_j).$$

Furthermore,

$$F_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}) \leq F_n(x) - F(x),$$

$$F_n(t_{j+1}-) - F(t_{j+1}) + F(t_{j+1}) - F(t_j) \geq F_n(x) - F(x).$$

By the construction of the partition, we have that

$$F_n(t_j) - F(t_j) - \frac{\epsilon}{2} \leq F_n(x) - F(x),$$
$$F_n(t_{j+1}-) - F(t_{j+1}) + \frac{\epsilon}{2} \geq F_n(x) - F(x).$$

For each j, let $N_j = N_j(\epsilon)$ be such that, for $n > N_j$,

$$F_n(t_j) - F(t_j) > -\frac{\epsilon}{2}.$$

Also, by equation (A.1), let $M_j = M_j(\epsilon)$ be such that, for $n > M_j$,

$$F_n(t_j-) - F(t_j) < \frac{\epsilon}{2}.$$

Let $N = \max_{1 \leq j \leq k} N_j \vee M_j$. For $n > N$ and any $x \in \mathbb{R}$, we then have that

$$|F_n(x) - F(x)| < \epsilon.$$

The desired result follows. □

Let W be the canonical graphon of a degree-identifiable ExGM. The marginal integral $g(u) \triangleq \int_0^1 W(u,v)\,dv$ must be strictly increasing, of which the range is not necessarily the whole $[0,1]$ interval. However, we will still use the notation g^{-1} to denote the corresponding CDF of the degree proportion random variable $g(U)$. As mentioned in Remark 2, g^{-1} will be a continuous function on $[0,1]$.

First we show that $\left|\hat{U}_i - U_i\right| \to 0$ in probability. Recall that

$$\hat{U}_i = \hat{F}\left(D_i\right) \text{ and } U_i = g^{-1}\left(g\left(U_i\right)\right),$$

so

$$\left|\hat{U}_i - U_i\right| \leq \left|\hat{F}\left(D_i\right) - g^{-1}\left(D_i\right)\right| + \left|g^{-1}\left(D_i\right) - g^{-1}\left(g\left(U_i\right)\right)\right|.$$

In the proof of Theorem 5 in Bickel et al. (2011), they prove that actually

$$\mathbb{E}\left(D_i - g\left(U_i\right)\right)^2 \to 0,$$

so, in particular, $D_i \to g\left(U_i\right)$ in probability and hence Continuous Mapping Theorem suggests

$$g^{-1}\left(D_i\right) \to g^{-1}\left(g\left(U_i\right)\right) \text{ in probability.}$$

Furthermore, Theorem 5 in Bickel et al. (2011) also suggest that $M_2\left(\hat{F}, g^{-1}\right) \to 0$ in probability. Here M_2 means the Mallows 2-distance between two distributions F_1 and F_2, which is defined by

$$M_2\left(F_1, F_2\right) \triangleq \min_F \left\{\sqrt{\mathbb{E}\left(X - Y\right)^2} \left|\left(X, Y\right) \backsim F, X \backsim F_1, Y \backsim F_2\right.\right\}.$$

Especially, $M_2\left(F_n, F\right) \to 0$ if and only only if $F_n \Rightarrow F$ in distribution, i.e.,

$$F_n\left(x\right) \to F\left(x\right) \text{ for all } x \text{ being the continuous point of } F,$$

and

$$\int x^2 dF_n\left(x\right) \to \int x^2 dF\left(x\right).$$

Hence, for every subsequence $N_1, ..., N_m$, there exists a further subsequence $N_{m_1}, ..., N_{m_k}$ such that $M_2\left(\hat{F}_{N_{m_k}}, g^{-1}\right) \to 0$ a.s. In particular, this means, for a.s. ω,

$$\hat{F}_{N_{m_k}}\left(x\right) \to g^{-1}\left(x\right) \qquad \forall x \in \left[0, 1\right].$$

Because g^{-1} is continuous, Lemma 20 easily extend this result to

$$\sup_{x\in[0,1]}\left|\hat{F}_{N_{m_k}}(x)-g^{-1}(x)\right|\to 0 \text{ for a.s. } \omega.$$

Thus we actually have

$$\sup_{x\in[0,1]}\left|\hat{F}(x)-g^{-1}(x)\right|\to 0 \text{ in probability,} \tag{A.3}$$

so this then implies that

$$\left|\hat{F}(D_i)-g^{-1}(D_i)\right|\to 0 \text{ in probability.}$$

The desired result follows.

Next we show that $\left|\hat{U}_i-\tilde{U}_i\right|\to 0$ in probability. We simply observe that

$$\begin{aligned}\left|\tilde{U}_i-\hat{U}_i\right| &\le \left|\hat{F}(D_i)-\hat{F}(D_i-)\right| \le \left|\hat{F}(D_i)-g^{-1}(D_i)\right|+\left|g^{-1}(D_i)-\hat{F}(D_i-)\right| \\ &\le \sup_{x\in[0,1]}\left|\hat{F}(x)-g^{-1}(x)\right|+\lim_{x\to D_i-}\left|g^{-1}(x)-\hat{F}(x)\right| \le 2\sup_{x\in[0,1]}\left|\hat{F}(x)-g^{-1}(x)\right|,\end{aligned}$$

so equation (A.3) guarantees that $\left|\hat{U}_i-\tilde{U}_i\right|\to 0$ in probability. This then finishes the proof of Theorem 3. □

Proof of Theorem 1. The USVT-A estimate for the canonical graphon is explicitly written as

$$\hat{W}(u,v)\triangleq\sum_{i,j=1}^{N}\hat{P}_{ij}1_{S_{ij}}(u,v),$$

where $S_{ij}\triangleq\left(\tilde{U}_i-1/N,\tilde{U}_i\right]\times\left(\tilde{U}_j-1/N,\tilde{U}_j\right]$ is the (i,j)-th square with spacing $1/N$ and $\hat{P}_{ij}$ is the USVT probability matrix estimation defined in Subsection 1.3.1.

To begin with, we define

$$\left\|\hat{W}-W\right\|^2\triangleq\mathbb{E}\left(\int_0^1\int_0^1\left(\hat{W}(u,v)-W(u,v)\right)^2dudv\right),$$

which can be decomposed into four pieces by inserting the following three objects

$$\begin{aligned} W_1(u,v) &\triangleq \sum_{i,j=1}^{N} W\left(\tilde{U}_i, \tilde{U}_j\right) 1_{S_{ij}}(u,v), \\ W_2(u,v) &\triangleq \sum_{i,j=1}^{N} W(U_i, U_j) 1_{S_{ij}}(u,v), \\ W_3(u,v) &\triangleq \sum_{i,j=1}^{N} P_{ij} 1_{S_{ij}}(u,v) \end{aligned}$$

then

$$\left\|\hat{W} - W\right\| \leq \|W - W_1\| + \|W_1 - W_2\| + \|W_2 - W_3\| + \left\|W_3 - \hat{W}\right\|.$$

Now we will look at these terms individually.

For the first term, since W_1 is just a stepwise W evaluated at the upright corner for each small square S_{ij}, $\|W - W_1\| \to 0$ as W is continuous on $[0,1]^2$ and hence uniformly continuous.*

For the third term, it's easy to see that, by the definition of P_{ij},

$$\begin{aligned} \|W_2 - W_3\|^2 &= \mathbb{E}\left(\frac{1}{N^2} \sum_{1 \leq i,j \leq N} \left(W(U_i, U_j) - P_{ij}\right)^2\right) \\ &= \mathbb{E}\left(\frac{1}{N^2} \sum_{i=1}^{N} W(U_i, U_i)^2\right) = \frac{1}{N} \mathbb{E}\left(W(U_i, U_i)^2\right) \to 0. \end{aligned}$$

For the final term, we refer to Theorem 2, where by definition the left hand side of equation (1.2) is the same as $\left\|W_3 - \hat{W}\right\|^2$.

Now we pay attention to $\|W_1 - W_2\|$, which depends only on the empirical degree sorting. Denote the event $\left|U_i - \tilde{U}_i\right| > \delta_\epsilon$ by E_i, where δ_ϵ is chosen by the uniform continuity of W such that

$$|u_1 - u_2| \leq \delta_\epsilon \text{ and } |v_1 - v_2| \leq \delta_\epsilon \Rightarrow |W(u_1, v_1) - W(u_2, v_2)| < \epsilon.$$

*Actually, what we need here is just a uniform continuity modulus δ such that $|W(u,v) - W(u',v')| < \epsilon$ whenever $|u - u'| < \delta$ and $|v - v'| < \delta$ in $[0,1]^2$. Thus, the continuity condition of W can be actually weaken by piecewise continuity on $[0,1]^2$ with only finitely many number of pieces.

Then we will have

$$
\begin{aligned}
&\mathbb{E}\left(\left|W\left(\tilde{U}_i,\tilde{U}_j\right)-W\left(U_i,U_j\right)\right|^2\right)\\
&=\mathbb{E}\left(\left|W\left(\tilde{U}_i,\tilde{U}_j\right)-W\left(U_i,U_j\right)\right|^2;E_i\cup E_j\right)+\mathbb{E}\left(\left|W\left(\tilde{U}_i,\tilde{U}_j\right)-W\left(U_i,U_j\right)\right|^2;E_i^c\cap E_j^c\right)\\
&\le 4\mathbb{P}\left(E_i\cup E_j\right)+\epsilon^2\le 4\left(\mathbb{P}\left(E_i\right)+\mathbb{P}\left(E_j\right)\right)+\epsilon^2,
\end{aligned}
$$

so

$$
\begin{aligned}
\|W_1-W_2\|^2&=\sum_{i,j=1}^{N}\frac{1}{N^2}\mathbb{E}\left(\left|W\left(\tilde{U}_i,\tilde{U}_j\right)-W\left(U_i,U_j\right)\right|^2\right)\le\sum_{i,j=1}^{N}\frac{1}{N^2}\left(4\left(\mathbb{P}\left(E_i\right)+\mathbb{P}\left(E_j\right)\right)+\epsilon^2\right)\\
&=\frac{8}{N}\sum_{i=1}^{N}\mathbb{P}\left(\left|U_i-\tilde{U}_i\right|>\delta_\epsilon\right)+\epsilon^2=8\mathbb{P}\left(\left|U_i-\tilde{U}_i\right|>\delta_\epsilon\right)+\epsilon^2,
\end{aligned}
$$

where the last equality clearly follows from the exchangeability and the definition of $\tilde{U}_i$. By Theorem 3, we indeed have $\mathbb{P}\left(\left|U_i-\tilde{U}_i\right|>\delta_\epsilon\right)\to 0$, so we have $\limsup_{N\to\infty}\|W_1-W_2\|\le\epsilon^2$. Letting $\epsilon\to 0$ implies $\limsup_{N\to\infty}\|W_1-W_2\|=0$, which then finishes the proof to the Theorem 1. □

A.2 Chapter 2

Proof of Lemma 4. We can write $\boldsymbol{\theta}=\boldsymbol{\mu}+\boldsymbol{Z}_1$ and $\boldsymbol{Y}=\boldsymbol{\theta}+\boldsymbol{Z}_2$, where $\boldsymbol{Z}_1\sim\mathcal{N}_p(\mathbf{0},\boldsymbol{B})$ and $\boldsymbol{Z}_2\sim\mathcal{N}_p(\mathbf{0},\boldsymbol{A})$ are independent. Jointly $\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{\theta}\end{pmatrix}$ is still multivariate normal with mean vector $\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{pmatrix}$ and covariance matrix $\begin{pmatrix}\boldsymbol{A}+\boldsymbol{B} & \boldsymbol{B}\\ \boldsymbol{B} & \boldsymbol{B}\end{pmatrix}$. The result follows immediately from the conditional distribution of a multivariate normal distribution. □

Proof of Theorem 5. We start from decomposing the difference between the URE and the actual loss as

$$
\mathrm{URE}\left(\boldsymbol{B},\boldsymbol{\mu}\right)-l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)
$$

$$= \mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{0}_p\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{0}_p}\right) - \frac{2}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\right) \tag{A.4}$$

$$= \frac{1}{p}\mathrm{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^T\right) - \frac{2}{p}\mathrm{tr}\left(\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}\boldsymbol{\theta}^T - \boldsymbol{A}\right)\right) \tag{A.5}$$

$$- \frac{2}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\right)$$

$$= (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).$$

To verify the first equality (A.4), note that

$$\begin{aligned}
&\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - \mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{0}_p\right)\\
&= \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\right\|^2 - \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{Y}\right\|^2\\
&= -\frac{1}{p}\mathrm{tr}\left(\boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)^T\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(2\boldsymbol{Y}-\boldsymbol{\mu}\right)\right),\\
&l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{0}_p}\right)\\
&= \frac{1}{p}\left\|\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu} - \boldsymbol{\theta}\right\|^2 - \frac{1}{p}\left\|\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y} - \boldsymbol{\theta}\right\|^2\\
&= \frac{1}{p}\mathrm{tr}\left(\boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)^T\left(2\left(\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y} - \boldsymbol{\theta}\right) + \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\right)\right).
\end{aligned}$$

(A.4) then follows by rearranging the terms. To verify the second equality (A.5), note

$$\begin{aligned}
&\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{0}_p\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{0}_p}\right)\\
&= \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{Y}\right\|^2 - \frac{1}{p}\left\|\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y} - \boldsymbol{\theta}\right\|^2 + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A} - 2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}\right)\\
&= \frac{1}{p}\mathrm{tr}\left(\left(\boldsymbol{Y} - 2\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y} + \boldsymbol{\theta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right) + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A} - 2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}\right)\\
&= \frac{1}{p}\mathrm{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^T\right) - \frac{2}{p}\mathrm{tr}\left(\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T - \boldsymbol{A}\right)\right).
\end{aligned}$$

With the decomposition, we want to prove separately the uniform L^1 convergence of the three terms (I), (II), and (III).

Proof for the case of Model I. The uniform L^2 convergence of (I) and (II) has been shown in

Theorem 3.1 of Xie et al. (2012) under our assumptions (A) and (B), so we focus on (III), i.e., we want to show that $\sup_{0\le\lambda\le\infty,\, \boldsymbol{\mu}\in\mathcal{L}} |(\mathrm{III})| \to 0$ in L^1 as $p\to\infty$.

Without loss of generality, let us assume $A_1 \le A_2 \le \cdots \le A_p$. We have

$$\begin{aligned}\sup_{0\le\lambda\le\infty,\, \boldsymbol{\mu}\in\mathcal{L}} |(\mathrm{III})| &= \frac{2}{p}\sup_{0\le\lambda\le\infty,\, \boldsymbol{\mu}\in\mathcal{L}} \left|\sum_{i=1}^{p} \frac{A_i}{A_i+\lambda}\mu_i\left(Y_i-\theta_i\right)\right| \\ &\le \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sup_{0\le c_1\le\cdots\le c_p\le 1}\left|\sum_{i=1}^{p} c_i\mu_i\left(Y_i-\theta_i\right)\right| = \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\max_{1\le j\le p}\left|\sum_{i=j}^{p}\mu_i\left(Y_i-\theta_i\right)\right|,\end{aligned}$$

where the last equality follows from Lemma 2.1 of Li (1986). For a generic p-dimensional vector $\boldsymbol{v}$, we denote $[\boldsymbol{v}]_{j:p} = (0,\ldots 0, v_j, v_{j+1},\ldots,v_p)$. Let $\boldsymbol{P_X} = \boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}$ be the projection matrix onto $\mathcal{L}_{\mathrm{row}}\left(\boldsymbol{X}\right)$. Then since $\mathcal{L}\subset\mathcal{L}_{\mathrm{row}}\left(\boldsymbol{X}\right)$, we have

$$\begin{aligned}&\frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\max_{1\le j\le p}\left|\sum_{i=j}^{p}\mu_i\left(Y_i-\theta_i\right)\right| = \frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|\boldsymbol{\mu}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right| \\ =&\frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|\boldsymbol{\mu}^T\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right| \le \frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\|\boldsymbol{\mu}\|\times\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\| \\ =&\frac{2}{p}\max_{1\le j\le p} Mp^{\kappa}\|\boldsymbol{Y}\|\times\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\|.\end{aligned}$$

Cauchy-Schwarz inequality thus gives

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\boldsymbol{\mu}\in\mathcal{L}}|(\mathrm{III})|\right) \le 2Mp^{\kappa-1}\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)}\times\sqrt{\mathbb{E}\left(\max_{1\le j\le p}\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\|^2\right)}. \tag{A.6}$$

It is straightforward to see that, by conditions (A) and (C),

$$\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)} = \sqrt{\mathbb{E}(\textstyle\sum_{i=1}^{p}Y_i^2)} = \sqrt{\textstyle\sum_{i=1}^{p}\left(\theta_i^2+A_i\right)} = O\left(p^{1/2}\right).$$

For the second term on the right hand side of (A.6), let $\boldsymbol{P_X} = \boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\Gamma}^T$ denote the spectral decomposition. Clearly,

$$\boldsymbol{D} = \mathrm{diag}\left(\underbrace{1,...,1}_{k\text{ copies}},\underbrace{0,...,0}_{p-k\text{ copies}}\right).$$

It follows that

$$
\begin{aligned}
&\mathbb{E}\left(\max_{1\le j\le p}\|\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\|^2\right)\\
&=\mathbb{E}\left(\max_{1\le j\le p}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}^T\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right)=\mathbb{E}\left(\max_{1\le j\le p}\operatorname{tr}\left(\boldsymbol{D}\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\left(\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right)^T\right)\right)\\
&=\mathbb{E}\left(\max_{1\le j\le p}\sum_{l=1}^{k}\left[\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right]_l^2\right)=\mathbb{E}\left(\max_{1\le j\le p}\sum_{l=1}^{k}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right)\\
&\le\mathbb{E}\left(\sum_{l=1}^{k}\max_{1\le j\le p}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right)=\sum_{l=1}^{k}\mathbb{E}\left(\max_{1\le j\le p}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right).
\end{aligned}
$$

For each l, $M_j^{(l)}=\sum_{m=p-j+1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)$ forms a martingale, so by Doob's L^p maximum inequality,

$$
\begin{aligned}
\mathbb{E}\left(\max_{1\le j\le p}\left(M_j^{(l)}\right)^2\right)\le 4\mathbb{E}\left(M_p^{(l)}\right)^2&=4\mathbb{E}\left(\sum_{m=1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\\
&=4\sum_{m=1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}^2A_m=4\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left(\max_{1\le j\le p}\|\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\|^2\right)&\le\sum_{l=1}^{k}4\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}=4\sum_{l=1}^{p}[\boldsymbol{D}]_{ll}\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}\\
&=4\operatorname{tr}\left(\boldsymbol{D}\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right)=4\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}\right)\\
&=4\operatorname{tr}\left(\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\right)=4\operatorname{tr}\left(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right)=O(1),
\end{aligned}
$$

where the last equality uses conditions (D) and (E). We finally obtain

$$
\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathcal{L}}|(\mathrm{III})|\right)\le o\left(p^{-1/2}\right)\times O\left(p^{1/2}\right)\times O(1)=o(1). \qquad \square
$$

Proof for the case of Model II. Under Model II, we know that

$$
\sum_{i=1}^{p}A_i\theta_i^2=\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}=\boldsymbol{\beta}^T(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T)\boldsymbol{\beta}=O(p)
$$

by condition (D). In other words, condition (D) implies condition (B). Therefore, we know that the term (I) → 0 in L^2 as shown in Theorem 3.1 of Xie et al. (2012), and we only need to show the uniform L^1 convergence of the other two terms, (II) and (III).

Recall that $\boldsymbol{B} \in \mathcal{B} = \left\{\lambda \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} : \lambda > 0\right\}$ has only rank k under Model II. We can reexpress (II) and (III) in terms of low rank matrices. Let $\boldsymbol{V} = \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}$. Woodbury formula gives

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{B})^{-1} = \left(\boldsymbol{A}+\lambda \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1} &= \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\lambda\boldsymbol{X}^T\left(\boldsymbol{W}^{-1}+\lambda\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\\ &= \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1},\end{aligned}$$

which tells us

$$\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1} = \boldsymbol{I}_p - \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1} = \lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1}.$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the spectral decomposition of $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}$, i.e., $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda} = \operatorname{diag}(d_1, ..., d_k)$ with $d_1 \leq \cdots \leq d_k$. Then

$$(\lambda\boldsymbol{W}+\boldsymbol{V})^{-1} = \boldsymbol{W}^{-1/2}\left(\lambda\boldsymbol{I}_k + \boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}\right)^{-1}\boldsymbol{W}^{-1/2} = \boldsymbol{W}^{-1/2}\boldsymbol{U}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2},$$

from which we obtain

$$\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1} = \lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1} = \lambda\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{U}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{A}^{-1}.$$

If we denote $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$, i.e., $\boldsymbol{Z}$ is the transformed covariate matrix, then $\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1} = \lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}$. It follows that

$$\begin{aligned}(\text{II}) &= -\frac{2}{p}\operatorname{tr}\left(\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}\boldsymbol{\theta}^T - \boldsymbol{A}\right)\right)\\ &= -\frac{2}{p}\operatorname{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}\boldsymbol{\theta}^T - \boldsymbol{A}\right)\right)\\ &= -\frac{2}{p}\operatorname{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}\boldsymbol{\theta}^T - \boldsymbol{A}\right)\boldsymbol{Z}^T\right),\end{aligned}$$

$$
\begin{aligned}
\text{(III)} &= -\frac{2}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right)\\
&= -\frac{2}{p}\mathrm{tr}\left(\left(\boldsymbol{I}_p-\lambda\boldsymbol{Z}^{T}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\right)\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right)\\
&= -\frac{2}{p}\mathrm{tr}\left(\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right)+\frac{2}{p}\mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\boldsymbol{Z}^{T}\right)\\
&= \text{(III)}_1+\text{(III)}_2\,.
\end{aligned}
$$

We will next show that (II), (III)_1, and (III)_2 all uniformly converge to zero in L^1, which will then complete our proof.

Let $\boldsymbol{\Xi}=\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^{T}-\boldsymbol{Y}\boldsymbol{\theta}^{T}-\boldsymbol{A}\right)\boldsymbol{Z}^{T}$. Then

$$
\begin{aligned}
\sup_{0\le\lambda\le\infty}\left|\text{(II)}\right| &= \frac{2}{p}\sup_{0\le\lambda\le\infty}\left|\sum_{i=1}^{k}\frac{\lambda d_i}{\lambda+d_i}\left[\boldsymbol{\Xi}\right]_{ii}\right|\\
&\le \frac{2}{p}\sup_{0\le c_1\le\cdots\le c_k\le d_k}\left|\sum_{i=1}^{k}c_i\left[\boldsymbol{\Xi}\right]_{ii}\right| = \frac{2}{p}\max_{1\le j\le k}\left|\sum_{i=j}^{k}d_k\left[\boldsymbol{\Xi}\right]_{ii}\right|,
\end{aligned}
$$

where the last equality follows as in Lemma 2.1 of Li (1986). As there are finite number of terms in the summation and the maximization, it suffices to show that

$$
d_k\left[\boldsymbol{\Xi}\right]_{ii}/p\to 0 \text{ in } L^2 \quad \text{for all } 1\le i\le k.
$$

To establish this, we note that $\left[\boldsymbol{\Xi}\right]_{ii}=\sum_{n=1}^{p}\sum_{m=1}^{p}\left(A_n^{-1}Y_n\left(Y_m-\theta_m\right)-\delta_{nm}\right)\left[\boldsymbol{Z}\right]_{in}\left[\boldsymbol{Z}\right]_{im}$,

$$
\begin{aligned}
\mathbb{E}\left(\left[\boldsymbol{\Xi}\right]_{ii}^{2}\right) = \sum_{n,m,n',m'}&\mathbb{E}\left(\left(A_n^{-1}Y_n\left(Y_m-\theta_m\right)-\delta_{nm}\right)\left(A_{n'}^{-1}Y_{n'}\left(Y_{m'}-\theta_{m'}\right)-\delta_{n'm'}\right)\right)\\
&\times\left[\boldsymbol{Z}\right]_{in}\left[\boldsymbol{Z}\right]_{im}\left[\boldsymbol{Z}\right]_{in'}\left[\boldsymbol{Z}\right]_{im'}\,.
\end{aligned}
$$

Depending on n,m,n',m' taking the same or distinct values, we can break the summation into 15 disjoint cases:

$$
\sum_{\text{all distinct}}+\sum_{\text{three distinct, }n=m}+\sum_{\text{three distinct, }n=n'}+\sum_{\text{three distinct, }n=m'}
$$

$$
\begin{aligned}
&+ \sum_{\text{three distinct, } m=n'} + \sum_{\text{three distinct, } m=m'} + \sum_{\text{three distinct, } n'=m'} + \sum_{\text{two distinct, } n=m,\ n'=m'} \\
&+ \sum_{\text{two distinct, } n=n',\ m=m'} + \sum_{\text{two distinct, } n=m',\ n'=m} + \sum_{\text{two distinct, } n=m=n'} + \sum_{\text{two distinct, } n=m=m'} \\
&+ \sum_{\text{two distinct, } n=n'=m'} + \sum_{\text{two distinct, } m=n'=m'} + \sum_{n=m=n'=m'} .
\end{aligned}
$$

Many terms are zero. Straightforward evaluation of each summation gives

$$
\begin{aligned}
\mathbb{E}\left([\boldsymbol{\Xi}]_{ii}^2\right) &= \sum_{n=1}^{p} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_n - \theta_n\right) - 1\right)^2\right) [\boldsymbol{Z}]_{in}^4 \\
&+ \sum_{n=1}^{p} \sum_{m \neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_m - \theta_m\right)\right)^2\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&+ \sum_{n=1}^{p} \sum_{m \neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_m - \theta_m\right)\right)\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&+ 2 \sum_{n=1}^{p} \sum_{m \neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_n - \theta_n\right) - 1\right)\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^3 [\boldsymbol{Z}]_{im} \\
&+ \sum_{n=1}^{p} \sum_{m \neq n', n' \neq n, m \neq n} \mathbb{E}\left(\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\left(A_{n'}^{-1} Y_{n'} \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'} \\
&= \sum_{n=1}^{p} \frac{2A_n + \theta_n^2}{A_n} [\boldsymbol{Z}]_{in}^4 + \sum_{n=1}^{p} \sum_{m \neq n} \frac{A_n A_m + A_n \theta_m^2}{A_m^2} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n=1}^{p} \sum_{m \neq n} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&+ 2 \sum_{n=1}^{p} \sum_{m \neq n} \frac{\theta_n \theta_m}{A_m} [\boldsymbol{Z}]_{in}^3 [\boldsymbol{Z}]_{im} + \sum_{n=1}^{p} \sum_{m \neq n', n' \neq n, m \neq n} \frac{A_n \theta_m \theta_{n'}}{A_m A_{n'}} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'} \\
&= \sum_{n,m=1}^{p} \frac{A_n}{A_m} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n,m=1}^{p} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n,m,n'=1}^{p} \frac{A_n \theta_m \theta_{n'}}{A_m A_{n'}} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'} .
\end{aligned}
$$

Using matrix notation, we can reexpress the above equation as

$$
\begin{aligned}
\mathbb{E}\left([\boldsymbol{\Xi}]_{ii}^2\right) &= \left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right]_{ii} \left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right]_{ii} + \left[\boldsymbol{Z}\boldsymbol{Z}^T\right]_{ii}^2 + \left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right]_{ii} \left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\theta}\right]_{i}^2 \\
&\leq \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right) \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right) + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^T\right)^2 + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right) \operatorname{tr}\left(\boldsymbol{\theta}^T\boldsymbol{A}^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\theta}\right) \\
&= \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right) + \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^T\right)^2 \\
&+ \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) \operatorname{tr}\left(\boldsymbol{\beta}^T\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\boldsymbol{W}\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\boldsymbol{\beta}\right),
\end{aligned}
$$

which is $O\left(p\right)O\left(p\right)+O\left(p\right)^2+O\left(p\right)O\left(p^2\right)=O\left(p^3\right)$ by conditions (D)-(F). Note also that condition (F) implies

$$d_k \le \sum_{i=1}^{k} d_i = \mathrm{tr}\left(\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}\right) = \mathrm{tr}\left(\boldsymbol{W}^{-1}\boldsymbol{V}\right) = \mathrm{tr}\left(\boldsymbol{W}^{-1}(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}\right) = O\left(p^{-1}\right).$$

Therefore, we have

$$\mathbb{E}\left(d_k^2\left[\boldsymbol{\Xi}\right]_{ii}^2/p^2\right) = O\left(p^{-2}\right)O\left(p^3\right)/p^2 = O\left(p^{-1}\right) \to 0,$$

which proves

$$\sup_{0\le\lambda\le\infty}\left|(\mathrm{II})\right| \to 0 \text{ in } L^2, \quad \text{as } p\to\infty.$$

To prove the uniform convergence of $(\mathrm{III})_1$ to zero in L^1, we note that

$$\begin{aligned}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|(\mathrm{III})_1\right| &= \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|\boldsymbol{\mu}^T\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right| = \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|\boldsymbol{\mu}^T\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right| \\ &\le \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left\|\boldsymbol{\mu}\right\|\times\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right\| = \frac{2}{p}Mp^{\kappa}\left\|\boldsymbol{Y}\right\|\times\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right\|,\end{aligned}$$

so by Cauchy-Schwarz inequality

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left|(\mathrm{III})_1\right|\right) \le 2Mp^{\kappa-1}\sqrt{\mathbb{E}\left(\left\|\boldsymbol{Y}\right\|^2\right)}\sqrt{\mathbb{E}\left(\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right\|^2\right)}. \tag{A.7}$$

Under Model II, $\boldsymbol{\theta}=\boldsymbol{X}^T\boldsymbol{\beta}$, so it follows that $\sum_{i=1}^p\theta_i^2=\left\|\boldsymbol{\theta}\right\|^2=\mathrm{tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}\boldsymbol{X}^T\right)=O\left(p\right)$ by condition (E). Hence $\sqrt{\mathbb{E}\left(\left\|\boldsymbol{Y}\right\|^2\right)}=\sqrt{\sum_{i=1}^p\left(\theta_i^2+A_i\right)}=O\left(p^{1/2}\right)$. For the second term on the right hand side of (A.7), note that

$$\begin{aligned}\mathbb{E}\left(\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right\|^2\right) &= \mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\right)\right) \\ &= \mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}\right) = \mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) = O\left(1\right)\end{aligned}$$

by conditions (D) and (E). Thus, in aggregate, we have

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathcal{L}} \left|(\mathrm{III})_1\right|\right) \leq 2Mp^{\kappa-1}O\left(p^{1/2}\right)O\left(1\right) = o\left(1\right).$$

We finally consider the $(\mathrm{III})_2$ term. We have

$$\begin{aligned}
\sup_{0\leq\lambda\leq\infty,\, \boldsymbol{\mu}\in\mathcal{L}} \left|(\mathrm{III})_2\right| &= \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sup_{0\leq\lambda\leq\infty}\left|\sum_{i=1}^{k}\frac{\lambda d_i}{\lambda+d_i}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&\leq \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\max_{1\leq j\leq k}\left|\sum_{i=j}^{k} d_k\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&\leq \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left|\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&= \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left|\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i\left[\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right]_i\right| \\
&\leq \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sqrt{\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2}\times\sqrt{\sum_{i=1}^{k}\left[\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right]_i^2}.
\end{aligned}$$

Thus, by Cauchy-Schwarz inequality

$$\mathbb{E}\left(\sup_{0\leq\lambda\leq\infty,\, \boldsymbol{\mu}\in\mathcal{L}}\left|(\mathrm{III})_2\right|\right) \leq \frac{2d_k}{p}\sqrt{\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2\right)}\times\sqrt{\mathbb{E}\left(\sum_{i=1}^{k}\left[\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right]_i^2\right)}.$$

Note that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2 = \sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left(\sum_{m=1}^{p}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\right]_{im}\left[\boldsymbol{\mu}\right]_m\right)^2 \\
&\leq \sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left(\sum_{m=1}^{p}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\right]_{im}^2\times\sum_{m=1}^{p}\left[\boldsymbol{\mu}\right]_m^2\right) = \sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left(\left[\boldsymbol{Z}\boldsymbol{A}^{-2}\boldsymbol{Z}^T\right]_{ii}\left\|\boldsymbol{\mu}\right\|^2\right) \\
&= \mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-2}\boldsymbol{Z}^T\right)\sup_{\boldsymbol{\mu}\in\mathcal{L}}\left\|\boldsymbol{\mu}\right\|^2 = \mathrm{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-2}\boldsymbol{X}^T\right)\left(Mp^{\kappa}\left\|\boldsymbol{Y}\right\|\right)^2 = o\left(p^2\right)\left\|\boldsymbol{Y}\right\|^2,
\end{aligned}$$

where the last equality uses condition (G). Thus,

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathcal{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2\right) = o\left(p^3\right).$$

Also note that

$$\mathbb{E}\left(\sum_{i=1}^{k}\left[\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right]_i^2\right)=\mathbb{E}\left(\operatorname{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\right)\right)=\operatorname{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\right)=\operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right)=O\left(p\right)$$

by condition (D). Recall that $d_k = O\left(p^{-1}\right)$ by condition (F). It follows that

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathcal{L}}\left|(\mathrm{III})_2\right|\right)\le\frac{2}{p}O\left(p^{-1}\right)o\left(p^{3/2}\right)O\left(p^{1/2}\right)=o\left(1\right). \qquad \square$$

This completes our proof with Theorem 5 under both Model I and II. □

Proof of Lemma 6. The fact that $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}\in\mathcal{L}$ is trivial as

$$\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}=\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{Y}=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y},$$

while the projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ has induced matrix 2-norm $\left\|\boldsymbol{P}_{\boldsymbol{X}}\right\|_2 = 1$. Thus, $\left\|\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}\right\| \le \left\|\boldsymbol{P}_{\boldsymbol{X}}\right\|_2\left\|\boldsymbol{Y}\right\| = \left\|\boldsymbol{Y}\right\|$. For $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, note that

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}&=\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}\\&=\boldsymbol{A}^{1/2}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)^T\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)\boldsymbol{A}^{-1/2}\boldsymbol{Y}\\&=\boldsymbol{A}^{1/2}\left(\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}^{-1/2}}\right)\boldsymbol{A}^{-1/2}\boldsymbol{Y},\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}^{-1/2}}$ is the ordinary projection matrix onto the row space of $\boldsymbol{X}\boldsymbol{A}^{-1/2}$ and has induced matrix 2-norm 1. It follows

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}\right\|\le\left\|\boldsymbol{A}^{1/2}\right\|_2\left\|\boldsymbol{P}_{\boldsymbol{A}^{-1/2}\boldsymbol{X}}\right\|_2\left\|\boldsymbol{A}^{-1/2}\right\|_2\left\|\boldsymbol{Y}\right\|=\max_{1\le i\le p}A_i^{1/2}\times\max_{1\le i\le p}A_i^{-1/2}\times\left\|\boldsymbol{Y}\right\|.$$

Condition (A) gives

$$\max_{1\le i\le p}A_i^{1/2}=(\max_{1\le i\le p}A_i^2)^{1/4}\le(\sum_{i=1}^{p}A_i^2)^{1/4}=O\left(p^{1/4}\right).$$

Similarly, condition (A$'$) gives

$$\max_{1\le i\le p} A_i^{-1/2} = (\max_{1\le i\le p} A_i^{-2-\delta})^{1/(4+2\delta)} \le (\sum_{i=1}^{p} A_i^{-2-\delta})^{1/(4+2\delta)} = O\left(p^{1/(4+2\delta)}\right).$$

We then have proved that

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}\right\| \le O\left(p^{1/4}\right) O\left(p^{1/(4+2\delta)}\right) \|\boldsymbol{Y}\| = O\left(p^{\kappa}\right) \|\boldsymbol{Y}\|. \qquad \square$$

Proof of Theorem 7. To prove the first assertion, note that

$$\mathrm{URE}\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}}\right) \le \mathrm{URE}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right)$$

by the definition of $\hat{\boldsymbol{B}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{URE}}$, so Theorem 5 implies that

$$\begin{aligned}
& l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \\
& \le l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - \mathrm{URE}\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}}\right) + \mathrm{URE}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \\
& \le 2 \sup_{\boldsymbol{B}\in\mathcal{B},\, \boldsymbol{\mu}\in\mathcal{L}} \left|\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \underset{p\to\infty}{\to} 0 \text{ in } L^1 \text{ and in probability,}
\end{aligned} \tag{A.8}$$

where the second inequality uses the condition that $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathcal{L}$. Thus, for any $\epsilon > 0$,

$$\mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \ge l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) + \epsilon\right) \le \mathbb{P}\left(2 \sup_{\boldsymbol{B}\in\mathcal{B},\, \boldsymbol{\mu}\in\mathcal{L}} \left|\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \ge \epsilon\right) \to 0.$$

To prove the second assertion, note that

$$l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \le l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right)$$

by the definition of $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ and the condition $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathcal{L}$. Thus, taking expectations on equation (A.8) easily gives the second assertion. □

Proof of Corollary 8. Simply note that

$$l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \le l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right)$$

by the definition of $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$. Thus,

$$l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right) \leq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right).$$

Then Theorem 7 clearly implies the desired result. □

Proof of Theorem 9. We observe that

$$\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) = \mathrm{URE}\left(\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) + \frac{2}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right).$$

Since

$$\sup_{\boldsymbol{B}\in\mathcal{B}}\left|\mathrm{URE}\left(\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right)\right| \leq \sup_{\boldsymbol{B}\in\mathcal{B},\ \boldsymbol{\mu}\in\mathcal{L}}\left|\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B}, \boldsymbol{\mu}}\right)\right| \to 0 \text{ in } L^1$$

by Theorem 5, we only need to show that

$$\sup_{\boldsymbol{B}\in\mathcal{B}}\left|\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right| \to 0 \quad \text{as } p\to\infty.$$

Under Model I,

$$\begin{aligned}
&\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) \\
&= \sum_{i=1}^{p}\frac{A_i}{A_i+\lambda}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii} \leq \left(\sum_{i=1}^{p}(\frac{A_i}{A_i+\lambda})^2 \times \sum_{i=1}^{p}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii}^2\right)^{1/2} \\
&\leq \left(p \times \sum_{i=1}^{p}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii}^2\right)^{1/2} \leq p^{1/2}\sqrt{\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A})^T\right)}, \qquad \text{for all } \lambda \geq 0,
\end{aligned}$$

but

$$\begin{aligned}
\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}^T\right) &= \mathrm{tr}\left(\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T\right)\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right) = O(1)
\end{aligned}$$

by (2.13) and condition (E). Therefore,

$$\sup_{\boldsymbol{B}\in\mathcal{B}}\left|\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right| = \frac{1}{p}O\left(p^{1/2}\right)O(1) = O(p^{-1/2}) \to 0.$$

Under Model II,

$$\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1} = \boldsymbol{I}_p - \lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1},$$

where $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}\left(d_1, ..., d_k\right)$ with $d_1 \leq \cdots \leq d_k$, and $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$ as defined in the proof of Theorem 5. Thus,

$$\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) = \mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) - \mathrm{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right).$$

We know that $\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) = \mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{X}^T\right)\right) = O(1)$ by the assumption (2.13).

$$\begin{aligned}\mathrm{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) &= \mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\boldsymbol{Z}^T\right)\\ &= \mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\right).\end{aligned}$$

The Cauchy-Schwarz inequality for matrix trace gives

$$\begin{aligned}&\left|\mathrm{tr}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\right)\right)\right|\\ &\leq \mathrm{tr}^{1/2}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)^2\right)\\ &\times \mathrm{tr}^{1/2}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right).\end{aligned}$$

Since

$$\mathrm{tr}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)^2\right) = \sum_{i=1}^{k}\left(\frac{\lambda d_i}{\lambda+d_i}\right)^2 \leq kd_k^2 = O\left(p^{-2}\right) \qquad \text{for all } \lambda \geq 0$$

as shown in the proof of Theorem 5 and

$$\begin{aligned}&\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right)\\ &= \mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\end{aligned}$$

$$= \operatorname{tr}\left(\left(\boldsymbol{XMX}^T\right)^{-1}(\boldsymbol{XMAX}^T)\boldsymbol{W}(\boldsymbol{XAMX}^T)\left(\boldsymbol{XMX}^T\right)^{-1}(\boldsymbol{XA}^{-1}\boldsymbol{X}^T)\boldsymbol{W}(\boldsymbol{XA}^{-1}\boldsymbol{X}^T)\right) = O(p^2)$$

from (2.13) and condition (F), we have

$$\sup_{\boldsymbol{B}\in\mathcal{B}}\left|\frac{1}{p}\operatorname{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right| = \frac{1}{p}\left(O(1)+\sqrt{O\left(p^{-2}\right)\times O(p^2)}\right) = O(p^{-1}) \to 0.$$

This completes our proof of (2.14). With this established, the rest of the proof is identical to that of Theorem 7 and Corollary 8. □

A.3 Chapter 3

Details of Remark 14. To see this, we first argue that P_t is a semimartingale with respect to the complete data filtration $\mathcal{F}_t \vee \mathcal{S}_t$, which includes the history of the price process itself ($\mathcal{F}_t$) and the history of all the hidden activities of events ($\mathcal{S}_t$). Precisely, $\mathcal{S}_t$ is the natural filtration generated by the process of random set S_t, which consists of all the surviving events (q, y) in the trawl at time t, where $q \leq t$ is the original arrival time of the event and y is its size. Then clearly $L\left(A_t\right) = \sum_{(q,y)\in S_t} y$ is a càdlàg adapted process (w.r.t. $\mathcal{S}_t$) of locally bounded variation if the underlying Lévy basis has finite activities.

Denote the natural filtration generated by the path of $L\left(B_t\right)$ as $\mathcal{L}_t$. Then from the definition of P_t the complete data information $\mathcal{F}_t \vee \mathcal{S}_t$ must be the same as $\mathcal{L}_t \vee \mathcal{S}_t \vee \sigma\left(V_0\right)$—the path of $L\left(B_t\right)$ will be completely revealed under $\mathcal{F}_t \vee \mathcal{S}_t$, where $\{\mathcal{L}_t\}$, $\{\mathcal{S}_t\}$ and V_0 are completely independent to each other. Thus,

$$M_t = L\left(B_t\right) - b\left(\sum_{y\in\mathbb{Z}\setminus\{0\}} y\nu\left(y\right)\right) t \in \mathcal{L}_t \subseteq \mathcal{F}_t \vee \mathcal{S}_t$$

must be a martingale w.r.t. $\mathcal{F}_t \vee \mathcal{S}_t$ because

$$\mathbb{E}\left(M_t|\mathcal{F}_s,\mathcal{S}_s\right) = \mathbb{E}\left(M_t|\mathcal{L}_s,\mathcal{S}_s,V_0\right) = \mathbb{E}\left(M_t|\mathcal{L}_s\right) = M_s,$$

where the second equality follows from the independence between $\{\mathcal{L}_t\}$, $\{\mathcal{S}_t\}$ and V_0.

Write

$$P_t = M_t + Q_t, \qquad Q_t = V_0 + L\left(A_t\right) + b\left(\sum_{y\in\mathbb{Z}\setminus\{0\}} y\nu\left(y\right)\right)t.$$

As V_0 can be revealed under $\mathcal{F}_0$ and $\mathcal{S}_0$, it is trivially in $\mathcal{F}_t \vee \mathcal{S}_t$, too. Then Q_t is also a càdlàg adapted process (w.r.t. $\mathcal{F}_t \vee \mathcal{S}_t$) of locally bounded variation. We then conclude that P_t is a semimartingale w.r.t. $\mathcal{F}_t \vee \mathcal{S}_t$. As the property of being a semimartingale is preserved under shrinking the filtration, P_t is a semimartingale w.r.t. $\mathcal{F}_t \subseteq \mathcal{F}_t \vee \mathcal{S}_t$. □

Proof of Theorem 14. We partition C_t and C_0 into three disjoint sets, one of which is in common:

$$C_t = \left(C_t \cap C_0\right) \cup \left(C_t \backslash C_0\right), \qquad C_0 = \left(C_t \cap C_0\right) \cup \left(C_0 \backslash C_t\right),$$

so this means that

$$P_t - P_0 = L\left(C_t \backslash C_0\right) - L(C_0 \backslash C_t).$$

$L\left(C_t \backslash C_0\right)$ is clearly independent of $L(C_0 \backslash C_t)$ due to the independence property of the Lévy basis and the disjointedness between $C_t \backslash C_0$ and $C_0 \backslash C_t$.

For any $t \geq 0$,

$$\begin{aligned}
C_t \backslash C_0 &= \left(A_t \backslash A\right) \cup B_t = \left(A_t \backslash A\right) \cup \left([0,b) \times (0,t]\right) \\
C_0 \backslash C_t &= A \backslash A_t, \\
leb\left(C_t \backslash C_0\right) &= leb\left(A_t \backslash A\right) + bt, \\
leb\left(C_0 \backslash C_t\right) &= leb\left(A \backslash A_t\right) = leb\left(A_t \backslash A\right).
\end{aligned}$$

Then

$$C\left(\theta \ddagger P_t - P_0\right) = C\left(\theta \ddagger L\left(C_t \backslash C_0\right)\right) + C\left(-\theta \ddagger L\left(C_0 \backslash C_t\right)\right),$$

$$= leb(C_t \backslash C_0) C\left(\theta \ddagger L_1\right) + leb(C_0 \backslash C_t) C\left(-\theta \ddagger L_1\right)$$

$$= btC\left(\theta \ddagger L_1\right) + leb(A_t \backslash A)\left(C\left(\theta \ddagger L_1\right) + C\left(-\theta \ddagger L_1\right)\right).$$

For any random variable X we always have

$$\kappa_j\left(X\right) = \frac{1}{\mathrm{i}^j} \frac{\partial^j}{\partial^j \theta} C\left(\theta \ddagger X\right)\bigg|_{\theta=0},$$

so using the equation above it is clear that

$$\kappa_j(P_t - P_0) = \left(bt + leb(A_t \backslash A)\left(1 + (-1)^j\right)\right)\kappa_j(L_1),$$

which is the required result. □

Proof of Theorem 15. For each $y \in \mathbb{Z}\backslash\left\{0\right\}$, the price process has a jump with size y if and only if either one event with size y arrives or one event with size $-y$ departures—thanks to the monotonicity of d. Thus, the probability of the arrival event can be characterized by the non-zero probability of a Poisson random variable with intensity

$$\nu\left(y\right) leb\left(R_t \backslash R_{t-\mathrm{d}t}\right) \approx \nu\left(y\right)\mathrm{d}t;$$

on the other hand, the probability of the departure event can be similarly depicted by the non-zero probability of a Poisson random variable with intensity

$$\nu\left(-y\right) leb\left(A_{t-\mathrm{d}t} \backslash A_t\right) \approx \nu\left(-y\right)\left(1-b\right)\mathrm{d}t.$$

Therefore, by noting that $\mathbb{P}\left(X > 0\right) = 1 - e^{-\lambda} \approx \lambda$ for $X \backsim \text{Pois}\left(\lambda\right)$ and small λ, we have

$$\mathbb{P}\left(\Delta P_t = y | \Delta P_t \neq 0\right) = \frac{\mathbb{P}\left(\Delta P_t = y\right)}{\sum_{y \in \mathbb{Z}\backslash\{0\}} \mathbb{P}\left(\Delta P_t = y\right)} = \frac{\nu\left(y\right)\mathrm{d}t + \nu\left(-y\right)\left(1-b\right)\mathrm{d}t}{\sum_{y \in \mathbb{Z}\backslash\{0\}}\left(\nu\left(y\right)\mathrm{d}t + \nu\left(-y\right)\left(1-b\right)\mathrm{d}t\right)}$$
$$= \frac{\nu\left(y\right) + \nu\left(-y\right)\left(1-b\right)}{\left(2-b\right)\left\|\nu\right\|}.$$
□

Proof of Theorem 16. We will use the following straightforward result on the increments of a process to prove Theorem 16.

Lemma 21. *Suppose Z_t, for $t \in \mathbb{R}$, has covariance stationary increments. For $\delta > 0$, $k = 1, 2, 3, ...$*

$$\begin{aligned}\gamma_k &= \operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_\delta - Z_0\right)\\ &= \frac{1}{2}\operatorname{Var}\left(Z_{(k+1)\delta} - Z_{k\delta}\right) - \operatorname{Var}\left(Z_{k\delta} - Z_0\right) + \frac{1}{2}\operatorname{Var}\left(Z_{(k-1)\delta} - Z_0\right).\end{aligned}$$

Proof. First note that

$$\begin{aligned}\operatorname{Var}\left(Z_{(k+1)\delta} - Z_0\right) &= \operatorname{Var}\left(\left(Z_{(k+1)\delta} - Z_{k\delta}\right) + \left(Z_{k\delta} - Z_0\right)\right)\\ &= \operatorname{Var}\left(Z_\delta - Z_0\right) + \operatorname{Var}\left(Z_{k\delta} - Z_0\right) + 2\operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_{k\delta} - Z_0\right).\end{aligned}$$

By rearranging, we have

$$2\gamma_k^* = 2\operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_{k\delta} - Z_0\right) = \operatorname{Var}\left(Z_{(k+1)\delta} - Z_0\right) - \operatorname{Var}\left(Z_\delta - Z_0\right) - \operatorname{Var}\left(Z_{k\delta} - Z_0\right).$$

If $k \geq 2$, then

$$\begin{aligned}2\gamma_k^* &= 2\operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_{k\delta} - Z_0\right)\\ &= 2\operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_{k\delta} - Z_\delta\right) + 2\operatorname{Cov}\left(Z_{(k+1)\delta} - Z_{k\delta}, Z_\delta - Z_0\right) = 2\gamma_{k-1}^* + 2\gamma_k.\end{aligned}$$

Hence,

$$\begin{aligned}\gamma_k = \frac{2\gamma_k^* - 2\gamma_{k-1}^*}{2} &= \frac{1}{2}\begin{pmatrix}\operatorname{Var}\left(Z_{(k+1)\delta} - Z_0\right) - \operatorname{Var}\left(Z_\delta - Z_0\right) - \operatorname{Var}\left(Z_{k\delta} - Z_0\right)\\ -\left(\operatorname{Var}\left(Z_{k\delta} - Z_0\right) - \operatorname{Var}\left(Z_\delta - Z_0\right) - \operatorname{Var}\left(Z_{(k-1)\delta} - Z_0\right)\right)\end{pmatrix}\\ &= \frac{1}{2}\operatorname{Var}\left(Z_{(k+1)\delta} - Z_0\right) - \operatorname{Var}\left(Z_{k\delta} - Z_0\right) + \frac{1}{2}\operatorname{Var}\left(Z_{(k-1)\delta} - Z_0\right),\end{aligned}$$

which is the required result. □

Combining Lemma 21 and Theorem 14 gives us

$$\begin{aligned}\gamma_k &= \frac{1}{2}\left(\mathrm{Var}\left(P_{(k+1)\delta} - P_0\right) - 2\mathrm{Var}\left(P_{k\delta} - P_0\right) + \mathrm{Var}\left(P_{(k-1)\delta} - P_0\right)\right) \\ &= \frac{1}{2}\left(\begin{array}{c} b\left(k+1\right)\delta + 2leb\left(A_{(k+1)\delta}\backslash A\right) - 2\left(bk\delta + 2leb\left(A_{k\delta}\backslash A\right)\right) \\ +b\left(k-1\right)\delta + 2leb\left(A_{(k-1)\delta}\backslash A\right)\end{array}\right)\kappa_2\left(L_1\right) \\ &= \left(leb\left(A_{(k+1)\delta}\backslash A\right) - 2leb\left(A_{k\delta}\backslash A\right) + leb\left(A_{(k-1)\delta}\backslash A\right)\right)\kappa_2\left(L_1\right), \\ \rho_k &= \frac{\gamma_k}{\mathrm{Var}\left(P_\delta - P_0\right)} = \frac{leb\left(A_{(k+1)\delta}\backslash A\right) - 2leb\left(A_{k\delta}\backslash A\right) + leb\left(A_{(k-1)\delta}\backslash A\right)}{b\delta + 2leb\left(A_\delta\backslash A\right)}.\end{aligned}$$

□

Proof of Corollary 17. From Proposition 13 we have

$$\frac{\partial}{\partial t} leb\left(A_t \cap A\right) = -\left(d\left(-t\right) - b\right),$$

so mean value theorem states that, for any $0 \le t_1 < t_2 < t_3$, there exist $t_{23} \in (t_2, t_3)$ and $t_{12} \in (t_1, t_2)$ such that

$$\begin{aligned}\frac{leb\left(A_{t_3} \cap A\right) - leb\left(A_{t_2} \cap A\right)}{t_3 - t_2} &= -\left(d\left(-t_{23}\right) - b\right) \le -\left(d\left(-t_{12}\right) - b\right) \\ &= \frac{leb\left(A_{t_2} \cap A\right) - leb\left(A_{t_1} \cap A\right)}{t_2 - t_1},\end{aligned} \tag{A.9}$$

where the second inequality follows from the monotonicity of d and $t_{12} < t_{23}$. This proves that $leb\left(A_t \cap A\right)$ is a convex function of t. Hence, equation (3.3) implies

$$\begin{aligned}&leb(A_{(k+1)\delta}\backslash A) - 2leb(A_{k\delta}\backslash A) + leb(A_{(k-1)\delta}\backslash A) \\ &\quad = -leb\left(A_{(k+1)\delta} \cap A\right) + 2leb\left(A_{k\delta} \cap A\right) - leb\left(A_{(k-1)\delta} \cap A\right) \le 0,\end{aligned} \tag{A.10}$$

as required.

When d is a strictly increasing function, the inequality (A.9) becomes strict, so $leb\left(A_t \cap A\right)$ becomes a strictly convect function of t, which further makes inequality (A.10) strict, as required.

□

Proof of Theorem 18. Arrivals are in R_t and so aggregated to $\Sigma(R_t; r)$, while departures only happen at most once due to the monotonicity of d. All the departures are in G_t and so aggregated to $\Sigma(G_t; r)$. Now

$$\begin{aligned}\mathbb{E}\left(\{P\}_t^{[r]}\right) &= \mathbb{E}\left(\Sigma\left(B_t; r\right)\right) + \mathbb{E}\left(\Sigma\left(H_t; r\right)\right) + \mathbb{E}\left(\Sigma\left(G_t; r\right)\right) \\ &= \left(leb(B_t) + leb(H_t) + leb(G_t)\right) \int_{-\infty}^{\infty} |y|^r \, \nu(\mathrm{d}y) \\ &= \left(bt + (1-b)t + (1-b)t\right) \int_{-\infty}^{\infty} |y|^r \, \nu(\mathrm{d}y) = (2-b)\, t \int_{-\infty}^{\infty} |y|^r \, \nu(\mathrm{d}y),\end{aligned}$$

where the third equality follows from

$$leb\left(G_t\right) = leb\left(H_t \cup A\right) - leb\left(A_t\right) = leb\left(H_t\right) + leb\left(A\right) - leb\left(A_t\right) = leb\left(H_t\right) = (1-b)\, t. \qquad \square$$

Proof of Proposition 19. Stationarity of returns and the definition that $\delta_n = T/n$ imply

$$\begin{aligned}\mathbb{E}\left(RV^{(n)}\right) &= \sum_{k=1}^{n} \mathbb{E}\left(P_{k\delta_n} - P_{(k-1)\delta_n}\right)^2 = n \operatorname{Var}\left(P_{\delta_n} - P_0\right) + n\left(\mathbb{E}\left(P_{\delta_n} - P_0\right)\right)^2 \\ &= n\left(b\delta_n + 2leb\left(A_{\delta_n} \backslash A\right)\right) \kappa_2\left(L_1\right) + n\left(b\delta_n \kappa_1\left(L_1\right)\right)^2 \\ &= \left(b + 2\frac{leb\left(A_{\delta_n} \backslash A\right)}{\delta_n}\right) T\kappa_2\left(L_1\right) + b^2 T \delta_n \kappa_1^2\left(L_1\right). \qquad \square\end{aligned}$$

Details of Remark 23. For any $r \geq 0$ plug-in (3.13) into the left-hand side of (3.11). Then

$$\begin{aligned}(2-b) \sum_{y \in \mathbb{Z} \backslash \{0\}} |y|^r \, \widehat{\nu\left(y\right)} &= (2-b) \sum_{y \in \mathbb{Z} \backslash \{0\}} |y|^r \left(\frac{\hat{\alpha}_y - (1-b)\, \hat{\alpha}_{-y}}{(2-b)\, b} \hat{\beta}_0\right) \\ &= \frac{\sum_{y \in \mathbb{Z} \backslash \{0\}} |y|^r \, \hat{\alpha}_y - (1-b) \sum_{y \in \mathbb{Z} \backslash \{0\}} |y|^r \, \hat{\alpha}_{-y}}{b} \hat{\beta}_0 = \sum_{y \in \mathbb{Z} \backslash \{0\}} |y|^r \, \hat{\alpha}_y \hat{\beta}_0,\end{aligned}$$

which has nothing to do with parameter b. $\square$

Page intentionally left blank.

B

Supplementary Details for Chapter 3

B.1 Computing probability mass functions of price changes

Let $a_1, ..., a_n$ be non-zero integers. We will demonstrate how the inverse fast Fourier transform (IFFT) can be used to calculate $p_y = \mathbb{P}\left(Y = y\right)$ of $Y = \sum_{k=1}^{n} a_k X_k \in \mathbb{Z}$, where X_k's are independent Poisson random variables with intensities λ_k.

The characteristic function of Y is:

$$\varphi\left(\theta \ddagger Y\right) = \mathbb{E}\left(e^{\mathrm{i}\theta Y}\right) = \mathbb{E}\left(e^{\sum_{k=1}^{n} \mathrm{i}\theta a_k X_k}\right) = \prod_{k=1}^{n} \varphi\left(\theta a_k \ddagger X_k\right) = \prod_{k=1}^{n} \exp\left(\lambda_k \left(e^{\mathrm{i}\theta a_k} - 1\right)\right).$$

As Y is discrete, the discrete IFFT can be used to get p_y. Note that $\varphi\left(\theta \ddagger Y\right) = \sum_{y=-\infty}^{\infty} e^{\mathrm{i}\theta y} p_y$, so the inverse Fourier transform is justified by, for $y = 0, 1, 2, ...$, as $N \to \infty$,

$$\begin{aligned}
\frac{1}{N}\sum_{k=0}^{N-1}\varphi\left(-\frac{2\pi k}{N}\ddagger Y\right)e^{\mathrm{i}2\pi ky/N} &= \frac{1}{N}\sum_{k=0}^{N-1}\sum_{y'=-\infty}^{\infty} p_{y'}e^{-\mathrm{i}2\pi ky'/N}e^{\mathrm{i}2\pi ky/N} \\
&= \frac{1}{N}\sum_{y'=-\infty}^{\infty} p_{y'}\sum_{k=0}^{N-1} e^{\mathrm{i}2\pi k(y-y')/N} \to \sum_{y'=-\infty}^{\infty} p_{y'}1_{\{y=y'\}} = p_y, \\
\frac{1}{N}\sum_{k=0}^{N-1}\varphi\left(\frac{2\pi k}{N}\ddagger Y\right)e^{\mathrm{i}2\pi ky/N} &= \frac{1}{N}\sum_{y'=-\infty}^{\infty} p_{y'}\sum_{k=0}^{N-1} e^{\mathrm{i}2\pi k(y+y')/N} \to \sum_{y'=-\infty}^{\infty} p_{y'}1_{\{y=-y'\}} = p_{-y},
\end{aligned}$$

where the approximation here comes from the Riemann sum

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{\mathrm{i}2\pi k\theta'/N} = \int_0^1 e^{\mathrm{i}2\pi\theta\theta'}\mathrm{d}\theta + O\left(N^{-1}\right) = 1_{\{\theta'=0\}} + O\left(N^{-1}\right).$$

Hence, the IFFT will take the input of

$$\left(\varphi\left(0 \ddagger Y\right), \varphi\left(-2\pi/N \ddagger Y\right), ..., \varphi\left(-\frac{2\pi\left(N-1\right)}{N}\ddagger Y\right)\right)^T$$

and give the output as $\left(p_0, ..., p_{N-1}\right)^T$ approximately; similarly, with the input of

$$\left(\varphi\left(0 \ddagger Y\right), \varphi\left(2\pi/N \ddagger Y\right), ..., \varphi\left(\frac{2\pi\left(N-1\right)}{N}\ddagger Y\right)\right),$$

the IFFT will give the output as $\left(p_0, p_{-1}, ..., p_{-(N-1)}\right)^T$ approximately.

In Figure 3.10, we take $N = 60$ in order to accurately compute p_y for $y \in \{-30, ..., 30\}$.

B.2 Cleaning of the empirical data

Here we discuss the preprocessing procedures for the raw empirical data. For each data set, our database has the current bid price (`bid`), bid size (`bidsz`), ask price (`ask`), ask size (`asksz`), trade price (`trade`), trade volume (`tradesz`) and the record logging time on the data server (`log_t`). The following events will be logged into the raw data set chronologically:

- A change of `bid` and `bidsz` (or `ask` and `asksz`), which will leave missing `ask`, `asksz` (or `bid`, `bidsz`), `trade` and `tradesz`.

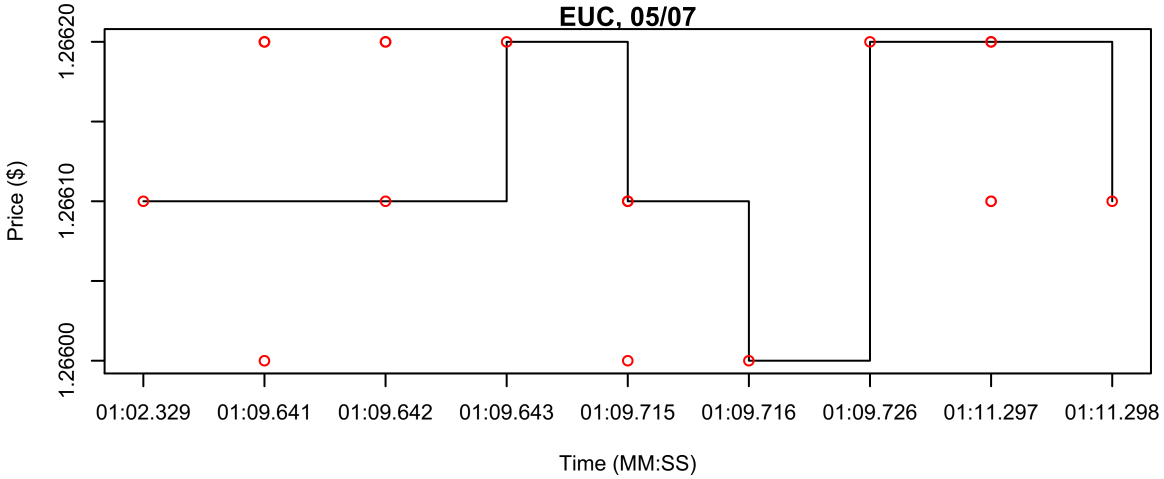


Figure B.1: Illustration of the price definition. The white points illustrate all the trade prices with the same time tag; the black solid line represent the unique price we define. This data is EUC1006 between 00:01:02 and 00:01:12 on May 7.

- A new instance of `trade` and `tradesz`, which will leave missing `bid`, `bidsz`, `ask` and `asksz`. This is usually followed by a record that shows the newest `bid` and `ask` status after the trading. Sometimes this updating record will be combined with its previous trading record.

Step 1: Remove the wrong records (Optional). We forward fill the missing values in columns `bid` and `ask`; after this, we examine whether the recorded trade price lies in the range from `bid` minus a factor `M` of tick sizes to `ask` plus `M` tick sizes. `M` here is manually chosen as 9.5 for the two EUC data sets, which is a conservative setting and will only remove those visually inspectable errors. We do not use this step for the two TNC data sets.

Step 2: Preserve only the trading activities. Since in this paper we are only concerned with the dynamics of the trade prices, we throw out all the other data records that are not directly associated with a trade, that is, those rows with missing `trade` and `tradesz`.

Step 3-1: Associate a unique price to a time tag. Occasionally several data feeds will be pushed into the data server almost at the same time but perhaps with different prices. Then we iteratively define a unique price for this particular time tag by the price that is closest to the price of the previous time tag. Figure B.1 illustrates this.

Step 3-2: Do nothing for an ambiguous case. When there are exactly two trade prices with the same time tag that are just one tick above and one tick below the previous price (e.g. at time 01:09.641 in Figure B.1), then we use the previous price as the price for the current time tag.

Step 4: Keep only jumps. Here it is sufficient to keep only the columns `Time` and `Price`, such that `Time` is always increasing without duplicates while `Price` have no two adjacent elements that take the same value. `Price` is always the value the price process takes immediately after a jump.

B.3 Non-parametric inference of the trawl function

Let $\tilde{d}\,(s)$ be the non-squashed trawl function with $\tilde{d}\,(-\infty) = 0$ such that $d\,(s) = b + (1-b)\,\tilde{d}\,(s)$. Then equation (3.3) implies $\partial_\delta leb\,(A_\delta \backslash A) = (1-b)\,\tilde{d}\,(-\delta)$. Hence,

$$\frac{\partial \widehat{\sigma_\delta^2}}{\partial \delta} = \left(\frac{b + 2\,(1-b)\,\tilde{d}\,(-\delta)}{2-b}\right) \sum_{y \in \mathbb{Z}\backslash\{0\}} y^2 \hat{\alpha}_y \hat{\beta}_0, \quad \frac{\partial \widehat{\sigma_\delta^2}}{\partial \delta}\,(\infty) = \left(\frac{b}{2-b}\right) \sum_{y \in \mathbb{Z}\backslash\{0\}} y^2 \hat{\alpha}_y \hat{\beta}_0,$$

which then gives us

$$b = \frac{2\left(\partial_\delta \widehat{\sigma_\delta^2}\right)(\infty)}{\left(\partial_\delta \widehat{\sigma_\delta^2}\right)(0) + \left(\partial_\delta \widehat{\sigma_\delta^2}\right)(\infty)}, \ \tilde{d}\,(-\delta) = \frac{\partial_\delta \widehat{\sigma_\delta^2} - \left(\partial_\delta \widehat{\sigma_\delta^2}\right)(\infty)}{\left(\partial_\delta \widehat{\sigma_\delta^2}\right)(0) - \left(\partial_\delta \widehat{\sigma_\delta^2}\right)(\infty)}, \ \left(\partial_\delta \widehat{\sigma_\delta^2}\right)(0) = \sum_{y \in \mathbb{Z}\backslash\{0\}} y^2 \hat{\alpha}_y \hat{\beta}_0.$$

Therefore, by estimating $\partial_\delta \widehat{\sigma_\delta^2}$ for every δ the trawl function is revealed non-parametrically.

In practice, it might be demanding to get $\left(\partial_\delta \widehat{\sigma_\delta^2}\right)(\infty)$, the asymptotic slope of the sample variogram $\widehat{\sigma_\delta^2}$ against δ, because as δ being larger, the sample size we use to calculate $\widehat{\sigma_\delta^2}$ is getting smaller. Is it possible to use other moment equations in Theorem 14 to identify b rather than through the boundary behavior of $\partial_\delta \widehat{\sigma_\delta^2}$ for $\delta \to \infty$? Unfortunately, the answer is no. b and d are not identifiable if we neither parameterize d nor adopt a boundary estimation for b at $\delta \to \infty$.

To justify this point, assume that one wants to employ all the other additional moment equations in Theorem 14 to identify b:

$$\begin{aligned}\kappa_j\,(P_\delta - P_0) &= \left(b\delta + \left(1 + (-1)^j\right) leb\,(A_\delta \backslash A)\right) \kappa_j\,(L_1) \\ &= \left(b\delta + \left(1 + (-1)^j\right) leb\,(A_\delta \backslash A)\right) \sum_{y \in \mathbb{Z}\backslash\{0\}} y^j \nu\,(y)\,, \qquad j \geq 3.\end{aligned}$$

Denote the sample j-th cumulant with sampling interval δ as $\widehat{\kappa_{j,\delta}}$. Then equation (3.13) implies that

$$\begin{aligned}
\frac{\partial\widehat{\kappa_{j,\delta}}}{\partial\delta} &= \left(b + \left(1 + (-1)^j\right) \frac{\partial}{\partial\delta} leb\left(A_\delta \backslash A\right)\right) \sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \frac{\hat{\alpha}_y - (1-b)\,\hat{\alpha}_{-y}}{(2-b)\,b} \hat{\beta}_0 \\
&= \left(b + \left(1 + (-1)^j\right)(1-b)\,\tilde{d}\,(-\delta)\right) \frac{\sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \hat{\alpha}_y - (1-b)\sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \hat{\alpha}_{-y}}{(2-b)\,b} \hat{\beta}_0 \\
&= \left(b + \left(1 + (-1)^j\right)(1-b)\,\tilde{d}\,(-\delta)\right) \frac{1 - (-1)^j\,(1-b)}{(2-b)\,b} \sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \hat{\alpha}_y \hat{\beta}_0 \\
&= \begin{cases} \sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \hat{\alpha}_y \hat{\beta}_0 & \text{, for } j \text{ odd} \\ \dfrac{\partial\widehat{\sigma_\delta^2}/\partial\delta}{\sum_{y\in\mathbb{Z}\backslash\{0\}} y^2 \hat{\alpha}_y \hat{\beta}_0} \sum_{y\in\mathbb{Z}\backslash\{0\}} y^j \hat{\alpha}_y \hat{\beta}_0 & \text{, for } j \text{ even} \end{cases},
\end{aligned}$$

which is still again independent of b.

Page intentionally left blank.

References

Airoldi, E. M., T. B. Costa, and S. H. Chan (2013, December). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, pp. 692–700.

Aldous, D. J. (1981, December). Representations for partially exchangeable arrays of random variables. Journal of Multivariate Analysis 11, 581–598.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of exchange rate volatility. Journal of the American Statistical Association 96, 42–55.

Bacry, E., S. Delattre, M. Hoffman, and J. F. Muzy (2013a). Modelling microstructure noise with mutually exciting point processes. Quantitative Finance 13, 65–77.

Bacry, E., S. Delattre, M. Hoffman, and J. F. Muzy (2013b). Some limit theorems for Hawkes processes and application to financial statistic. Stochastic Processes and their Applications 123, 2475–2499.

Ball, C. A. (1988). Estimation bias induced by discrete security pricing. Journal of Finance 43, 841–865.

Ball, C. A., W. N. Torous, and A. E. Tschoegl (1985). The degree of price resolution: The case of the gold market. Journal of Futures Markets 5, 29–43.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. Econometrica 76, 1481–1536.

Barndorff-Nielsen, O. E., A. Lunde, N. Shephard, and A. E. D. Veraart (2014). Integer-valued trawl processes: A class of stationary infinitely divisible processes. Scandinavian Journal of Statistics 41, 693–724.

Barndorff-Nielsen, O. E., D. G. Pollard, and N. Shephard (2012). Integer-valued Lévy processes and low latency financial econometrics. Quantitative Finance 12, 587–605.

Barndorff-Nielsen, O. E. and N. Shephard (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society, Series B 64, 253–280.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Power and bipower variation with stochastic volatility and jumps (with discussion). Journal of Financial Econometrics 2, 1–48.

Bartlett, M. S. (1978). An Introduction to Stochastic Processes, with Special Reference to Methods and Applications (3 ed.). Cambridge: Cambridge University Press.

Berger, J. O. and W. E. Strawderman (1996, 06). Choice of hierarchical priors: Admissibility in estimation of normal means. The Annals of Statistics 24(3), 931–951.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and newman-girvan and other modularities. In Proceedings of the National Academy of Sciences of the United States of America, Volume 106, pp. 21068–21073.

Bickel, P. J., A. Chen, and E. Levina (2011). The method of moments and degree distributions for network models. Annals of Statistics 39(5), 2280–2301.

Borgs, C., J. T. Chayes, and L. Lovász (2010). Moments of two-variable functions and the uniqueness of graph limits. Geometric and Functional Analysis 19, 1597–1619.

Borgs, C., J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi (2008). Convergent sequences of dense graph I: Subgraph frequencies, metric properties and testing. Advances in Mathematics 219, 1801–1851.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. Journal of Econometrics 141, 876–912.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. Annals of Applied Statistics 2(1), 113–152.

Chan, S. H. and E. M. Airoldi (2014). A consistent histogram estimator for exchangeable graph models. Journal of Machine Learning Research 32(1), 208–216.

Chan, S. H., R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen (2011). An augmented lagrangian method for total variation video restoration. IEEE Transactions on Image Processing 20(11), 3097–3111.

Chatterjee, S. Matrix estimation by universal singular value thresholding. ArXiv:1212.1247. 2012.

Chatterjee, S. (2012). Matrix estimation by universal singular value thresholding. ArXiv:1212.1247. Unpublished manuscript.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012). Stochastic blockmodels with a growing number of classes. Biometrika 99, 273–284.

Cipollini, F., R. F. Engle, and G. Gallo (2009). A model for multivariate non-negative valued processes in financial econometrics. Available at SSRN: http://ssrn.com/abstract=1333869 or http://dx.doi.org/10.2139/ssrn.1333869.

Copas, J. B. (1983). Regression, prediction and shrinkage. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 45(3), 311–354.

Delattre, S. and J. Jacod (1997). A central limit theorem for normalized functions of the increments of a diffusion process in the presence of round-off errors. Bernoulli 3, 1–28.

Diaconis, P. and S. Janson (2008). Graph limits and exchangeable random graphs. Rendiconti di Matematica edelle sue Applicazioni, Series VII 28, 33–61.

Efron, B. and C. Morris (1972, August). Empirical Bayes on vector observations: An extension of Stein's method. Biometrika 59(2), 335–347.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. Journal of the American Statistical Association 68(341), 117–130.

Efron, B. and C. Morris (1975, June). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70(350), 311–319.

Engle, R. F. (2000). The econometrics of ultra-high frequency data. Econometrica 68, 1–22.

Engle, R. F. and J. R. Russell (1998). Forecasting transaction rates: The autoregressive conditional duration model. Econometrica 66, 1127–1162.

Fauth, A. and C. A. Tudor (2012). Modeling first line of an order book with multivariate marked point processes. ArXiv e-prints. Unpublished paper, SAMM, Université Paris 1 Panthéon-Sorbonne, November.

Fearn, T. (1975, April). A Bayesian approach to growth curves. Biometrika 62(1), 89–100.

Fodra, P. and H. Pham (2013a). High frequency trading in a Markov renewal model. ArXiv e-prints. Unpublished paper, Laboratoire de Probabilités et, Université Paris 7 Diderot, October.

Fodra, P. and H. Pham (2013b). Semi Markov model for market microstructure. ArXiv e-prints. Unpublished paper, Laboratoire de Probabilités et, Université Paris 7 Diderot, May.

Fuchs, F. and R. Stelzer (2013). Mixing conditions for multivariate infinitely divisible processes with an application to mixed moving averages and the supOU stochastic volatility model. ESAIM: Probability and Statistics 17, 455–471.

Goldenberg, A., A. Zheng, S. Fienberg, and E. Airoldi (2009). A survey of statistical network models. Foundations and Trends in Machine Learning 2, 129–233.

Gottlieb, G. and A. Kalay (1985). Implications of the discreteness of observed stock prices. Journal of Finance 40, 135–153.

Green, E. J. and W. E. Strawderman (1985, December). The use of Bayes/empirical Bayes estimation in individual tree volume equation development. Forest Science 31(4), 975–990.

Griffin, J. E. and R. C. A. Oomen (2008). Sampling returns for realized variance calculations: Tick time or transaction time? Econometric Review 27, 230–253.

Hamilton, J. D. and O. Jordá (2002). A model of the federal funds rate target. Journal of Political Economy 110, 1135–1167.

Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise (with discussion). Journal of Business and Economic Statistics 24, 127–218.

Harris, L. (1990). Estimation of stock price variances and serial covariances from discrete observations. Journal of Financial Quantative Analysis 25, 291–306.

Hasbrouck, J. (1999). The dynamics of discrete bid and ask quotes. Journal of Finance 54, 2109–2142.

Hautsch, N. (2012). Econometrics of Financial High-Frequency Data. Berlin Heidelberg: Springer.

Hawkes, A. G. (1972). Spectra of some mutually exciting point processes with associated variables. In P. A. W. Lewis (Ed.), Stochastic Point Processes, pp. 261–271. New York: Wiley.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ.

Hui, S. L. and J. O. Berger (1983, December). Empirical Bayes estimation of rates in longitudinal studies. Journal of the American Statistical Association 78(384), 753–760.

Jacod, J. (1996). La variation quadratique du Brownian en presence d'erreurs d'arrondi. Asterisque 236, 155–162.

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009). Microstructure noise in the continuous case: The pre-averaging approach. Stochastic Processes and Their Applications 119, 2249–2276.

James, W. and C. M. Stein (1961). Estimation with quadratic loss. Proceedings of 4th Berkeley Symposium on Probability and Statistics I, 367–379.

Jiang, J., T. Nguyen, and J. S. Rao (2011). Best predictive small area estimation. Journal of the American Statistical Association 106(494), 732–745.

Jones, K. (1991). Specifying and estimating multi-level models for geographical research. Transactions of the Institute of British Geographers 16(2), 148–159.

Kallenberg, O. (1989). On the representation theorem for exchangeable arrays. Journal of Multivariate Analysis 30(1), 137–154.

Kerss, A., N. Leonenko, and A. Sikorskii (2014). Fractional Skellam processes with applications to finance. Fractional Calculus and Applied Analysis 17, 532–551.

Kingman, J. F. C. (1993). Poisson Processes. New York: Oxford University Press.

Kolassa, J. and P. McCullagh (1990). Edgeworth series for lattice distributions. Annals of Statistics 18, 981–985.

Large, J. (2011). Estimating quadratic variation when quoted prices jump by a constant increment. Journal of Econometrics 160, 2–11.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Annals of Statistics 14(3), 1101–1102.

Li, Y. and P. A. Mykland (2014). Rounding errors and volatility estimation. Journal of Financial Econometrics.

Liesenfeld, R., I. Nolte, and W. Pohmeier (2006). Modelling financial transaction price movements: a dynamic integer count model. Empirical Economics 30, 795–825.

Lindley, D. V. (1956). The estimation of velocity distributions from counts. In Proceedings of the Internationl Congress of Mathematicians, Volume 3, pp. 427–444. Amsterdam: North-Holland.

Lindley, D. V. (1962). Discussion of a paper by C. Stein. Journal of the Royal Statistical Society. Series B (Methodological) 24, 285–287.

Lindley, D. V. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society. Series B (Methodological) 34(1), 1–41.

Morris, C. N. (1983, March). Parametric empirical Bayes inference: Theory and applications. Journal of the American Statistical Association 78(381), 47–55.

Morris, C. N. and M. Lysy (2012). Shrinkage estimation in multilevel normal models. Statistical Science 27(1), 115–134.

Mykland, P. A. and L. Zhang (2012). The econometrics of high frequency data. In M. Kessler, A. Lindner, and M. Sørensen (Eds.), Statistical Methods for Stochastic Differential Equations, pp. 109–190. New York: Chapman & Hall/CRC Press. Forthcoming.

Normand, S.-L. T., M. E. Glickman, and C. A. Gatsonis (1997, September). Statistical methods for profiling providers of medical care: Issues and applications. Journal of the American Statistical Association 92(439), 803–814.

Olding, B. P. and P. J. Wolfe (2009, June). Inference for graphs and networks: Extending classical tools to modern data. ArXiv:0906.4980. Unpublished manuscript.

Omen, S. D. (1982, November). Shrinking towards subspaces in multiple linear regression. Technometrics 24(4), 307–311. 1982.

Oomen, R. C. A. (2005). Properties of bias-corrected realized variance under alternative sampling schemes. Journal of Financial Econometrics 3, 555–577.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. Journal of Business & Economic Statistics 24, 219–237.

Press, S. J. (1967). A compound events model for security prices. Journal of Business 40, 317–335.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997, March). Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92(437), 179–191.

Reynolds, J. F. (1968). On the autocorrelation and spectral functions of queues. Journal of Applied Probability 5, 467–475.

Robbins, H. (1956). An empirical Bayes approach to statistics. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, pp. 157–163. University of California Press.

Rosenbaum, M. (2009). Integrated volatility and round-off error. Bernoulli 15, 687–720.

Rubin, D. B. (1980, December). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association 75(372), 801–816.

Russell, J. R. (1999). Econometric modeling of multivariate irregularly-spaced high-frequency data. Unpublished paper, Booth School of Business, University of Chicago.

Russell, J. R. and R. F. Engle (2006). A discrete-state continuous-time model of financial transaction prices and times. Journal of Business and Economic Statistics 23, 166–180.

Russell, J. R. and R. F. Engle (2010). Analysis of high-frequency data. In Y. Ait-Sahalia and L. P. Hansen (Eds.), Handbook of Financial Econometrics: Tools and techniques, pp. 383–426. Amsterdam: North-Holland.

Rydberg, T. H. and N. Shephard (2003). Dynamics of trade-by-trade price movements: Decomposition and models. Journal of Financial Econometrics 1, 2–25.

Sclove, S. L., C. Morris, and R. Radhakrishnan (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. The Annals of Mathematical Statistics 43(5), 1481–1490.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, pp. 197–206. University of California Press.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In F. J. Neyman (Ed.), Research Papers in Statistics, pp. 351–366. London: Wiley.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 24, 265–296.

Strenio, J. F., H. I. Weisberg, and A. S. Bryk (1983, March). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. Biometrics 39(1), 71–86.

Surgailis, D., J. Rosinski, V. Mandrekar, and S. Cambanis (1993). Stable mixed moving averages. Probability Theory and Related Fields 97, 543–558.

Tan, Z. (2015). Steinized empirical Bayes estimation for heteroscedastic data. Statistica Sinica. Forthcoming.

Wolfe, P. J. and S. C. Olhede (2013, September). Nonparametric graphon estimation. ArXiv:1309.5936. Unpublished manuscript.

Wolpert, R. L. and L. D. Brown (2011). Stationary infinitely-divisible Markov processes with non-negative integer values. Working paper, Department of Staistics, Duke University.

Wolpert, R. L. and M. S. Taqqu (2005). Fractional Ornstein-Uhlenbeck Lévy processes and the telecom process: Upstairs and downstairs. Signal Processing 85, 1523–1545.

Xie, X., S. C. Kou, and L. D. Brown (2012, December). SURE estimates for a heteroscedastic hierarchical model. Journal of the American Statistical Association 107(500), 1465–1479.

Xie, X., S. C. Kou, and L. D. Brown (2015). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. Annals of Statistics. Forthcoming.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multiscale approach. Bernoulli 12, 1019–1043.

Zhang, M. Y., J. R. Russell, and R. Tsay (2001). Determinants of bid and ask quotes and implications for the cost of trading. Journal of Empirical Finance 15, 656–678.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. Journal of Business and Economic Statistics 14, 45–52.