KERNEL ESTIMATION OF THE DRIFT COEFFICIENT OF A DIFFUSION PROCESS IN THE PRESENCE OF MEASUREMENT ERROR

by

WOOYONG LEE

B.Econ., Korea University, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

June 2014

Abstract

Diffusion processes, a class of continuous-time stochastic processes, can be used to model time-series data observed at discrete time points. A diffusion process can be completely characterized by two functions, called the drift coefficient and the diffusion coefficient. For the nonparametric estimation of these two functions, Bandi and Phillips (2003) proved consistency and asymptotic normality of Nadaraya-Watson kernel estimators of the drift and the diffusion coefficient.

In some cases, we observe the time-series data with measurement error. For instance, it is a well-known fact that we observe the financial time-series data with measurement errors (Zhou, 1996). For the nonparametric estimation of the drift and the diffusion coefficients in the presence of measurement error, some works are done for the estimation of integrated volatility, which is the integral of the diffusion coefficient over a fixed period of time, but little work exists on the estimation of the drift and the diffusion coefficients themselves. In this thesis, we focus on the estimation of the drift coefficient, and we propose a consistent and asymptotically normal Nadaraya-Watson type kernel estimator of the drift coefficient in the presence of measurement error.

Preface

This thesis is an original and unpublished work of the author, Wooyong Lee, under the supervision of Dr. Nancy Heckman and Dr. Priscilla Greenwood.

The research question and the estimator are established earlier by Nancy Heckman, Priscilla Greenwood and Dr. Wolfgang Wefelmeyer. Based on that, I have written a full proof for consistency and asymptotic normality of the estimator, proposed a criterion for choosing the appropriate bandwidths and performed a simulation study of the estimator.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I express my deep gratitude to my supervisors, Dr. Nancy Heckman and Dr. Priscilla Greenwood, and my second reader, Dr. Alexandre Bouchard-Côté. Nancy and Cindy happily invested a lot of their time for my thesis and for my academic training. Their help led to great academic improvement of myself during my stay at UBC. Especially, I learned from them how to write an academic paper, without which I can never become a good researcher. Alex taught me three courses, two of which were core courses in graduate level statistics. Taking Alex's courses were of great fun, and what I learned from him were extremely useful in reading technical papers cited in this thesis.

I also thank all the other faculty members and fellow graduate students in the department for introducing to me a lot of interesting fields in statistics in their classes and talks. It was one of the greatest decision that I have made to join the statistics department at UBC.

To my family

Chapter 1

Introduction

A continuous-time stochastic process can be used to model time-series data that are observed at discrete time points. For example, Felsenstein (1985) uses Brownian motion to model evolutionary history of species, and Andersen et al. (2001) use a continuous-time semimartingale to model variability of exchange rates.

In this thesis, we focus on a specific type of continuous-time stochastic process, namely, a diffusion process. A diffusion process is a solution to a stochastic differential equation, and it is used to describe many kinds of time-series data such as price data of financial instruments (see e.g. Aït-Sahalia, 1996). A stochastic differential equation has the form

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t,$$

where the function μ and the nonnegative function σ are two deterministic functions from $[0, \infty) \times \mathbb{R}$ to $\mathbb{R}$, called the drift coefficient and the diffusion coefficient respectively, and W_t is a stochastic process called Brownian motion. A solution to this stochastic differential equation with an initial value random variable Y is a stochastic process $\{X_t \mid t \geq 0\}$ satisfying $X_0 = Y$ and

$$X_t = Y + \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s, \quad t \geq 0.$$

The solution must satisfy additional conditions, introduced later in Definition 1.6. The integral $\int_0^t \sigma(s, X_s)dW_s$ is an example of what is called stochastic integration, which we define later in

Definition 1.5.

As a solution to a stochastic differential equation, a diffusion process can be completely characterized by the drift coefficient μ and the diffusion coefficient σ. In addition, μ determines the expected value of the (random) change in X_t over an infinitesimal amount of time, and σ determines the variance of the (random) change in X_t over an infinitesimal amount of time. Therefore, the statistical goal when using a diffusion model is to estimate these two functions.

The theme of this thesis is to propose a statistical method to estimate the drift coefficient nonparametrically in the presence of measurement error, in which case the discrete-time observations do not provide exact values of the latent continuous-time process. In addition, we consider the simpler form of the stochastic differential equation, that μ and σ are not functions of t and X_t but functions of X_t only, in which case the stochastic differential equation is said to be time-homogeneous:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t. \tag{1.1}$$

In this chapter, we introduce background knowledge used to formally define our research question and our estimator. We first introduce Brownian motion, which is used to define the stochastic integrals considered here. Then we use stochastic integration to construct a stochastic differential equation and its solution, which is called a diffusion process. After that, we discuss stationarity and ergodicity for diffusion processes, the properties we assume in our study in later chapters. Lastly, we discuss kernel estimation, a nonparametric estimation method we use in order to estimate the drift coefficient μ. The review of existing literature and the statement of our research question in relation to the literature will be given in Section 2.1.

For discussing relevant background knowledge related to stochastic processes, we use Karatzas and Shreve (1991), Øksendal (1992) and Kutoyants (2004). Øksendal (1992) is a textbook on stochastic integration and stochastic differential equations intended for graduate students and non-experts while Karatzas and Shreve (1991) offer a more abstract and rigorous treatment of these areas. Kutoyants (2004) studies statistical problems for stationary and ergodic diffusion processes. For an introduction to kernel estimation, we use Simonoff (1996) and Hardle (1990) and the references therein, which give an overview for graduate students and applied statisticians while giving further references for more advanced treatment of the

subject.

1.1 Brownian Motion

We introduce Brownian motion first. The sequence of subsets $\{\mathcal{F}_t \mid t \geq 0\}$ of a σ-algebra $\mathcal{F}$ will denote a filtration with the time-index t. All random processes are assumed to be defined on the same probability space, when required.

Definition 1.1 (Karatzas and Shreve, 1991, page 47) *Defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a stochastic process is said to be Brownian motion if it is a continuous $\mathcal{F}_t$-adapted stochastic process $\{W_t, \mathcal{F}_t \,;\, t \geq 0\}$ such that*

1. *$W_0 = 0$ almost surely, and*
2. *For any $0 \leq s < t$, the increment $W_t - W_s$ is independent of $\mathcal{F}_s$ and is normally distributed with mean 0 and variance $t - s$.*

Brownian motion is also called the Wiener process, which is why it is usually denoted with a W. From condition 2 of the above, for discrete times $t_1 < t_2 < \cdots < t_n$, the Brownian motion increments $\{W_{t_2} - W_{t_1}, W_{t_3} - W_{t_2}, \ldots, W_{t_n} - W_{t_{n-1}}\}$ are independent and normally distributed random variables.

1.2 Stochastic Integration

Stochastic integration is an integration with respect to a stochastic process, in contrast to Lebesgue integration which is an integration with respect to a measure. In this section, we restrict our discussion to stochastic integration with respect to Brownian motion, although stochastic integration is defined for more general classes of stochastic processes including martingales.

To understand the definition of stochastic integration, recall that the Riemann integral of an integrable function can be characterized by the limit of integrals of step functions which converge to the integrable function. The stochastic integral is defined in a similar way: we first define the stochastic integral of a simple process, which is similar to a step function, and

then we define the stochastic integral of a stochastic process as a limit of the stochastic integrals of simple processes that converge to the process in a suitable norm.

Now we introduce this construction formally. For ease of exposition, we only consider stochastic integration over the time-interval $[0, T]$. Its extension to a generic time-interval $[S, T]$ is straightforward. We first define a simple process and then define the stochastic integral of a simple process.

Definition 1.2 (Karatzas and Shreve, 1991, page 132) *Defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a stochastic process $\{S_t \mid t \geq 0\}$ is said to be a simple process if there exists a strictly increasing sequence of real numbers $\{t_i\}_{i=0}^{\infty}$ with $t_0 = 0$ and $t_n \to \infty$ as well as a sequence of random variables $\{\xi_i\}_{i=0}^{\infty}$ such that*

$$S_t = \xi_k \quad for \quad t_k \leq t < t_{k+1},$$

where $\sup_{n \geq 0} |\xi_n(\omega)| \leq C < \infty$ for every $\omega \in \Omega$ and ξ_n is $\mathcal{F}_{t_n}$-measurable for every $n \geq 0$.

Definition 1.3 (Karatzas and Shreve, 1991, page 132) *The stochastic integral of the simple process $\{S_t \mid t \geq 0\}$ with respect to Brownian motion $\{W_t \mid t \geq 0\}$ over $[0, T]$ is defined as*

$$\int_0^T S_t \, dW_t \equiv \sum_{k=0}^{N-1} \xi_k \left(W_{t_{k+1}} - W_{t_k} \right) + \xi_N \left(W_T - W_{t_N} \right),$$

where N is an integer such that $t_N \leq T < t_{N+1}$.

Note that the ξ's and $\{W_t\}$ need not be independent and that the stochastic integral is a random variable. Having defined stochastic integrals for simple processes, the next step is to define a class of stochastic processes that are well approximated by simple processes. We introduce the following class of stochastic processes, $\mathcal{L}^2([0, T])$, which is conceptually similar to the L^2 space of random variables.

Definition 1.4 (Øksendal, 1992, page 18) *$\mathcal{L}^2([0, T])$ is defined as the class of $\mathcal{F}_t$-adapted stochastic processes $\{V_t, \mathcal{F}_t \, ; \, t \geq 0, \; V_t \text{ defined on } (\Omega, \mathcal{F}, \mathbb{P})\}$ such that*

$$\mathbb{E}\left(\int_0^T V_t^2 \, dt \right) < \infty.$$

The following result states that every stochastic process in $\mathcal{L}^2([0,T])$ can be approximated by a simple process, in an "$\mathcal{L}^2$ norm".

Lemma 1.1 (Øksendal, 1992, page 19) *For $\{V_t \mid t \geq 0\} \in \mathcal{L}^2([0,T])$, there exists a sequence of simple processes $\left\{ \{S_t^{(n)} \mid t \geq 0\} \middle| n = 1, 2, \ldots \right\} \subseteq \mathcal{L}^2([0,T])$ such that*

$$\lim_{n\to\infty} \mathbb{E}\left(\int_0^T \left(V_t - S_t^{(n)} \right)^2 dt \right) = 0.$$

Then we define the stochastic integral of a stochastic process in $\mathcal{L}^2([0,T])$ as a limit of the stochastic integrals of the simple processes that converge to the process, as follows.

Definition 1.5 (Øksendal, 1992, page 21) *The stochastic integral of $\{V_t \mid t \geq 0\} \in \mathcal{L}^2([0,T])$ with respect to Brownian motion $\{W_t \mid t \geq 0\}$ over $[0,T]$ is defined as*

$$\int_0^T V_t \, dW_t \equiv \lim_{n\to\infty} \int_0^T S_t^{(n)} dW_t,$$

where the limit is the almost sure limit and $\left\{ \{S_t^{(n)} \mid t \geq 0\} \middle| n = 1, 2, \ldots \right\} \subseteq \mathcal{L}^2([0,T])$ is a sequence of simple processes as in Lemma 1.1.

One can show that this stochastic integral is well-defined, that is, that the limit is independent of the choice of $S_t^{(n)}$'s (Øksendal, 1992, page 21). In addition, the limiting random variable has a finite second moment (Øksendal, 1992, page 21).

We conclude this section by stating some basic properties of stochastic integration. These properties are obvious when the stochastic process in the integrand is simple. For generic stochastic processes in $\mathcal{L}^2([0,T])$, the properties can be proven via the limit argument.

Lemma 1.2 (Øksendal, 1992, page 22) *For real $S \leq R \leq T$ and α and β, stochastic processes $\{V_t \mid t \geq 0\}$ and $\{U_t \mid t \geq 0\}$ in $\mathcal{L}^2([0,T])$ and Brownian motion $\{W_t \mid t \geq 0\}$, the following*

hold almost surely.

$$\begin{aligned}
\int_S^T V_t\, dW_t &= \int_S^R V_t\, dW_t + \int_R^T V_t\, dW_t, \\
\int_S^T (\alpha V_t + \beta U_t)\, dW_t &= \alpha \int_S^T V_t\, dW_t + \beta \int_S^T U_t\, dW_t, \\
\mathbb{E}\left(\int_S^T V_t\, dW_t\right) &= 0, \\
\mathbb{E}\left(\left[\int_S^T V_t\, dW_t\right]^2\right) &= \mathbb{E}\left(\int_S^T V_t^2\, dt\right).
\end{aligned}$$

1.3 Stochastic Differential Equation

In this section, we formally introduce stochastic differential equations and their solutions, which are called diffusion processes. Then, in the next section, we define stationarity and ergodicity for diffusion processes, the properties we assume in our study in later chapters. As stated earlier, we consider the time-homogeneous stochastic differential equation given in (1.1).

Let $\{W_t \mid t \geq 0\}$ be Brownian motion defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and Y be a real-valued random variable also defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and independent of Brownian motion. We define an augmented filtration $\mathcal{F}_t$, $t \geq 0$, based on the filtration

$$\mathcal{G}_t \equiv \sigma\big(Y\,,\ \{W_s \mid 0 \leq s \leq t\}\big), \quad t \geq 0,$$

and the collection of all subsets of measure zero:

$$\mathcal{N} \equiv \Big\{N \subseteq \Omega \;\Big|\; \exists G \in \bigcup_{t \geq 0} \mathcal{G}_t \text{ with } N \subseteq G \text{ and } \mathbb{P}(G) = 0\Big\}.$$

Then an augmented filtration $\mathcal{F}_t$, $t \geq 0$, is defined by

$$\mathcal{F}_t \equiv \sigma\big(\mathcal{G}_t \cup \mathcal{N}\big), \quad t \geq 0.$$

With this notation, we define a (strong) solution to a stochastic differential equation as follows.

Definition 1.6 (Karatzas and Shreve, 1991, page 285) *A strong solution* $\{X_t \mid t \geq 0\}$ *to the stochastic differential equation in (1.1) on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *with respect to Brownian motion* $\{W_t \mid t \geq 0\}$ *and the initial value random variable* Y *is defined as a stochastic process with continuous sample paths such that*

1. X_t *is adapted to the augmented filtration* $\mathcal{F}_t$.
2. $\mathbb{P}(X_0 = Y) = 1$.
3. $\mathbb{P}\left(\int_0^t \left(|\mu(X_s)| + \sigma^2(X_s)\right) ds < \infty\right) = 1$ *for every* $0 \leq t < \infty$.
4. *The following holds almost surely for all* $t \in [0, \infty)$*:*

$$X_t = X_0 + \int_0^t \mu(X_s) ds + \int_0^t \sigma(X_s) dW_s.$$

In contrast, a weak solution is a stochastic process that has the same distribution as a strong solution but that is not necessarily adapted to the augmented filtration $\mathcal{F}_t$, that is, not necessarily a function of $\{W_t\}$ and Y both defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

As in the study of ordinary differential equations, we are interested in existence and uniqueness conditions for solutions of stochastic differential equations. First, the following defines the uniqueness of a strong solution.

Definition 1.7 (Karatzas and Shreve, 1991, page 286) *We say strong uniqueness holds for the pair* (μ, σ) *if, when* $\{X_t \mid t \geq 0\}$ *and* $\{Y_t \mid t \geq 0\}$ *are both strong solutions to the stochastic differential equation in (1.1) with the initial value random variable* Z*, we have* $\mathbb{P}(X_t = Y_t \,;\, t \geq 0) = 1$.

One well-known condition that ensures existence of the unique strong solution is the following.

Theorem 1.1 (Øksendal, 1992, page 48) *Suppose that the initial value random variable* Y *is independent of Brownian motion* $\{W_t \mid t \geq 0\}$ *and satisfies* $\mathbb{E}(Y^2) < \infty$*. Also suppose that, for every* $x, y \in \mathbb{R}$*, there exist constants* C *and* D *such that*

$$\begin{aligned} |\mu(x) - \mu(y)| \; + \; |\sigma(x) - \sigma(y)| \;&\leq\; C|x - y|, \quad \textit{and} \\ |\mu(x)| \; + \; |\sigma(y)| \;&\leq\; D(1 + |x|). \end{aligned}$$

Then the stochastic differential equation in (1.1) has a unique strong solution.

There are other conditions that give the existence and uniqueness of a strong solution. For example, as Bandi and Phillips (2003, page 244) point out, if μ and σ are twice continuously differentiable and if $\sigma^2(x) > 0$ for all x, then a unique strong solution exists by the following theorems in Karatzas and Shreve (1991): Theorem 2.5 (page 187), Theorem 5.15 (page 341) and Corollary 3.23 (page 310).

1.4 Stationarity and Ergodicity

In this section, we define stationarity and ergodicity of diffusion processes, which we assume in later chapters. In order to define them, we first define recurrence, positive recurrence and null recurrence, which are defined not only for a diffusion process but also for a generic real-valued stochastic process.

Definition 1.8 (Kutoyants, 2004, page 39) *Let $\{V_t\}$ be a real-valued stochastic process, and let $\tau_a \equiv \inf_{t\geq 0}\{V_t = a\}$ and $\tau_a^b \equiv \inf_{t\geq\tau_a}\{V_t = b\}$. We define $\inf\phi \equiv \infty$.*

1. *The process $\{V_t\}$ is said to be recurrent if $\mathbb{P}(\tau_a^b < \infty) = 1$ for all $a, b \in \mathbb{R}$.*
2. *The process $\{V_t\}$ is said to be positive recurrent if it is recurrent and $\mathbb{E}(\tau_a^b) < \infty$ for all $a, b \in \mathbb{R}$.*
3. *The process $\{V_t\}$ is said to be null recurrent if it is recurrent and $\mathbb{E}(\tau_a^b) = \infty$ for all $a, b \in \mathbb{R}$.*

When it comes to a strong solution of a time-homogeneous stochastic differential equation given in (1.1), there are conditions on μ and σ that are related to recurrence, positive recurrence and null recurrence of the corresponding strong solution. Below we give a necessary and sufficient condition on μ and σ for a strong solution to be recurrent or positive recurrent. Note that we only have a sufficient condition (but not a necessary condition) for the null recurrence.

Lemma 1.3 (Kutoyants, 2004, page 40) *A strong solution $\{X_t \mid t \geq 0\}$ to the time-homogeneous stochastic differential equation in (1.1) is recurrent if and only if*

$$S(x) \equiv \int_0^x \exp\left\{-2\int_0^y \frac{\mu(z)}{\sigma^2(z)}dz\right\}dy$$

satisfies $\lim_{x\to-\infty} S(x) = -\infty$ *and* $\lim_{x\to\infty} S(x) = \infty$.

In addition, $\{X_t \mid t \geq 0\}$ *is positive recurrent if and only if it additionally satisfies*

$$G \equiv \int_{-\infty}^{\infty} \frac{1}{\sigma^2(y)} \exp\left\{2\int_0^y \frac{\mu(z)}{\sigma^2(z)}dz\right\} dy < \infty.$$

Also, the solution process is null recurrent if it is recurrent and $G = \infty$.

A positive recurrent strong solution $\{X_t\}$ has the following properties. First, there exists a random variable X whose probability density function is f_X, called the invariant density, such that $X_t \xrightarrow{d} X$ as $t \to \infty$. In addition, a positive recurrent strong solution $\{X_t\}$ is ergodic, that is, for any measurable function h such that $\int |h(x)| f_X(x)dx < \infty$, we have, almost surely,

$$\frac{1}{T}\int_0^T h(X_s)ds \longrightarrow \int h(x)f_X(x)dx \quad \text{as } T \to \infty.$$

The following theorem summarizes this discussion and gives the analytical form of the invariant density f_X for a positive recurrent strong solution $\{X_t\}$.

Theorem 1.2 (Kutoyants, 2004, page 40) *If a strong solution* $\{X_t \mid t \geq 0\}$ *to the time-homogeneous stochastic differential equation in (1.1) is positive recurrent, then* $\{X_t \mid t \geq 0\}$ *is ergodic with the invariant density*

$$f_X(x) = \frac{1}{G\sigma^2(x)} \exp\left\{2\int_0^x \frac{\mu(y)}{\sigma^2(y)}dy\right\},$$

where G is as in Lemma 1.3.

Now we discuss stationarity. We first define stationarity for a generic stochastic process. A (strictly) stationary process is a stochastic process whose joint probability distributions are invariant under the shift of the time-indices, as defined below.

Definition 1.9 (Karatzas and Shreve, 1991, page 103) *A stochastic process* $\{V_t\}$ *is said to be strictly stationary if, for any* $n \in \mathbb{N}$, *any time-indices* $t_1, \ldots, t_n$ *and any* $s \in \mathbb{R}$,

$$(V_{t_1}, \ldots, V_{t_n}) \stackrel{d}{=} (V_{t_1+s}, \ldots, V_{t_n+s}),$$

where the symbol "$\stackrel{d}{=}$*" means both sides have the same distribution.*

Now we relate stationarity to a diffusion process. If a strong solution to (1.1), $\{X_t\}$, is positive recurrent (so that the invariant density f_X, defined in Theorem 1.2, exists) and the initial value random variable Y has the density function equal to f_X, then the strong solution $\{X_t\}$ is strictly stationary (Kutoyants, 2004, page 2). For this reason, the invariant density f_X is also called the stationary density.

In the next section, we discuss kernel estimation, the last background information we need to provide in order to formally define our research question and our estimator.

1.5 Kernel Estimation

We will introduce what is called the Nadaraya-Watson estimator as our estimator for the drift coefficient. The Nadaraya-Watson estimator is the first widely used kernel estimator for cross-sectional data (Nadaraya, 1964, and Watson, 1964). Suppose that we have independent bivariate data, $(x_1, y_1), \ldots, (x_n, y_n)$, from the distribution of (X, Y) from the regression model

$$Y = m(X) + \varepsilon \tag{1.2}$$

where m is a function and ε is a random variable such that $\mathbb{E}(\varepsilon|X = x) = 0$ and $\text{Var}(\varepsilon|X = x) = \sigma^2(x)$. Therefore, the function m represents the conditional expectation of Y given X. The Nadaraya-Watson estimator estimates $m(x) = \mathbb{E}(Y|X = x)$ for each fixed x. In this section, we introduce the estimator using the overview of Simonoff (1996, Chapter 5) and the references therein.

Note that the conditional expectation $\mathbb{E}(Y|X = x)$ is given by

$$\mathbb{E}(Y|X = x) = \int y f_{Y|X=x}(y)dy = \int y \frac{f_{X,Y}(x, y)}{f_X(x)}dy, \tag{1.3}$$

where $f_{Y|X=x}$, $f_{X,Y}$ and f_X are conditional, joint and marginal densities, respectively. We can obtain the Nadaraya-Watson estimator if we substitute for $f_X(x)$ and $f_{X,Y}(x, y)$ in (1.3) with the kernel density estimates, which we define in what follows.

We first define the kernel density estimate of $f_X(x)$. Note that we have observed independent and identically distributed data, $x_1, \ldots, x_n$, where $x_i \in \mathbb{R}$ for each i, having a common

density f_X. The kernel density estimate $\hat{f}_n(x)$ of $f_X(x)$ is defined as

$$\hat{f}_n(x) \equiv \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{1.4}$$

where $K : \mathbb{R} \to \mathbb{R}$ is called the kernel function and h is a positive constant called the bandwidth. Both K and h are chosen by the user. Parzen (1962) proved that $\hat{f}_n(x)$ is a consistent estimator of $f_X(x)$ in the L^2 norm if f_X is continuous at x. To emphasize that we choose h according to n but choose K independent of n, we will sometimes write $h = h_n$.

Theorem 1.3 (Parzen, 1962, page 1069) *Suppose that the kernel $K : \mathbb{R} \to \mathbb{R}$ is a bounded Borel measurable function such that*

$$\lim_{z\to\infty} |zK(z)| = 0, \quad \int_{-\infty}^{\infty} |K(y)|dy < \infty \quad \textit{and} \quad \int_{-\infty}^{\infty} K(y)dy = 1.$$

If $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then

$$\lim_{n\to\infty} \mathbb{E}\left(\left(\hat{f}_n(x) - f_X(x)\right)^2\right) = 0$$

for every x at which f_X is continuous.

Parzen (1962) gave the order of the asymptotic variance of $\hat{f}_n(x)$, but not that of the asymptotic bias. Rosenblatt (1956) derived orders of the asymptotic bias and the variance for nonnegative K and twice differentiable f_X. In addition, he found that using a symmetric K makes the bias converge to 0 in a higher order, which is why people often use symmetric kernels.

We can generalize the univariate kernel density estimate to a multivariate density estimate. Here we introduce a special case of the bivariate kernel density estimate used to derive the Nadaraya-Watson estimator. The product kernel density estimate of the joint density from the independent and identically distributed data, $(x_1, y_1), \ldots, (x_n, y_n)$, where $(x_i, y_i) \in \mathbb{R}^2$ for each i, is defined as

$$\hat{f}_n(x, y) \equiv \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x_i - x}{h_x}\right) K_y\left(\frac{y_i - y}{h_y}\right), \tag{1.5}$$

where K_x and K_y are kernels and h_x and h_y are bandwidths. Discussion of a more general form of the multivariate kernel density estimate can be found in Simonoff (1996, Chapter 4).

Now we define the Nadaraya-Watson estimator following the derivation of Simonoff (1996, page 134), which is a simplified version of the derivation of Watson (1964). Recall the expression (1.3) of the conditional expectation $\mathbb{E}(Y|X = x)$. If we substitute for $f_X(x)$ and $f_{X,Y}(x, y)$ in (1.3) with the kernel density estimates (1.4) and (1.5), set h_x in (1.5) to be equal to h in (1.4) and choose K_y so that $\int K_y(z)dz = 1$ and that $\int zK_y(z)dz = 0$ (for instance, if K_y is symmetric about zero), we derive the Nadaraya-Watson estimator:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}. \tag{1.6}$$

We note that Watson (1964) provided the estimator for the case of $x_i \in \mathbb{R}^d$ where $d \in \mathbb{N}$, in which the kernel K is appropriately defined according to the value of d.

Nadaraya (1964) proved consistency of $\hat{m}(x)$ when Y is bounded. Among many other results of the consistency of $\hat{m}(x)$ for unbounded Y, we state the following.

Theorem 1.4 (Hardle, 1990, Proposition 3.1.1) *Suppose that the following three conditions hold:*

1. *the regression model (X, Y) satisfies $f_X(x) > 0$ and $\mathbb{E}(Y^2) < \infty$,*
2. *the kernel K satisfies $\int |K(u)|du < \infty$ and $\lim_{|u|\to\infty} uK(u) = 0$,*
3. *the sequence of bandwidths $\{h_n\}$ satisfies $h_n \to 0$ and $nh_n \to \infty$.*

Then $\hat{m}(x) \xrightarrow{p} m(x)$ as $n \to \infty$ for every x at which all of $m(\cdot)$, $f_X(\cdot)$ and $\sigma^2(\cdot)$ are continuous.

We note that the use of the Nadaraya-Watson estimator is not restricted to model (1.2). For example, Hall and Hart (1990) studied estimating $\mathbb{E}(Y|X = x)$ by the Nadaraya-Watson estimator when ε_i's in the data are not independent, but rather a stationary process indexed by i. Robinson (1983) considered time-series data, $z_1, \ldots, z_n$, from a discrete-time stochastic process $\{Z_i\}_{i=1}^n$. He studied estimating $\mathbb{E}(Z_{i+p} \mid Z_i, \ldots, Z_{i+p-1})$ for some $p \in \mathbb{N}$ by the Nadaraya-Watson estimator, setting $y_i = z_{i+p}$ and $x_i = (z_i, \ldots, z_{i+p-1})$. Researchers also studied estimating statistical objects in continuous-time models, including diffusion processes, by the Nadaraya-Watson estimator. We refer to studies that used the Nadaraya-Watson estimator for estimation of the drift and the diffusion coefficients of a diffusion process in Section 2.1.

The Nadaraya-Watson estimator $\hat{m}(x)$, defined in (1.6), can be generalized to what is called the local polynomial estimator. Note that, for a fixed x, the estimator $\hat{m}(x)$ is the solution to the following weighted least square problem:

$$\hat{m}(x) = \underset{z}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - z)^2 K\left(\frac{x_i - x}{h_n}\right).$$

Generalizing this, the local polynomial estimator of degree $p \geq 0$ is defined (Stone, 1977, and Cleveland, 1979), for each x, as

$$\hat{m}_{LP}(x) \equiv \hat{\beta}_0^x + \hat{\beta}_1^x (x - x_i) + \ldots + \hat{\beta}_p^x (x - x_i)^p$$

where

$$(\beta_0^x, \ldots, \beta_p^x) = \underset{\beta_0, \ldots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 (x - x_i)^2 - \ldots - \beta_p (x - x_i)^p\right)^2 K\left(\frac{x_i - x}{h_n}\right).$$

If $p = 0$, then $\hat{m}_{LP}$ equals $\hat{m}$. Stone (1977) proved that $\hat{m}_{LP}(x)$ is consistent when $p = 1$. Ruppert and Wand (1994) derived the asymptotic bias and variance of $\hat{m}_{LP}(x)$ for $p \geq 1$. Their key assumptions are that m is $(p+2)$-times differentiable at x with continuous $(p+2)^{nd}$ derivative, that x is an interior point of the support of f_X, that f_X is continuous at x and that the kernel K has compact support (Ruppert and Wand, 1994, Theorem 4.1). In addition, they also derived the asymptotic bias and variance of the multivariate generalization of $\hat{m}_{LP}(x)$ when $x \in \mathbb{R}^k$ and $p = 1, 2$, under similar key assumptions (Ruppert and Wand, 1994, Theorem 3.2).

Lastly, we discuss the choice of the bandwidth h. From Theorem 1.4, we can see that many sequences of h_n's satisfy the conditions of Theorem 1.4. Therefore, given the sample size n, we have great freedom in choosing h_n. But this choice is important: as Simonoff (1996) writes (page 151), the shapes of the function estimates $\hat{m}$ and $\hat{m}_{LP}$ are strongly dependent on h. Larger h leads to a function estimate that is close to the least squares degree p polynomial. Therefore, we require a finite sample method of choosing h.

We will discuss the bandwidth choice criteria in detail in Section 2.4, so we don't give the details here. To summarize that section, the goal of the choice of h is to minimize the mean squared error of the estimator. We consider the bandwidth h be good if either it minimizes the

asymptotic mean squared error or it minimizes what is called the prediction error. We introduce a bandwidth choice method, called "cross-validation", in Section 2.4. In cross-validation, we choose the bandwidth as a minimizer of an estimate of prediction error.

We note that the dependence structure of the ε's affects the asymptotic mean squared error of the estimator and thus our choice of the bandwidth, if we choose the bandwidth as a minimizer of the asymptotic mean squared error. For example, Hall and Hart (1990) proved that the asymptotic variance of the Nadaraya-Watson estimator depends on the dependence structure of the ε's when we observe data $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_i = i/n$.

If we use cross-validation, we should use an appropriate estimate of prediction error according to the dependence structure of the ε's. For model (1.2), a widely-used estimate of prediction error is the estimate computed by the "leave-one-out" cross-validation:

$$\widehat{PE}(h) \equiv \sum_{i=1}^{n} \left(\hat{m}^{(-i)}(x_i) - y_i \right)^2, \tag{1.7}$$

where

$$\hat{m}^{(-i)}(x) \equiv \frac{\sum_{j \in A_i} K\left(\frac{x_j - x}{h}\right) y_j}{\sum_{j \in A_i} K\left(\frac{x_j - x}{h}\right)} \tag{1.8}$$

and $A_i \equiv \{1, \ldots, n\} \cap \{i\}^C$. That is, $\hat{m}^{(-i)}$ is the Nadaraya-Watson estimator, defined in (1.6), computed with the i^{th} observation removed.

However, (1.7) may give an inaccurate estimate of prediction error if the data are generated from a model other than (1.2), which may lead to an unsatisfactory choice of the bandwidth. For example, if ε_i's in model (1.2) are correlated, using (1.7) for such model tends to give a bandwidth that undersmooths the data (i.e. too small bandwidth) when the ε's are positively correlated and give one that oversmooths the data when negatively correlated (see e.g. Chu and Marron, 1991, and Hart, 1994, and the references therein). Chu and Marron (1991) modified (1.7) for use for the dependent ε's, which Burman, Chow, and Nolan (1994) also proposed for use in the analysis of time-series data. We will introduce their estimate in Section 2.4. For another approach, Hart (1994) proposed to, roughly speaking, modify the set A_i in (1.8) to be $A_i = \{1, \ldots, i-1\}$ and also modify (1.7) according to the dependence structure of the ε's.

We have now introduced all the concepts necessary to introduce our research question

and our estimator. In the next chapter, we discuss our research question, the estimation of the drift coefficient μ of a positive recurrent and strictly stationary diffusion process when we observe X_{t_i}'s with additive measurement errors at discrete times $t_1, \ldots, t_n$, and we introduce our Nadaraya-Watson estimator of μ.

Chapter 2

Kernel Estimation of the Drift Coefficient of a Diffusion Process in the Presence of Measurement Error

2.1 Introduction

Financial time-series data such as stock prices, interest rates and derivative prices can be modeled as diffusion processes. A diffusion process is completely characterized by two functions, the drift coefficient which is related to the expected return of an asset for an infinitesimal amount of time and the diffusion coefficient which is related to the variance of the return for an infinitesimal amount of time. When we model time-series data as a diffusion process, we are interested in estimating these two functions as they completely characterize the underlying process. In addition, the diffusion coefficient integrated over time, which is called integrated volatility, has also received attention as a risk measure of an asset (see e.g. Andersen et al., 2001).

Recently, analysis of ultra-high frequency data revealed an ugly fact that we observe financial time series data with measurement errors, called microstructure noise in the financial econometrics literature, which is negligible compared to the observed return in low sampling frequency but has a significant effect in high sampling frequency (Zhou, 1996). While there are

approaches that deal with the measurement error problem in the integrated volatiliy estimation literature (see e.g. Zhang, Mykland, and Aït-Sahalia, 2005), there are few papers, to our knowledge, that incorporate the measurement error problem in the estimation of the drift and the diffusion coefficients. In this chapter, we focus on estimation of the drift coefficient, and we provide a nonparametric estimator of the drift coefficient that is consistent and asymptotically normal in the presence of measurement error under the assumption that the underlying process is stationary.

Integrated volatility is defined as the integral of the squared diffusion coefficient with respect to time over a fixed time period, which is identical to the integrated quadratic variation of the process. Integrated volatility represents variability of a financial instrument for a given period of time, for example, variability of a stock price within a day. A widely used estimator of integrated volatility proposed by Andersen et al. (2001) is the realized volatility estimator, which is simply the sum of squared instantaneous returns. The theory of quadratic variation tells that the realized volatility estimator is an unbiased and consistent estimator of integrated volatility when there is no measurement error. For details about the realized volatility estimator, see e.g. Andersen et al. (2009).

However, according to Zhang, Mykland, and Aït-Sahalia (2005), researchers knew that the performance of the realized volatility estimator is not satisfactory when the measurement error is present and the data are sampled at high frequency. So the researchers purposely used low-frequency data to avoid estimation problems. Zhang, Mykland, and Aït-Sahalia (2005) formalized this approach, which they call the subsampling method, and proposed an estimator of integrated volatility which uses the subsampling method. They first chose a subsampling frequency by minimizing the mean squared error of the realized volatility estimator when the measurement error is present. Then they split the high frequency data into subdata with the chosen subsampling frequency and with different starting times. For example, if the data are sampled hourly and the subsampling frequency is 24 hours, they would create 24 subdata where the k^{th} subdata contain values at hour k every day. After that, they obtained an estimate by using the realized volatility estimates obtained from all subdata.

Other approaches proposed to deal with the measurement error problem in the context of integrated volatility estimation are that of Barndorff-Nielsen et al. (2008), who proposed the

realized kernel estimator which computes the kernel-weighted average of autocorrelations of the process, and that of Jacod et al. (2009) who proposed the preaveraging estimator which uses an average of the returns computed at low sampling frequency.

In the literature of the estimation of the drift and the diffusion coefficients, the diffusion process is usually assumed to be time-homogeneous, that is, the drift and the diffusion coefficients are not functions of time, but rather functions of the value of the process only, and that the process is stationary. Early work on the nonparametric estimation of the two functions includes that of Florens-Zmirou (1993) who provided a Nadaraya-Watson kernel estimator of the diffusion coefficient with the uniform kernel, Aït-Sahalia (1996) who estimated the diffusion coefficient nonparametrically under the parametric specification of the drift coefficient, and Stanton (1997) who proposed a Nadaraya-Watson kernel estimator of the drift and the diffusion coefficients. Later, Bandi and Phillips (2003) provided Nadaraya-Watson kernel estimators of the drift and the diffusion coefficients under more general conditions, including non-stationarity, and proved consistency and asymptotic normality of their estimators.

As we stated earlier, in contrast to estimation of integrated volatility, there are few studies, to our knowledge, that consider estimation of the drift and the diffusion coefficients in the presence of measurement error. An exception is Bandi, Corradi, and Moloche (2009), who consider a standard deviation of the measurement error that converges to zero as the sampling frequency increases to infinity. In our paper, we focus on estimation of the drift coefficient and consider a less restrictive form of the measurement error. We extend the result of Bandi and Phillips (2003) and propose a Nadaraya-Watson type kernel estimator of the drift coefficient which is consistent and asymptotically normal in the presence of independent measurement errors of mean zero and bounded variance.

The structure of the chapter is as follows. In Section 2.2, we introduce our assumptions and define our estimator, and we state the consistency and asymptotic normality of our estimator. In Section 2.3, we compare our estimator to the existing nonparametric estimators of the drift coefficient, especially those proposed by Bandi and Phillips (2003). In Section 2.4, we discuss the bandwidth choice problem of our estimator. In Section 2.5, we describe our simulation study. We will prove the consistency and asymptotic normality result in Section 2.6.

2.2 Statement of the Main Result

We consider the following stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t \tag{2.1}$$

where μ and σ are real-valued functions called the drift and the diffusion coefficients respectively, $\{W_t \mid t \geq 0\}$ is Brownian motion defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A real-valued initial value random variable, X_0, is also defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and independent of Brownian motion. To define a strong solution (a sample path solution) to (2.1), we define an augmented filtration $\mathcal{F}_t$, $t \geq 0$, based on the filtration

$$\mathcal{G}_t \equiv \sigma\big(X_0\,,\ \{W_s \mid 0 \leq s \leq t\}\big), \quad t \geq 0,$$

and the collection of all subsets of measure zero:

$$\mathcal{N} \equiv \Big\{ N \subseteq \Omega \;\Big|\; \exists G \in \bigcup_{t \geq 0} \mathcal{G}_t \text{ with } N \subseteq G \text{ and } \mathbb{P}(G) = 0 \Big\}.$$

Then an augmented filtration $\mathcal{F}_t$, $t \geq 0$, is defined by

$$\mathcal{F}_t \equiv \sigma\big(\mathcal{G}_t \cup \mathcal{N}\big), \quad t \geq 0.$$

A strong solution to (2.1) is a process $\{X_t \mid t \geq 0\}$ adapted to the augmented filtration $\{\mathcal{F}_t \mid t \geq 0\}$ such that the following almost surely holds:

$$X_t = X_0 + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dW_s, \tag{2.2}$$

which is the integrated version of (2.1). See Karatzas and Shreve (1991, page 285) for a formal definition of a strong solution.

Our objective is to provide a consistent and asymptotically normal Nadaraya-Watson type kernel estimator of $\mu(x)$ from observations of a sample path of the solution process $\{X_t \mid t \geq 0\}$ sampled discretely in time and with additive measurement error. To formalize the setting,

suppose that we observe the values of a sample path of the solution process $\{X_t\}$ at times $t \in \{t_1, \ldots, t_n \mid t_k \in [0, T]\}$ for some time span $T > 0$ and that the times are equispaced, i.e. $t_i = i\Delta$ for some $\Delta > 0$. Then we suppose that we observe $\{Y_{i\Delta}\}_{i=1}^n$ such that

$$Y_{i\Delta} \equiv X_{i\Delta} + \varepsilon_{i\Delta} \tag{2.3}$$

where $\{\varepsilon_{i\Delta}\}_{i=1}^n$ are values from a process $\{\varepsilon_t \mid t \geq 0\}$ which is independent of $\{X_t\}$.

Our objective is to estimate $\mu(x)$ from $\{Y_{i\Delta}\}_{i=1}^n$. Our key idea is to estimate $\mu(x)$ by averaging the $Y_{i\Delta}$'s neighboring in time, expecting that the averaging reduces the noise caused by $\varepsilon_{i\Delta}$'s and reveals the latent solution process $\{X_t \mid t \geq 0\}$. Formally speaking, we construct a new stochastic process by averaging the $Y_{i\Delta}$'s in m blocks, each of size r, as in Definition 2.1 below.

Definition 2.1 *For fixed $\Delta > 0$ and r and $n \in \mathbb{N}$, let $\bar{Y}_j^{r,\Delta}$ be the arithmetic average of the $Y_{i\Delta}$'s over i such that $(j-1)r < i \leq jr$. In other words, for $j = 1, \ldots, m \equiv \lfloor n/r \rfloor$ (the largest integer no greater than n/r),*

$$\bar{Y}_j^{r,\Delta} \equiv \frac{1}{r} \sum_{i=1}^{r} Y_{[(j-1)r+i]\Delta}.$$

In addition, we define $\bar{X}_j^{r,\Delta}$ and $\bar{\varepsilon}_j^{r,\Delta}$ similarly as the arithmetic averages of the $X_{i\Delta}$'s and the $\varepsilon_{i\Delta}$'s over i such that $(j-1)r < i \leq jr$, respectively.

Our estimator of $\mu(x)$, given in Definition 2.2 below, is a weighted average of the discrete slopes $(\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta})/r\Delta$ with $j = 1, \ldots, m-2$.

Definition 2.2 *Let $K : \mathbb{R} \to \mathbb{R}$ be a known function and $h > 0$. Let*

$$\hat{\mu}_{\bar{Y}}(x) \equiv \frac{\frac{1}{m-2} \sum_{j=1}^{m-2} \frac{\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta}}{r\Delta} \frac{1}{h} K\left(\frac{\bar{Y}_j^{r,\Delta} - x}{h}\right)}{\frac{1}{m-2} \sum_{j=1}^{m-2} \frac{1}{h} K\left(\frac{\bar{Y}_j^{r,\Delta} - x}{h}\right)} \equiv \frac{\mathcal{N}_{\bar{Y}_1, \ldots, \bar{Y}_m}(x)}{\mathcal{D}_{\bar{Y}_1, \ldots, \bar{Y}_{m-2}}(x)} \equiv \frac{\mathcal{N}_{\bar{Y}}(x)}{\mathcal{D}_{\bar{Y}}(x)}.$$

Note that the j^{th} summand of $\mathcal{N}_{\bar{Y}}(x)$ contains $\bar{Y}_j^{r,\Delta}$, in the argument of K, and the difference $\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta}$ (not $\bar{Y}_{j+1}^{r,\Delta} - \bar{Y}_j^{r,\Delta}$). We chose these indices (j, $j+1$, $j+2$) to make our asymptotic calculations easier: with these indices, the difference $\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta}$ depends on values of $X_t + \varepsilon_t$

with t in $[(jr+1)\Delta, (j+2)r\Delta]$, while $\bar{Y}_j^{r,\Delta}$ depends on values of $X_t + \varepsilon_t$ for t in a different interval. When it comes to the finite-sample performance, we saw in our simulation, which is not included in Section 2.5, that the shifts increase the mean squared error of our pre-averaging estimator. We will discuss this issue more concretely in Chapter 3.

Our proof of consistency and asymptotic normality of $\hat{\mu}_{\bar{Y}}(x)$ requires that the observation time lag Δ tends to zero and the observation time span $T = n\Delta$ tends to infinity as the number of observations n tends to infinity. These assumptions are necessary. As Bandi and Phillips (2003) note, without the condition $\Delta \to 0$, we suffer from what is called the aliasing problem: "different continuous-time processes may be indistinguishable when we observe the process discretely in time." If Δ is fixed and n tends to infinity, the data form a discrete-time process. We may be able to deduce some properties of the discrete time process. But we cannot identify what continuous-time process generated the data, as there is usually more than one continuous-time process that can generate the discrete-time process.

They also note that, without the condition $n\Delta \to \infty$, we cannot obtain a consistent estimator of $\mu(x)$ in general, even if the process is observed without measurement error. In addition to these assumptions on Δ, our proof requires similar conditions for the averaged process: the time lag between the two adjacent averages, $r\Delta$, tends to zero and the number of blocks n/r tends to infinity.

Now we introduce the assumptions.

Assumption 2.1 *As $n \to \infty$, the sequence of positive real numbers $\{\Delta_n\}_{n=1}^{\infty}$ and the sequence of positive integers $\{r_n\}_{n=1}^{\infty}$ satisfy $\Delta_n \to 0$, $n\Delta_n \to \infty$, $r_n \to \infty$, $r_n\Delta_n \to 0$ and $n/r_n \to \infty$.*

We will often denote $m_n \equiv n/r_n$, which represents the number of blocks (which equals the number of $\bar{Y}_j^{r,\Delta}$'s).

Assumption 2.2 *The functions μ and σ are Borel-measurable and twice continuously differentiable on $\mathbb{R}$. In addition, $\sigma^2(x) > 0$ for all $x \in \mathbb{R}$.*

Assumption 2.2 is a sufficient condition for the existence and uniqueness of a strong solution of the stochastic differential equation (2.1), as discussed in Bandi and Phillips (2003, page 244).

Assumption 2.3 *The solution process $\{X_t\}$ is positive recurrent and strictly stationary. Let f_X be the stationary density. The functions μ, σ and f_X satisfy*

$$\int \mu^2(x) f_X(x) dx < \infty \quad \text{and} \quad \int \sigma^2(x) f_X(x) dx < \infty.$$

Note that, if the solution process $\{X_t\}$ is positive recurrent, there exists a random variable X whose probability density function is f_X, called the stationary density, such that $X_t \xrightarrow{d} X$ as $t \to \infty$. If we let the initial value random variable X_0 have the density function f_X, then $\{X_t\}$ is strictly stationary (Kutoyants, 2004, page 2).

Note that Assumption 2.2 and the positive recurrence assumption in Assumption 2.3 imply that f_X is continuous: if $\{X_t\}$ is positive recurrent, f_X is given by

$$f_X(x) = \frac{1}{G\sigma^2(x)} \exp\left\{2 \int_0^x \frac{\mu(y)}{\sigma^2(y)} dy\right\}, \tag{2.4}$$

where G is a normalizing constant (Kutoyants, 2004, Theorem 1.16, page 40).

Assumption 2.4 *The kernel $K \in L^2(\mathbb{R})$ is bounded, symmetric, nonnegative and continuously differentiable. Its derivative, K', is bounded and is in $L^1(\mathbb{R})$. In addition,*

$$\int_{-\infty}^{\infty} K(x) dx = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} s^2 K(s) ds < \infty.$$

Assumption 2.5 *The error process $\{\varepsilon_t\}$ is independent of $\{X_t\}$, and the ε_t's are independent across t. Also, $\mathbb{E}(\varepsilon_t) = 0$ for all t, and there exists a finite, positive constant σ_ε^2 such that $\sup_t Var(\varepsilon_t) \leq \sigma_\varepsilon^2$.*

In the literature, the $\varepsilon_{i\Delta_n}$'s with i in $1, \ldots, n$ are usually assumed to be independent and identically distributed and that $\mathbb{E}(\varepsilon_{i\Delta_n}) = 0$. Some authors assume that $\text{Var}(\varepsilon_{i\Delta_n}) = \sigma_\varepsilon^2$ for all i and n (see Zhang, Mykland, and Aït-Sahalia, 2005, among others). In contrast, some papers in the literature, and most of the papers in the rounding error literature according to Jacod et al. (2009), assume that $\text{Var}(\varepsilon_{i\Delta_n}) = a_n \sigma_\varepsilon^2$ for all i where $a_n \to 0$ as $n \to \infty$ (see Bandi, Corradi, and Moloche, 2009, among others). Our Assumption 2.5 includes both specifications as special cases.

Now we introduce our main result.

Theorem 2.1 *Suppose Assumptions 2.1 to 2.5 hold, and suppose*

$$(i) \qquad \left(\frac{n\Delta_n}{h_n}\right)^2 r_n\Delta_n \ln(1/r_n\Delta_n) = o(1),$$

$$(ii) \qquad n\Delta_n h_n \to \infty, \quad \textit{and}$$

$$(iii) \qquad \frac{n}{h_n^3 r_n^2} = o(1).$$

Let $K_2 \equiv \int K^2(s)ds$ *and* $\nu_2 \equiv \int s^2 K(s)ds$. *Then the following consistency and asymptotic normality results hold for every* $x \in \{y \mid f_X(y) > 0\}$.

1. $\hat{\mu}_{\bar{Y}}(x) \longrightarrow \mu(x)$ *in probability as* $n \to \infty$.

2. *If* $n\Delta_n h_n^5 = o(1)$, *then*

$$\sqrt{(n-r_n)\Delta_n h_n}\ \{\hat{\mu}_{\bar{Y}}(x) - \mu(x)\} \xrightarrow{d} N\left(0, K_2\frac{\sigma^2(x)}{f_X(x)}\right).$$

3. *If* $n\Delta_n h_n^5 = O(1)$, *then*

$$\sqrt{(n-r_n)\Delta_n h_n}\ \left\{\hat{\mu}_{\bar{Y}}(x) - \mu(x) - h_n^2\Gamma_\mu(x)\right\} \xrightarrow{d} N\left(0, K_2\frac{\sigma^2(x)}{f_X(x)}\right)$$

where

$$\Gamma_\mu(x) = \nu_2 \times \left(\mu'(x)\frac{f_X'(x)}{f_X(x)} + \frac{1}{2}\mu''(x)\right).$$

The conclusions 1 and 2 give consistency and asymptotic normality of $\hat{\mu}_{\bar{Y}}(x)$. The conclusion 3 gives asymptotic bias and variance, which are useful for the choice of the bandwidth h.

Bandi and Phillips (2003) provided consistency and asymptotic normality of the Nadaraya-Watson estimator of μ when one observes a sample path of a *recurrent* diffusion process $\{X_t\}$ sampled discretely in time and *without measurement error*. We prove consistency and asymptotic normality of our estimator by showing that the difference between our estimator and their estimator converges to 0 asymptotically, under our stronger assumptions. The full proof of Theorem 2.1 can be found in Section 2.6.

We finish this section by considering simple sufficient conditions so that Δ_n, r_n and h_n satisfy the conditions of Theorem 2.1. Suppose $\Delta_n = n^{-\delta}$, $r_n = n^{\rho}$ and $h_n = n^{-\eta}$ for some positive real numbers δ, ρ and η. We study conditions on δ, ρ and η so that the conditions of Theorem 2.1 are satisfied.

To begin with, from Assumption 2.1, we have

$$0 < \rho < \delta < 1. \tag{2.5}$$

Also, from condition (i) of Theorem 2.1, we have

$$(\rho - \delta) n^{2-3\delta+2\eta+\rho} \ln n = o(1),$$

that is, since $\rho \neq \delta$ by (2.5), we have $2 - 3\delta + 2\eta + \rho < 0$, or

$$\eta < \frac{3}{2}\delta - \frac{1}{2}\rho - 1. \tag{2.6}$$

Condition (ii) becomes $1 - \delta - \eta > 0$, or

$$\eta < 1 - \delta. \tag{2.7}$$

Condition (iii) becomes $1 + 3\eta - 2\rho < 0$, or

$$\eta < \frac{1}{3}(2\rho - 1). \tag{2.8}$$

In addition, for the condition of Theorem 2.1's conclusion 2, that $n\Delta_n h_n^5 = o(1)$, we require

$$\eta > \frac{1}{5}(1 - \delta). \tag{2.9}$$

Lastly, for the condition of Theorem 2.1's conclusion 3, that $n\Delta_n h_n^5 = O(1)$, we require

$$\eta \geq \frac{1}{5}(1 - \delta). \tag{2.10}$$

For example, suppose that $\delta = 0.9$ and $\rho = 0.58$, which satisfies (2.5). Then the conditions in Equations (2.6) to (2.8) are satisfied provided

$$0.02 \leq \eta < 0.0533.$$

Equation (2.9) is satisfied if $\eta > 0.02$ and Equation (2.10) is satisfied if $\eta = 0.02$. Table 2.1 shows values of Δ_n, r_n and h_n for $\delta = 0.9$, $\rho = 0.58$ and $\eta = 0.02$, when $n = 5000,\ 10000,\ 15000$.

Table 2.1: The values of Δ_n, r_n, m_n and h_n when $\delta = 0.9$, $\rho = 0.58$ and $\eta = 0.02$.

n	$\Delta_n = n^{-0.9}$	$r_n = n^{0.58}$	$m_n \approx n^{0.42}$	$r_n\Delta_n = n^{-0.32}$	$h_n = n^{-0.02}$
5,000	0.00047	139	35	0.0655	0.8434
10,000	0.00025	208	48	0.0525	0.8318
15,000	0.00017	264	56	0.0461	0.8250

Remark: The distribution of $\{\bar{X}_j^{r,\Delta}\}$ is not clear, although we suspect that the marginal distribution of $\bar{X}_j^{r,\Delta}$ converges in distribution to the marginal distribution of X_t under the conditions of Theorem 2.1. We can prove adapting the proof of Theorem 2.1 that

$$\frac{\frac{1}{m_n-2}\sum_{j=1}^{m_n-2}\frac{\bar{X}_{j+2}^{r_n,\Delta_n}-\bar{X}_{j+1}^{r_n,\Delta_n}}{r_n\Delta_n}\frac{1}{h_n}K\left(\frac{\bar{X}_j^{r_n,\Delta_n}-x}{h_n}\right)}{\frac{1}{m_n-2}\sum_{j=1}^{m_n-2}\frac{1}{h_n}K\left(\frac{\bar{X}_j^{r_n,\Delta_n}-x}{h_n}\right)} - \frac{\frac{1}{m_n-2}\sum_{j=1}^{m_n-2}\frac{X_{(j+1)r_n\Delta_n}-X_{jr_n\Delta_n}}{r_n\Delta_n}\frac{1}{h_n}K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)}{\frac{1}{m_n-2}\sum_{j=1}^{m_n-2}\frac{1}{h_n}K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)}$$

converges to 0 in probability as $n \to \infty$, where the two terms on the left-hand side are constructed by replacing $\bar{Y}_j$'s in Definition 2.2 with $\bar{X}_j$'s and with X_t's, respectively.

2.3 Comparison to the Existing Estimators

Recall that our Theorem 2.1 is based on the result of Bandi and Phillips (2003). Our pre-averaging estimator is similar to their estimator which they call the "double-smoothing estimator", and we compare the two in this section. After that, we discuss the "subsampling method", which is used to estimate integrated volatility when measurement error is present. The subsampling method can also be applied to estimation of the drift coefficient, so we will compare it to our pre-averaging estimator.

We first compare the double-smoothing estimator of Bandi and Phillips (2003) with our estimator. Suppose that we observe the time-equispaced data $\{Y_{i\Delta}\}_{i=1}^{n}$. Both estimators are of the form

$$\frac{\sum_{j=1}^{w} \frac{1}{h} K\left(\frac{W_j^{kern}-x}{h}\right) W_j^{slope}}{\sum_{j=1}^{w} \frac{1}{h} K\left(\frac{W_j^{kern}-x}{h}\right)}.$$

In our pre-averaging estimator,

$$w = m-2, \quad W_j^{kern} = \bar{Y}_j^{r,\Delta} \quad \text{and} \quad W_j^{slope} = \frac{\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta}}{r\Delta}.$$

In the double-smoothing estimator of Bandi and Phillips (2003),

$$w = n-1, \quad W_j^{kern} = Y_{j\Delta} \quad \text{and} \quad W_j^{slope} = \frac{1}{N_j^{l,\Delta}} \sum_{k:|Y_{k\Delta}-Y_{j\Delta}|\leq l} \frac{Y_{(k+1)\Delta} - Y_{k\Delta}}{\Delta}, \tag{2.11}$$

where $l \in \mathbb{R}^+$ and $N_j^{l,\Delta}$ is the number of $Y_{k\Delta}$'s such that $|Y_{k\Delta} - Y_{j\Delta}| \leq l$. Note that l can be interpreted as the bandwidth of the uniform kernel.

Bandi and Phillips (2003) also define the usual Nadaraya-Watson estimator and call it the "single-smoothing estimator" to contrast it to the double-smoothing estimator. The single-smoothing estimator (that is, the usual Nadaraya-Watson estimator) is defined by setting

$$w = n-1, \quad W_j^{kern} = Y_{j\Delta} \quad \text{and} \quad W_j^{slope} = \frac{Y_{(j+1)\Delta} - Y_{j\Delta}}{\Delta}. \tag{2.12}$$

Note that the single-smoothing and the double-smoothing estimators were introduced for the case of no measurement error, that is, the case where $\varepsilon_t = 0$ for all t, in which case we have $Y_{i\Delta} = X_{i\Delta}$.

We note three differences between the double-smoothing estimator and our estimator. First, the double-smoothing estimator pre-averages $Y_{i\Delta}$'s such that $|Y_{i\Delta} - Y_{j\Delta}| \leq l$ for each j before computing the kernel-weighted average. This is similar to our pre-averaging $Y_{i\Delta}$'s such that $(j-1)r < i \leq jr$ for each j. A difference is that the double-smoothing estimator pre-averages $Y_{i\Delta}$'s according to their values while our estimator pre-averages $Y_{i\Delta}$'s according to their time-indices (the $i\Delta$'s).

Second, while both estimators use averaging for W_j^{slope}, our estimator uses the averaged value, $\bar{Y}_j^{r,\Delta}$, for W_j^{kern} while the double-smoothing estimator uses a single observation, $Y_{j\Delta}$. Third, the double-smoothing estimator was introduced and studied for the case of no measurement error, so its consistency in the presence of measurement error is not yet established. In contrast, Theorem 2.1 states that our estimator is consistent in the presence of measurement error as well as in the case of no measurement error (that $\varepsilon_t = 0$ for all t satisfies Assumption 2.5). Our simulation study will indicate that, for finite samples, the double-smoothing estimator has higher mean squared error than our estimator when there is measurement error.

Now we discuss the subsampling method, which Zhang, Mykland, and Aït-Sahalia (2005) studied for estimation of integrated volatility, and compare it to our pre-averaging approach. Using notation in (2.1), integrated volatility is defined as $\int_a^b \sigma^2(X_t)dt$ for some fixed time period $[a, b]$. When we observe data $\{X_{t_i}\}_{i=1}^n$ where $a = t_1 < t_2 < \ldots < t_n = b$, Andersen et al. (2001) proposed $\sum_{i=1}^{n-1}(X_{t_{i+1}} - X_{t_i})^2$, called the realized volatility estimator, as an estimator of the integrated volatility. They showed that $\sum_{i=1}^{n-1}(X_{t_{i+1}} - X_{t_i})^2$ converges to $\int_a^b \sigma^2(X_t)dt$ in probability as n tends to infinity.

However, the realized volatility estimator does not give an accurate estimate of the integrated volatility when the data are observed with measurement errors and when the data are sampled at high frequency, i.e. when the $t_{i+1} - t_i$'s are small. Zhang, Mykland, and Aït-Sahalia (2005) showed that, when we observe data $\{Y_{t_i} = X_{t_i} + \varepsilon_{t_i}\}_{i=1}^n$ where the ε_{t_i}'s are independent and identically distributed with mean zero and variance s_ε^2,

$$\sum_{i=1}^{n-1}(Y_{t_{i+1}} - Y_{t_i})^2 = 2ns_\varepsilon^2 + O_p(n^{1/2})$$

(Zhang, Mykland, and Aït-Sahalia, 2005, page 1395, Equation 5). This proves that the realized volatility estimator does not converge in probability to the integrated volatility in the presence of measurement error.

As Zhang, Mykland, and Aït-Sahalia (2005) noted, in order to avoid this estimation problem of the realized volatility estimator, researchers used the data sampled at lower frequency, namely, the data $\{Y_{t_k}, Y_{t_{2k}}, \ldots\}$ for some $k > 1$ instead of $\{Y_{t_1}, Y_{t_2}, \ldots\}$, to compute the realized volatility estimate. When considering $Y_{t_{(i+1)k}} - Y_{t_{ik}} = (X_{t_{(i+1)k}} - X_{t_{ik}}) + (\varepsilon_{t_{(i+1)k}} - \varepsilon_{t_{ik}})$ for each i

for some large k, the difference $X_{t_{(i+1)k}} - X_{t_{ik}}$ is relatively larger in magnitude than $\varepsilon_{t_{(i+1)k}} - \varepsilon_{t_{ik}}$. This yields $\sum_i (Y_{t_{(i+1)k}} - Y_{t_{ik}})^2 \approx \sum_i (X_{t_{(i+1)k}} - X_{t_{ik}})^2$ when k is large. Recall that $\sum_i (X_{t_{(i+1)k}} - X_{t_{ik}})^2$ is a consistent estimator of the integrated volatility. Zhang, Mykland, and Aït-Sahalia (2005) formalized this ad-hoc approach and proposed to choose k as the minimizer of the asymptotic mean squared error of the estimator $\sum_i (Y_{t_{(i+1)k}} - Y_{t_{ik}})^2$, and they called this approach the subsampling method.

We compare the subsampling method to our pre-averaging approach by considering hourly-observed stock price data. Our pre-averaging approach with the block size of a day obtains $\bar{Y}$'s that are average daily prices. In contrast, the equivalent subsampling method uses daily closing prices to construct the subsampled data.

Recall that the estimators of Bandi and Phillips (2003), defined in (2.11) and (2.12), use $\{Y_{i\Delta}\}_{i=1}^n$ as data while our pre-averaging estimator, defined in Definition 2.2, uses $\{\bar{Y}_j^{r,\Delta}\}_{j=1}^m$ where $m = n/r$ and r is the block size. We can apply the subsampling method to the estimators of Bandi and Phillips (2003) by using $\{Y_{jr\Delta}\}_{j=1}^m$ as the data instead of $\{Y_{i\Delta}\}_{i=1}^n$. Our simulation study, which will be given in Section 2.5, indicates that applying the subsampling method leads to much lower mean squared errors for the estimators of Bandi and Phillips (2003). The estimators of Bandi and Phillips (2003) without subsampling have higher mean squared errors than our estimator. However, the mean squared errors of the estimators with subsampling are about the same as ours.

2.4 Bandwidth Choices

In Section 2.2, we stated Theorem 2.1, that $\hat{\mu}_{\bar{Y}}$ is a consistent and asymptotically normal estimator of μ in the presence of measurement error under some conditions. We can see that many sequences of $\{h_n\}$ and $\{r_n\}$ satisfy the conditions of the theorem. Hence, it is desirable to have a principle of the choice of h and r given the sample size n. We discuss the choice of h in Section 2.4.1 and the choice of r in Section 2.4.2.

2.4.1 Choice of the kernel bandwidth h

The kernel bandwidth choice criterion has been extensively studied in the literature, at least in certain circumstances such as analysis of cross-sectional data. For an overview of bandwidth choice methods, see e.g. Jones, Marron, and Sheather (1996) and the references therein. In this subsection, we discuss the two popular methods to choose h, the plug-in method and the cross-validation method, in the context of our estimator. After that, we briefly introduce a bandwidth choice method recently proposed by Bandi, Corradi, and Moloche (2009). This method is explicitly intended for kernel estimators of the drift coefficient μ and the diffusion coefficient σ of a diffusion process.

The plug-in method requires an expression for the asymptotic mean squared error of the estimator and the existence of h_{opt}, the h that minimizes the asymptotic mean squared error. We then "plug into" the expression for h_{opt} estimates of all unknown quantities. This yields $\hat{h}_{opt}$, the plug-in bandwidth. In the context of our estimator, we first obtain the asymptotic bias and the asymptotic variance of $\hat{\mu}_{\tilde{Y}}(x)$ using conclusion 3 of Theorem 2.1:

$$\text{asymptotic bias} = h^2 \Gamma_\mu(x) \tag{2.13}$$

and

$$\text{asymptotic variance} = \frac{K_2 \sigma^2(x)/f_X(x)}{(n-r)\Delta h}.$$

The asymptotic mean squared error (AMSE) of our estimator is, then, the sum of the squared asymptotic bias and the asymptotic variance:

$$\text{AMSE}(x) = h^4 \Gamma_\mu^2(x) + \frac{K_2 \sigma^2(x)/f_X(x)}{(n-r)\Delta h}.$$

Now we find the bandwidth h that minimizes the AMSE. When $\Gamma_\mu^2(x) \neq 0$, differentiation yields the minimizer of the AMSE, $h_{opt}(x)$:

$$h_{opt}(x) = \left(\frac{K_2 \sigma^2(x)/f_X(x)}{4\Gamma_\mu^2(x)} \right)^{1/5} \times \left(\frac{n}{n-r} \right)^{1/5} \times \left(\frac{1}{n\Delta} \right)^{1/5}. \tag{2.14}$$

When $\Gamma_\mu^2(x) = 0$, the AMSE decreases to 0 as h approaches infinity.

While it is reasonable to take $h_{opt}(x)$ given by (2.14) as the bandwidth for our estimator, in practice the values of $\Gamma_\mu(x)$, $\sigma^2(x)$ and $f_X(x)$ are not known. The plug-in bandwidth is gotten by plugging estimates of $\Gamma_\mu(x)$, $\sigma^2(x)$ and $f_X(x)$ into (2.14). The estimates of these unknowns can be obtained either parametrically or nonparametrically. The "ideal" bandwidth, $h_{opt}(x)$, is called the oracle bandwidth and is often used in simulation studies as a gold standard.

Our simulation study in Section 2.5 indicates that the cross-validation bandwidth, which is discussed below, exhibits better finite sample performance than the oracle bandwidth. Therefore, we do not consider estimating $\Gamma_\mu(x)$, $\sigma^2(x)$ and $f_X(x)$ to calculate the plug-in bandwidth, and we recommend using the cross-validation bandwidth instead of the plug-in bandwidth.

The cross-validation method of choosing h attempts to minimize what is called the prediction error as follows. We can think of our estimator $\hat{\mu}_{\bar{Y}}(x)$ as a predictor of $(X_{t+\delta} - X_t)/\delta$ given that $X_t = x$. In this perspective, we choose the bandwidth h so that the resulting estimator $\hat{\mu}_{\bar{Y}}(x)$ has the least error in predicting $(X_{t+\delta} - X_t)/\delta$ given $X_t = x$, where the prediction error is estimated using the data.

To compute an estimate of the prediction error, we can use what is called the H-block cross-validation method proposed by Chu and Marron (1991) and further developed by Burman, Chow, and Nolan (1994), who coined the name "H-block". The H-block cross-validation modifies the well-known leave-one-out cross-validation (Stone, 1974), used for independent data, for use with stationary time-series data. In cross-validation, in order to estimate the prediction error, one predicts a data value by using information in a portion of the data set, called the training data. Ideally, the training data and the target data value are independent. In leave-one-out cross-validation, one constructs the training data by omitting one observation. In H-block cross-validation, one omits $2H$ more observations, H neighboring observations in the past and H neighboring observations in the future. The objective of omitting the additional $2H$ observations is to weaken the dependence between the target data value and the training data: if the time between the two is large enough, we expect that the autocorrelation between the two is close to zero. This expectation is valid if the time-series data are stationary, which we assume in Assumption 2.3.

Now we formally describe the cross-validation bandwidth choice method based on H-

block cross-validation in the context of our estimator. We estimate the prediction error by

$$\widehat{PE}(h\,;\,H) \equiv \sum_{k=1}^{m-1} \left(\hat{\mu}_{\bar{Y}}^{(k,H)} \left(\bar{Y}_k^{r,\Delta} \right) - \frac{\bar{Y}_{k+1}^{r,\Delta} - \bar{Y}_k^{r,\Delta}}{r\Delta} \right)^2 \tag{2.15}$$

where H is an integer and $\hat{\mu}_{\bar{Y}}^{(k,H)} \left(\bar{Y}_k^{r,\Delta} \right)$ is our drift coefficient estimator $\hat{\mu}_{\bar{Y}}(x)$ evaluated at $x = \bar{Y}_k^{r,\Delta}$ calculated by removing the terms $j = k-H, \ldots, k+H$ of the sums in Definition 2.2, that is,

$$\hat{\mu}_{\bar{Y}}^{(k,H)} \left(\bar{Y}_k^{r,\Delta} \right) \equiv \frac{\sum_{j \in \mathcal{A}_k^H} \frac{\bar{Y}_{j+2}^{r,\Delta} - \bar{Y}_{j+1}^{r,\Delta}}{r\Delta} \frac{1}{h} K \left(\frac{\bar{Y}_j^{r,\Delta} - \bar{Y}_k^{r,\Delta}}{h} \right)}{\sum_{j \in \mathcal{A}_k^H} \frac{1}{h} K \left(\frac{\bar{Y}_j^{r,\Delta} - \bar{Y}_k^{r,\Delta}}{h} \right)}$$

for the set of indices $\mathcal{A}_k^H = \{1, \ldots, m-2\} \bigcap \{k-H, \ldots, k+H\}^C$. The integer H is chosen so that the correlation between $\bar{Y}_k^{r,\Delta}$ and $\bar{Y}_j^{r,\Delta}$'s with $j \in \mathcal{A}_k^H$ is "weak enough". For the implementation, H can be chosen by looking at the empirical autocorrelation function.

The value $(\bar{Y}_{k+1}^{r,\Delta} - \bar{Y}_k^{r,\Delta})/r\Delta$ in (2.15) is called the target data value. Recall that we want to choose h so that the resulting estimator $\hat{\mu}_{\bar{Y}}(x)$ has the least error in predicting $(X_{t+\delta} - X_t)/\delta$ given $X_t = x$. The target data value $(\bar{Y}_{k+1}^{r,\Delta} - \bar{Y}_k^{r,\Delta})/r\Delta$ is considered an estimate of the value $(X_{t+\delta} - X_t)/\delta$ when $t = (k-1)r\Delta$ and $\delta = r\Delta$. Since we cannot observe the underlying process $\{X_t\}$, we use the pre-averaged process for the target data value expecting that the pre-averaged process is close to the underlying process. We choose the target data value not to depend on h in order to prevent interaction between $\hat{\mu}_{\bar{Y}}^{(k,H)} \left(\bar{Y}_k^{r,\Delta} \right)$ and the target data value.

The cross-validation bandwidth h_{cv} is the minimizer of $\widehat{PE}(h\,;\,H)$:

$$h_{cv} \equiv \underset{h}{\operatorname{argmin}}\, \widehat{PE}(h\,;\,H). \tag{2.16}$$

The simulation study in Section 2.5 indicates that the mean integrated squared error of our pre-averaging estimator with h_{cv} is smaller than that of our estimator with the oracle bandwidth $h_{opt}(x)$ in (2.14). Based on this simulation result, we recommend using (2.16) as a bandwidth choice criterion over the plug-in bandwidth.

We finish this subsection by introducing a recently proposed bandwidth choice method in Bandi, Corradi, and Moloche (2009), which is explicitly developed to jointly choose the band-

widths for the estimators of the drift and the diffusion coefficients of a diffusion process. Their method relies on residuals of the fits being approximately independent and normally distributed. Their method consists of two stages, and the first stage is as follows. Given the time-equispaced time-series data $\{X_{i\Delta}\}_{i=1}^{n}$ generated from a diffusion process and the bandwidths (h_{dr}, h_{dif}) applied to the estimators of the drift and the diffusion coefficients respectively, they define the scaled residuals

$$\hat{r}_{i\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta} - \hat{\mu}(X_{i\Delta}; h_{dr})\Delta}{\hat{\sigma}(X_{i\Delta}; h_{dif})\sqrt{\Delta}}$$

for $i = 1, \ldots, n-1$ where $\hat{\mu}(.\,; h_{dr})$ and $\hat{\sigma}(.\,; h_{dif})$ are kernel estimates of the drift and the diffusion coefficients. Then they choose $(h_{dr}^*, h_{dif}^*) \in (0, \infty) \times (0, \infty)$ by

$$(h_{dr}^*, h_{dif}^*) = \underset{h_{dr}, h_{dif}}{\operatorname{argmin}} \sup_{x} |\mathbb{F}_{\hat{r}}(x) - \Phi(x)| \tag{2.17}$$

where $\mathbb{F}_{\hat{r}}$ is the empirical distribution function of $\hat{r}$ and Φ is the distribution function of a standard normal random variable. The justification for this first step is as follows. For small Δ, the drift coefficient μ and the diffusion coefficient σ can be treated as constants in each time interval of $[i\Delta, (i+1)\Delta]$, $i = 1, \ldots, n-1$. We denote such constants by $\mu_{i\Delta}$ and $\sigma_{i\Delta}$. Then, from (2.2), we have

$$X_{(i+1)\Delta} - X_{i\Delta} \approx \int_{i\Delta}^{(i+1)\Delta} \mu_{i\Delta} dt + \int_{i\Delta}^{(i+1)\Delta} \sigma_{i\Delta} dW_t,$$

and so

$$\frac{X_{(i+1)\Delta} - X_{i\Delta} - \mu_{i\Delta}\Delta}{\sigma_{i\Delta}\sqrt{\Delta}} \approx \frac{1}{\sqrt{\Delta}} \int_{i\Delta}^{(i+1)\Delta} dW_t = \frac{W_{(i+1)\Delta} - W_{i\Delta}}{\sqrt{\Delta}} \overset{d}{=} Z_i,$$

where $\{Z_i\}_{i=1}^{n-1}$ are independent and identically distributed standard normal random variables.

After choosing the bandwidth in (2.17), they proceed to the second stage. Here we just introduce the main idea of their second stage. The kernel estimators of the drift and the diffusion coefficients have conditions on the convergence rates of the bandwidth for consistency and asymptotic normality, for example, conditions (i), (ii) and (iii) and the additional condition in conclusion 2 of Theorem 2.1 for our estimator. They construct three random variables

that depend on (h^*_{dr}, h^*_{dif}), the bandwidths chosen in the first step. They derive the asymptotic distribution of a functional of these random variables when (h^*_{dr}, h^*_{dif}) do not satisfy at least one of the conditions. They use this functional and its asymptotic distribution to determine if any of the conditions are violated. If so, they use the values of the three random variables to determine how to adjust h^*_{dr} and h^*_{dif}.

2.4.2 Choice of the block size r

Note first that pre-averaging is a form of smoothing. According to our simulation study using the oracle bandwidth $h_{opt}(x)$ defined in (2.14), there is a bias-variance tradeoff in choosing the block size r (see Figure 2.8). This tradeoff is similar to the well-known tradeoff for the choice of the bandwidth h: a large value of r is likely to produce a constant function estimate of μ and result in large bias but small variance. On the other hand, a small value of r is likely to result in small bias but large variance.

Therefore, one might think of choosing r by the plug-in method or the cross-validation method. However, there are complications in using these approaches. For the plug-in method, we cannot use AMSE for the choice of r because we do not have an asymptotic result that contains r. For the cross-validation method, there are two complications. First, there is a computational issue. If we use the cross-validation method, we should minimize the prediction error with respect to both r and h. However, this two-dimensional optimization problem is computationally burdensome considering that the measurement error problem is often considered for high-frequency data. Second, finding target data values that do not depend on r is not easy. In choosing h, we proposed in (2.15) to use the $\bar{Y}_j^{r,\Delta}$'s. However, when we choose both h and r, using these $\bar{Y}_j^{r,\Delta}$'s in the targets may lead to unsatisfactory choices.

Because of these problems, we recommend using an ad-hoc choice of r, just as researchers estimating integrated volatility used when choosing the subsampling frequency before Zhang, Mykland, and Aït-Sahalia (2005) formalized the subsampling approach. A sensible ad hoc choice of r would be one considering any periodicity of the data. For example, in our simulation study in which we generated daily observed data with five business days per week, we chose $r = 5$ to yield weekly averages.

2.5 Simulation Study

In this section, we carry out a simulation study to assess finite sample performance of our estimator. We simulate data with two kinds of underlying models for the drift coefficient, μ, one with linear drift coefficient and one with nonlinear drift coefficient, and with the methods of choosing h and r discussed in Section 2.4. We consider the following two kinds of underlying models:

$$dX_t = 0.858 \times (0.086 - X_t)dt + 0.157\sqrt{X_t}\, dW_t, \tag{2.18}$$

$$dX_t = -(X_t - 1)(X_t + 1)^2 dt + 2dW_t. \tag{2.19}$$

The process defined by (2.18) is called a Cox-Ingersoll-Ross (CIR) process and is used as an underlying model for a short-term interest rate process. The value of a CIR process at time t equals the annual interest rate, and the time is measured in days, with a year being 250 days (counting business days only). Following the parameter choice of Chapman and Pearson (2000), we use the parameter values $(0.858, 0.086, 0.157)$ in (2.18) to match the solution process's monthly (i.e. 21st-order) autocorrelation, unconditional mean and unconditional variance to the corresponding sample quantities of the dataset of Aït-Sahalia (1996). The dataset is seven-day Eurodollar deposit rates observed daily from June 1, 1973 to February 25, 1995 (total of 5505 observations). The Eurodollar deposit rate is known to move in close connection with short-term interest rates such as T-bill rates (Aït-Sahalia, 1996, page 539). We use the process defined by (2.19) to study the performance of our estimator when the true drift coefficient is nonlinear. It is straightforward to check that each of the models (2.18) and (2.19) satisfies Assumptions 2.2 and 2.3, so that the unique solution process exists and is positive recurrent. We generated 1,000 discretely-observed independent sample paths for each of (2.18) and (2.19) at time increments of $\Delta = 1/250$, which represents daily observations assuming 250 business days a year, and with the number of observations $n = 5505$, which is the sample size of the dataset of Aït-Sahalia (1996). The top panels in Figures 2.1 and 2.2 depict sample paths of the processes defined by (2.18) and (2.19) with these values of Δ and n.

In order to generate sample paths of the model (2.18), we first note that the analytical forms

of the stationary density and the transition density are known. When generating each sample path, we used the package sde in R (Iacus, 2009) to generate an initial value by a random draw from the stationary density, to generate the first observation by a random draw from the transition density given the initial value, to generate the second observation by a random draw from the transition density given the first observation, and so on. Then the data generated by this procedure have the same distribution as the distribution of the discretely observed data of model (2.18).

For model (2.19), we note that the analytical form of the stationary density is known (Equation 2.4), but the analytical form of the transition density is unknown. When generating each sample path, we again used the package sde in R, which first obtains an initial value by a random draw from the stationary density, then implements a numerical approximation method proposed by Milstein to generate the discretely-observed sample paths. Milstein's method uses the first-order and the second-order derivatives of μ and σ. For details on Milstein's method, see e.g. Iacus (2008, Chapter 2, page 81, Equation 29).

We then added independent and identically normally distributed measurement errors to the generated discretely-observed sample paths. For model (2.18), we took 0.002 as the standard deviation of our measurement errors. This value is an estimate of the standard deviation of the measurement error of the dataset of Aït-Sahalia (1996), proposed by Jones (2003, page 812). We note that the value 0.002 is 5.7% of the unconditional standard deviation of the solution process of (2.18). We also set the standard deviation of the measurement error added to model (2.19) to be 5.7% of the unconditional standard deviation of the solution process of (2.19), that is, to be 0.0661. The second panels of Figures 2.1 and 2.2 depict sample paths of the processes defined by (2.18) and (2.19) with measurement errors added.

Using these sample paths with additive measurement errors, we estimated the drift coefficient by our pre-averaging estimator in Definition 2.2 (which we denote "Avg") and the double-smoothing and the single-smoothing estimator of Bandi and Phillips (2003), which are defined in (2.11) and (2.12) respectively and which we denote "BPD" and "BPS", respectively. We also combined the subsampling method explained in Section 2.3 with the BPS and the BPD estimators, and we denote these as "BPSs" and "BPDs". In Avg, we chose $r = 5$ (see Definition 2.1 for the definition of r), which means we took weekly averages assuming 5 business

days a week. In BPSs and BPDs, we used the weekly closing prices (i.e. every fifth value) in order to construct subsampled data. The third and fourth panels of each of Figures 2.1 and 2.2 depict, respectively, averaged and subsampled sample paths of the process defined by each of (2.18) and (2.19).

For all estimators, we used the standard normal kernel for estimation. The estimators were evaluated pointwise at each point in the grid which consists of 100 equispaced points ranging from the 20th percentile to the 80th percentile of the invariant density f_X defined in Assumption 2.3.

For each candidate estimator, we used the oracle bandwidth defined in (2.14). Note that all estimators have the same oracle bandwidths because they have the same asymptotic biases and variances. We used the cross-validation bandwidths defined in (2.16) for Avg, BPSs and BPDs estimators. We didn't use the cross-validation bandwidths for BPS and BPD due to the high computational cost. However, it will become evident from the simulation result using the oracle bandwidths, summarized in Table 2.2, that BPS and BPD have much larger mean squared errors than those of Avg, BPSs and BPDs. When calculating the cross validation bandwidth defined in (2.16) for Avg, BPSs and BPDs, we set $H = 150$ by the observation that, for most sample paths, the empirical autocorrelation functions of the averaged and the subsampled data reached zero before the time lag reaches 150. In addition, for BPSs and BPDs, we used Y's instead of $\bar{Y}$'s in the target data value of (2.15). In other words, the $\widehat{PE}$ for BPSs and BPDs using the subsampled data $\{Y_{jr\Delta}\}_{j=1}^{m}$ is defined by

$$\widehat{PE}(h\,;\,H) \equiv \sum_{j=1}^{m-1} \left(\hat{\mu}_{BP}^{(j,H)}\left(Y_{jr\Delta}\right) - \frac{Y_{(j+1)r\Delta} - Y_{jr\Delta}}{r\Delta} \right)^2,$$

where $\hat{\mu}_{BP}^{(j,H)}$ is either the BPSs or the BPDs estimator. The superscript (j, H) has the same meaning as that of $\hat{\mu}_{\bar{Y}}^{(k,H)}\left(\bar{Y}_k^{r,\Delta}\right)$ in (2.15) for the Avg estimator.

In order to solve the minimization problem of (2.16), for each sample path, we first calculated $D = \max_i Y_{i\Delta} - \min_i Y_{i\Delta}$ where $Y_{i\Delta}$ is defined in (2.3), and we found the local minimum by looking at bandwidths in the grid of length 30, $\{D/30,\ 2D/30, \ldots,\ D\}$. If there were multiple bandwidths that attain local minima, we took the the largest bandwidth, which is a

common practice when using cross-validation. In our simulation result, a grid of 30 values was fine enough to detect the local minima. We obtained an interior minimizer of $\widehat{PE}(\cdot\ ;\ H)$ for Avg and BPSs estimators for every sample path we generated. Figure 2.3 contains density plots of the selected bandwidths. For the sample paths generated by model (2.18), the average value of D was 0.176. For model (2.19), the average value was 5.41. These values are reflected in the scales of the horizontal axes in Figure 2.3.

For BPDs, recall that we need to choose both h and l, where l is defined in (2.11). We chose to minimize $\widehat{PE}(\cdot\ ;\ H)$ with respect to h with the restriction that $l = h$. This is motivated by the fact that, as Bandi and Phillips (2003) noted, choosing $\{l_n\}$, the bandwidth sequence of l according to the sample size n, so that $h_n/l_n \to C > 0$ yields smaller asymptotic variance for the double-smoothing estimator than that of the single-smoothing estimator (Bandi and Phillips, 2003, Remark 5). In our simulation study, for some sample paths, the curve $h \mapsto \widehat{PE}(h\ ;\ H)$ for the BPDs estimator evaluated at the grid of the bandwidths $\{D/30,\ 2D/30, \ldots,\ D\}$ was monotonically decreasing in h. In this case, we picked D as the bandwidth. We see this in Figure 2.3 by the additional modes on the right side of the density plot of the BPDs bandwidths. Since D is random, the bandwidths equal to D form a smooth mode in the density plots. The number of h's whose values were set to D was 365 for model (2.18) and 214 for model (2.19). Whenever $h = D$, the BPDs function estimate was a constant function. There was very little variation in the intercept across such constant function estimates. For example, the standard deviation of the intercepts for model (2.19) was 0.07, which is small considering that the drift coefficient of model (2.19) ranges from -1 to 1 at our evaluation points.

Now we present the simulation results. Table 2.2 summarizes the estimated expected integrated squared errors (ISE) of the estimators. For each combination of the model, the estimator and the bandwidth choice method, we approximated the ISE for each sample path by an invariant-density-weighted sum of the squared errors over the equispaced grid of length 100 on which the estimators were evaluated, and we provided the mean of the 1,000 ISEs along with the standard error of the mean in Table 2.2.

According to the table, if we use the oracle bandwidth, our Avg estimator has a smaller mean ISE than any other listed estimators except for the BPDs estimator. If we use the cross-validation bandwidth, our Avg estimator has smaller mean ISE than the BPSs estimator and

	ISE, Model (2.18)		ISE, Model (2.19)	
Estimator	Oracle	CV	Oracle	CV
BPS	1.474 (0.048)	—	94.8 (2.1)	—
BPD	0.628 (0.022)	—	50.9 (1.5)	—
BPSs	0.479 (0.012)	0.187 (0.010)	48.9 (1.1)	27.84 (0.8)
BPDs	0.194 (0.016)	0.283 (0.011)	27.2 (0.8)	22.67 (0.8)
Avg	0.327 (0.008)	0.138 (0.006)	38.0 (1.0)	24.37 (0.7)

Table 2.2: Means (and standard errors, i.e. standard deviations/$\sqrt{1000}$) of the integrated squared errors (ISEs) of candidate estimators over 1,000 sample paths. Labels "BPS" and "BPD" stand for the single-smoothing and the double-smoothing estimators of Bandi and Phillips (2003), respectively. Label "Avg" stands for the pre-averaging estimator. The "s" after a label means the estimator is combined with the subsampling method.

has smaller mean ISE than BPDs for model (2.18) and larger for model (2.19). Except for BPDs in model (2.18), the cross-validation bandwidths have smaller mean ISEs than the oracle bandwidths.

Figures 2.4 to 2.7 depict the pointwise mean squared errors (MSEs), i.e. the means of the 1,000 pointwise squared errors, for each estimator over the grid of evaluation points, for both bandwidths and for both models (2.18) and (2.19). According to the MSE plots, an estimator which has smaller mean ISE than another estimator in Table 2.2 tends to have smaller pointwise MSEs at almost all evaluation points.

We see that the MSE is small around $x = 0.072$ in Figures 2.4 and 2.5 and around $x = -0.2$ and $x = 1$ in Figures 2.6 and 2.7. To understand why the MSEs are small around these points, in the asymptotic bias and variance defined in (2.13), we set h equal to the oracle bandwidth $h_{opt}(x)$, defined in (2.14). Then we obtain that

$$\text{AMSE}(x) \propto \frac{\sigma^{8/5}(x)\Gamma_\mu^{2/5}(x)}{f_X^{4/5}(x)}.$$

Note that $\Gamma_\mu^{2/5}$ is nonnegative as Γ_μ^2 is nonnegative. It follows that the points where the MSE is small correspond to points where $f_X(x)$ is large or $\Gamma_\mu(x)$ is small. In Figures 2.4 and 2.5, around $x = 0.072$, $\Gamma_\mu(x)$ equals zero and f_X attains its maximum. In Figures 2.6 and 2.7, $\Gamma_\mu(x)$ equals zero around $x = -0.2$ and f_X attains its maximum around $x = 1$.

We can rewrite $\Gamma_\mu(x)$ using the analytical form of $f_X(x)$ in (2.4), calculating $f_X'(x)$ and

simplifying the expression for $\Gamma_\mu(x)$ to

$$\Gamma_\mu(x) = \nu_2 \times \left(2\mu'(x) \times \frac{\mu(x)/\sigma(x) - \sigma'(x)}{\sigma(x)} + \frac{1}{2}\mu''(x) \right).$$

From this expression, we can see that $\Gamma_\mu(x) = 0$ whenever $4\mu'(x) \times (\mu(x) - \sigma(x)\sigma'(x)) + \mu''(x)\sigma^2(x) = 0$. In particular, if the drift coefficient μ is linear so that μ' is constant and μ'' is zero, then $\Gamma_\mu(x) = 0$ if and only if $\mu(x) = \sigma(x)\sigma'(x)$. This condition is equivalent to the condition that $f'_X(x)$ is zero.

Considering the oracle bandwidth, we see from the bottom panels of Figures 2.4 and 2.6 and the Γ_μ curve in Figure 2.9 that $h_{opt}(x)$ is very large when $\Gamma_\mu(x)$ is close to zero. Recall that the asymptotic bias obtained from Theorem 2.1 is equal to $h^2\Gamma_\mu(x)$, as in (2.13). That $\Gamma_\mu(x)$ is close to zero means that the asymptotic bias is very small, which means we choose large h in order to reduce the asymptotic variance without suffering much from the increase in the asymptotic bias.

One may notice that the MSE curve slightly increases at the point where $h_{opt}(x)$ attains its maximum. The point corresponds to the one at which $\Gamma_\mu(x)$, and hence the asymptotic bias, is nearly zero. When $\Gamma_\mu(x)$ is nearly zero, if we choose a very large h, the asymptotic variance and hence the AMSE are close to zero. However, the finite-sample bias is not necessarily zero even if the asymptotic bias is zero. Therefore, unlike the AMSE, the MSE curve calculated from the simulations does not equal to zero. In fact, we see a slight increase in MSE for a very high choice of the bandwidth.

Next, we present a simulation result that indicates the bias-variance tradeoff of the block size r, where r is defined in Equation (2.3). Figure 2.8 depicts pointwise squared bias and variance of the Avg estimator with the oracle bandwidth under different values of r, namely $r = 2, 5, 10, 20, 40, 60$, for model (2.19). It is clear from Figure 2.8 that the squared bias is increasing in r and the variance is decreasing in r. Note that the sharp increase in the bias and the sharp decrease in the variance at $x \approx -0.2$ is due to a very high value of the oracle bandwidth, h_{opt}.

We provide the plots only for (2.19) because we can clearly see by plots the bias-variance tradeoff of r for (2.19) as its drift coefficient is nonlinear. We obtained similar results for (2.18),

that the bias is increasing in r and that the variance is decreasing in r.

Returning to Table 2.2, it indicates whether an estimator has smaller ISE than another estimator "on average". To make "pathwise" comparisons of ISEs, we construct Figures 2.10 to 2.17. Each plot is a scatterplot of 1,000 points, each point representing a pair of pathwise ISEs of the two estimators indicated in the plot. For instance, Figure 2.10 is a scatterplot of 1,000 pairs of pathwise ISEs of the Avg and the BPSs estimators, where each pair (i.e. each point in the scatterplot) corresponds to each of 1,000 sample paths of model (2.18). According to the figures, the pathwise comparisons give conclusions that are consistent with Table 2.2, in other words, an estimator which has smaller mean ISE than another estimator tends to have smaller pathwise ISEs. From Figures 2.10, 2.12, 2.14 and 2.16, the Avg estimator has smaller pathwise ISE than BPSs for, respectively, 825, 513, 679 and 549 out of 1,000 sample paths. From Figures 2.11, 2.13, 2.15 and 2.17, the ISE is smaller than BPDs for, respectively, 267, 782, 282 and 464 out of 1,000 sample paths.

We can make similar statements about the oracle and the cross-validation bandwidth, that is, the cross-validation bandwidth tends to have smaller pathwise ISEs than the oracle bandwidth. As an example, Figures 2.18 and 2.19 compare pathwise ISEs of the Avg estimator with the oracle and the cross-validation bandwidths for models (2.18) and (2.19). From Figures 2.18 and 2.19, the Avg estimator with the cross-validation bandwidth has smaller pathwise ISEs than the estimator with the oracle bandwidth for, respectively, 741 and 852 out of 1,000 sample paths.

One may notice that some points in Figures 2.13 and 2.17 form straight horizontal lines in the plots. Those points correspond to sample paths for which the bandwidth h for the BPDs estimator was equal to $D = \max_i Y_{i\Delta} - \min_i Y_{i\Delta}$, where $Y_{i\Delta}$ is defined in (2.3) (recall the discussion about the additional modes of the density plot of the BPDs bandwidths in Figure 2.3). We mentioned earlier that, when we chose $h = D$, the resulting function estimate was a constant function estimate with little variation in its intercept across such constant function estimates. Therefore, when we take D as the bandwidth, the pathwise ISE of the BPDs estimator has little variation across the sample paths. Hence the points form a straight horizontal line, where the values of pathwise ISEs of the BPDs estimator are almost fixed at some level at the Y-axis.

2.6 Proof of Theorem 2.1

In this section, we provide a full proof of Theorem 2.1. In order to clarify the argument, we first, in Section 2.6.1, give a overall scheme of the proof and derive key statements that are sufficient to prove Theorem 2.1. Then we prove those key statements in Sections 2.6.2 and 2.6.3.

2.6.1 Structure of the proof

Bandi and Phillips (2003) showed that their single-smoothing estimator, defined in (2.12), is a consistent and asymptotically normal estimator of μ when we observe a sample path of a recurrent solution process $\{X_t\}$ sampled discretely in time and without measurement error. We prove Theorem 2.1 by showing, under the conditions of this theorem, that the difference between our estimator computed from the $Y_{i\Delta}$'s and the single-smoothing estimator of Bandi and Phillips (2003) computed from the $X_{i\Delta}$'s converges to 0 in probability. This subsection presents the overall scheme of this proof.

To begin with, we introduce the single-smoothing estimator of Bandi and Phillips (2003) more explicitly and state their results of its consistency and asymptotic normality. For notational convenience, in order to relate the consistency and asymptotic normality results for their estimator to our estimator, we present the estimator of Bandi and Phillips (2003) when the observation time lag is $r\Delta$ and the number of observations is $m-1$, that is, when we observe $\{X_{(j-1)r_n\Delta_n}\}_{j=1}^{m-1}$.

Definition 2.3 *The single-smoothing estimator $\hat{\mu}_X$ of the drift coefficient μ proposed by Bandi and Phillips (2003) is defined by*

$$\begin{aligned}\hat{\mu}_X(x) &\equiv \frac{\frac{1}{m-2}\sum_{j=1}^{m-2}\frac{X_{jr\Delta}-X_{(j-1)r\Delta}}{r\Delta}\frac{1}{h}K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)}{\frac{1}{m-2}\sum_{j=1}^{m-2}\frac{1}{h}K\left(\frac{X_{(j-1)r\Delta}-x}{h}\right)} \\ &\equiv \frac{\mathcal{N}_{X_0,\ldots,X_{(m-2)r\Delta}}(x)}{\mathcal{D}_{X_0,\ldots,X_{(m-3)r\Delta}}(x)} \equiv \frac{\mathcal{N}_X(x)}{\mathcal{D}_X(x)}.\end{aligned}$$

We state the consistency and asymptotic normality result of $\hat{\mu}_X(x)$ not for a recurrent diffusion process but for a positive recurrent diffusion process, in order to relate the result to our estimator.

Theorem 2.2 (Bandi and Phillips, 2003) *Suppose that*

(i) $\Delta_n \to \infty$ *and* $n\Delta_n \to \infty$ *as* $n \to \infty$,

(ii) Assumption 2.2 holds,

(iii) The solution process $\{X_t\}$ *is positive recurrent (so that the stationary density* f_X *exists), and*

(iv) The kernel K satisfies Assumption 2.4 except for our additional requirement that K' *is bounded.*

In addition, suppose

$$\left(\frac{n\Delta_n}{h_n}\right)^2 r_n\Delta_n \ln(1/r_n\Delta_n) = o(1) \quad \text{and} \quad n\Delta_n h_n \to \infty.$$

Then the following consistency and asymptotic normality results hold whenever $f_X(x) > 0$.

1. $\hat{\mu}_X(x) \longrightarrow \mu(x)$ *almost surely as* $n \to \infty$.

2. *If* $n\Delta_n h_n^5 = o(1)$, *then*

$$\sqrt{n\Delta_n h_n}\ \{\hat{\mu}_X(x) - \mu(x)\} \xrightarrow{d} N\left(0, K_2\frac{\sigma^2(x)}{f_X(x)}\right),$$

   where K_2 *is as in Theorem 2.1.*

3. *If* $n\Delta_n h_n^5 = O(1)$, *then*

$$\sqrt{n\Delta_n h_n}\ \{\hat{\mu}_X(x) - \mu(x) - h_n^2\Gamma_\mu(x)\} \xrightarrow{d} N\left(0, K_2\frac{\sigma^2(x)}{f_X(x)}\right),$$

   where $\Gamma_\mu(x)$ *is as in Theorem 2.1.*

Note that Assumptions 2.1 and 2.3 include conditions (i) and (iii) of Theorem 2.2, respectively. Thus Theorem 2.2 holds under Assumptions 2.1 to 2.4 as well.

We will prove consistency and asymptotic normality of our estimator $\hat{\mu}_{\bar{Y}}(x)$ based on Theorem 2.2 and the following theorem.

Theorem 2.3 (Bandi and Phillips, 2003) *Suppose conditions* $(i) - (iv)$ *of Theorem 2.2 hold, and suppose*

$$\left(\frac{1}{h_n}\right)^2 r_n\Delta_n \ln(1/r_n\Delta_n) = o(1).$$

Then, for each x such that $f_X(x) > 0$, we have

$$\mathcal{D}_X(x) \longrightarrow f_X(x) \quad \textit{almost surely,}$$

where $f_X(x)$ is as in Assumption 2.3.

In what follows, we will prove that the following statements hold under Assumptions 2.1 to 2.5 in Section 2.2 and conditions (i), (ii) and (iii) of Theorem 2.1. Recall the definition of $\hat{\mu}_{\bar{Y}}(x) = \mathcal{N}_{\bar{Y}}(x)/\mathcal{D}_{\bar{Y}}(x)$, in Definition 2.2.

$$\mathcal{D}_{\bar{Y}}(x) - \mathcal{D}_X(x) = o_p(1) \quad \text{and} \tag{2.20}$$

$$\sqrt{n\Delta_n h_n}\ \{\mathcal{N}_{\bar{Y}}(x) - \mathcal{N}_X(x)\} = o_p(1). \tag{2.21}$$

Then Theorem 2.1 follows from these statements and Theorems 2.2 and 2.3. First, since $n\Delta_n h_n \to \infty$ as in condition (ii) of Theorem 2.1, Equation (2.21) implies

$$\mathcal{N}_{\bar{Y}}(x) - \mathcal{N}_X(x) = o_p(1). \tag{2.22}$$

Then (2.20) and (2.22) along with Theorems 2.2 and 2.3 imply Theorem 2.1's conclusion 1 since, for every $x \in \{y \mid f_X(y) > 0\}$,

$$\hat{\mu}_{\bar{Y}}(x) = \frac{\mathcal{N}_{\bar{Y}}(x)}{\mathcal{D}_{\bar{Y}}(x)} = \frac{\mathcal{N}_X(x) + o_p(1)}{\mathcal{D}_X(x) + o_p(1)} \xrightarrow{p} \mu(x).$$

To prove Theorem 2.1's conclusion 2, where $n\Delta_n h_n^5 = o(1)$, we use (2.20) and (2.21) along with Theorems 2.2 and 2.3 to write

$$\begin{aligned}
\sqrt{n\Delta_n h_n}\ \{\hat{\mu}_{\bar{Y}}(x) - \mu(x)\} &= \frac{\sqrt{n\Delta_n h_n}\,\mathcal{N}_{\bar{Y}}(x)}{\mathcal{D}_{\bar{Y}}(x)} - \sqrt{n\Delta_n h_n}\,\mu(x) \\
&= \frac{\sqrt{n\Delta_n h_n}\,\mathcal{N}_X(x) + o_p(1)}{\mathcal{D}_X(x) + o_p(1)} - \sqrt{n\Delta_n h_n}\,\mu(x) \\
&= \sqrt{n\Delta_n h_n}\left\{\frac{\mathcal{N}_X(x) + o_p(1)}{\mathcal{D}_X(x) + o_p(1)} - \mu(x)\right\} \\
&\xrightarrow{d} N\left(0, K_2\frac{\sigma^2(x)}{f_X(x)}\right),
\end{aligned}$$

by Theorem 2.2's conclusion 2. We can prove Theorem 2.1's conclusion 3 similarly.

In summary, if we prove (2.20) and (2.21) under Assumptions 2.1 to 2.5 and conditions (i), (ii), (iii) of Theorem 2.1, then the conclusions of Theorem 2.1 follow. In the remainder of this section, we prove (2.20) and (2.21). The proof uses the following preliminary lemmas.

2.6.2 Preliminary lemmas

Lemma 2.1 *The following hold.*

1. *Under Assumption 2.4, K is globally Lipschitz, that is, there exists a finite constant $M > 0$ such that $|K(x) - K(y)| \leq M|x - y|$ for all $x, y \in \mathbb{R}$. In addition, we can take $M \equiv \sup_x K'(x)$.*

2. *Under Assumption 2.5, the following inequalities hold for all j and n:*

$$\left(\mathbb{E}(|\bar{\varepsilon}_j^{r_n,\Delta_n}|)\right)^2 \leq \mathbb{E}([\bar{\varepsilon}_j^{r_n,\Delta_n}]^2) \leq \frac{\sigma_\varepsilon^2}{r_n}.$$

Proof :

1. The conclusion follows directly from the fact that K is continuously differentiable and that K' is bounded.

2. The first inequality is from the fact that the L^1 norm (that is, expectation of the absolute value) is bounded by the L^2 norm. For the second inequality, since $\mathbb{E}(\bar{\varepsilon}_j^{r_n,\Delta_n}) = 0$, we have $\mathbb{E}([\bar{\varepsilon}_j^{r_n,\Delta_n}]^2) = \mathrm{Var}(\bar{\varepsilon}_j^{r_n,\Delta_n})$. Then independence of ε_t's and the boundedness of $\mathrm{Var}(\varepsilon_t)$ allow us to write

$$\mathrm{Var}(\bar{\varepsilon}_j^{r_n,\Delta_n}) = \frac{1}{r_n^2}\sum_{i=1}^{r_n}\mathrm{Var}(\varepsilon_{[(j-1)r_n+i]\Delta_n}) \leq \frac{\sigma_\varepsilon^2}{r_n}. \quad \blacksquare$$

Lemma 2.2 *Under Assumptions 2.2 and 2.3, for any real $0 \leq a < b$,*

$$\begin{aligned}
\left\{\mathbb{E}\left(\left|\int_a^b \mu(X_s)ds\right|\right)\right\}^2 &\leq \mathbb{E}\left(\left[\int_a^b \mu(X_s)ds\right]^2\right) \leq \mathbb{E}(\mu^2(X_0))\,(b-a)^2, \\
\left\{\mathbb{E}\left(\left|\int_a^b \sigma(X_s)dW_s\right|\right)\right\}^2 &\leq \mathbb{E}\left(\left[\int_a^b \sigma(X_s)dW_s\right]^2\right) \leq \mathbb{E}(\sigma^2(X_0))\,(b-a).
\end{aligned}$$

Proof : The first inequality of each line is by the fact that the L^1 norm is bounded by the L^2 norm. We now prove the second inequality of the first line. By applying Hölder's inequality to $|\mu| \times 1$,

$$\left[\int_a^b \mu(X_s)ds\right]^2 \leq \left[\int_a^b |\mu(X_s)|ds\right]^2 \leq \int_a^b \mu^2(X_s)ds \int_a^b ds = (b-a)\int_a^b \mu^2(X_s)ds.$$

Therefore,

$$\mathbb{E}\left(\left[\int_a^b \mu(X_s)ds\right]^2\right) \;\leq\; (b-a)\mathbb{E}\left(\int_a^b \mu^2(X_s)ds\right) \;=\; (b-a)\int_a^b \mathbb{E}(\mu^2(X_s))ds.$$

In addition, since $\mathbb{E}(\mu^2(X_s)) = \mathbb{E}(\mu^2(X_0))$ as $\{X_t\}$ is stationary,

$$(b-a)\int_a^b \mathbb{E}(\mu^2(X_s))ds \;=\; (b-a)\int_a^b \mathbb{E}(\mu^2(X_0))ds \;=\; \mathbb{E}(\mu^2(X_0))(b-a)^2.$$

This proves the second inequality of the first line. We can use the same reasoning to prove the second inequality of the second line if we first use the following property of stochastic integration:

$$\mathbb{E}\left(\left[\int_a^b \sigma(X_s)dW_s\right]^2\right) = \mathbb{E}\left(\int_a^b \sigma^2(X_s)ds\right). \quad \blacksquare$$

Remark : We denote

$$\mathcal{M}_a^b \equiv \int_{a\Delta}^{b\Delta} \mu(X_s)ds \quad \text{and} \quad \mathcal{W}_a^b \equiv \int_{a\Delta}^{b\Delta} \sigma(X_s)dW_s. \tag{2.23}$$

Then Lemma 2.2 can be rewritten as $\mathbb{E}\left(\left(\mathcal{M}_a^b\right)^2\right) \leq \mathbb{E}(\mu^2(X_0))\ (b-a)^2\Delta^2$ and $\mathbb{E}\left(\left(\mathcal{W}_a^b\right)^2\right) \leq \mathbb{E}(\sigma^2(X_0))\ (b-a)\Delta$.

Lemma 2.3 *Suppose that Assumptions 2.1 to 2.3 hold. Define*

$$\kappa_n \equiv \max_{j \leq m_n} \sup_{(j-1)r_n\Delta_n \leq s \leq jr_n\Delta_n} |X_s - X_{(j-1)r_n\Delta_n}|$$

and

$$\gamma_n \equiv \max_{j \leq m_n} \mathbb{E}\left(\left(\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n}\right)^2\right).$$

Then the following hold.

(i) *For any nonnegative numerical sequence* $\{a_n\}$ *such that* $\lim_{n\to\infty} a_n(r_n\Delta_n \ln(1/r_n\Delta_n))^{1/2} = 0$, $\lim_{n\to\infty} a_n\kappa_n = 0$ *a.s.*

(ii) $\max_{j\le m_n} |\bar{X}_j^{r_n\Delta_n} - X_{(j-1)r_n\Delta_n}| \le \kappa_n$.

(iii) $\max_{j\le m_n} |\bar{X}_{j+1}^{r_n\Delta_n} - \bar{X}_j^{r_n\Delta_n}| \le 3\kappa_n$.

(iv) *There exists a finite constant* β *such that* $\gamma_n \le \beta r_n\Delta_n$.

(v) $\max_{j\le m_n} \mathbb{E}(|\bar{X}_j^{r_n\Delta_n} - X_{(j-1)r_n\Delta_n}|) \le \sqrt{\gamma_n}$.

(vi) $\max_{j\le m_n} \mathbb{E}(|\bar{X}_{j+1}^{r_n\Delta_n} - \bar{X}_j^{r_n\Delta_n}|) \le 3\sqrt{\gamma_n}$.

Proof : We first prove (i). As Bandi and Phillips (2003, page 267) point out, by Levy's modulus of continuity of diffusions, we have

$$\mathbb{P}\Big(\limsup_{n\to\infty} \frac{\kappa_n}{(r_n\Delta_n \ln(1/r_n\Delta_n))^{1/2}} = C\Big) = 1$$

where C is a suitable constant (Karatzas and Shreve, 1991, Theorem 9.25, Chapter 2, page 114). Therefore, we have, for a nonnegative sequence a_n such that $a_n(r_n\Delta_n \ln(1/r_n\Delta_n))^{1/2} \to 0$,

$$\mathbb{P}\Big(\limsup_{n\to\infty} a_n\kappa_n = 0\Big) = 1.$$

Since both a_n and κ_n are nonnegative, (i) follows. Next, to prove (ii), it suffices to notice that, for any j,

$$|\bar{X}_j^{r_n\Delta_n} - X_{(j-1)r_n\Delta_n}| \le \frac{1}{r_n}\sum_{i=1}^{r_n} |X_{(j-1)r_n\Delta_n + i\Delta_n} - X_{(j-1)r_n\Delta_n}| \;\le\; \frac{1}{r_n}\sum_{i=1}^{r_n} \kappa_n = \kappa_n.$$

We can prove (iii) using (ii) and the definition of κ_n as follows: for any j,

$$|\bar{X}_{j+1}^{r_n\Delta_n} - \bar{X}_j^{r_n\Delta_n}| \le |\bar{X}_{j+1}^{r_n\Delta_n} - X_{jr_n\Delta_n}| + |X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}| + |\bar{X}_j^{r_n\Delta_n} - X_{(j-1)r_n\Delta_n}| \le 3\kappa_n.$$

Now we prove (iv). Note first that, by (2.2), we have the following on a set of probability

1:

$$\begin{aligned}\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n} &= \frac{1}{r_n}\sum_{i=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+i\Delta_n}\mu(X_s)ds + \frac{1}{r_n}\sum_{i=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+i\Delta_n}\sigma(X_s)dW_s \\ &\equiv A + B. \end{aligned} \tag{2.24}$$

Then we have $\mathbb{E}\left((\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n})^2\right) \leq 2\mathbb{E}(A^2) + 2\mathbb{E}(B^2)$, for A and B defined in (2.24), since the above equality holds almost surely and since $(A+B)^2 \leq 2A^2 + 2B^2$. We now expand A^2:

$$\left(\frac{1}{r_n}\sum_{i=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+i\Delta_n}\mu(X_s)ds\right)^2 = \frac{1}{r_n^2}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+i\Delta_n}\mu(X_s)ds\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+k\Delta_n}\mu(X_s)ds. \tag{2.25}$$

We use the Cauchy-Schwarz inequality, Lemma 2.2 and the fact that $i, k \leq r_n$ to bound the expectation of the absolute value of the ik^{th} summand of (2.25) by

$$\sqrt{\mathbb{E}\left(\left[\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+i\Delta_n}\mu(X_s)ds\right]^2\right)}\sqrt{\mathbb{E}\left(\left[\int_{(j-1)r_n\Delta_n}^{(j-1)r_n\Delta_n+k\Delta_n}\mu(X_s)ds\right]^2\right)} \;\leq\; \mathbb{E}(\mu^2(X_0))r_n^2\Delta_n^2.$$

This bound is uniform in i and k, so $\mathbb{E}(A^2)$ is bounded by $\mathbb{E}(\mu^2(X_0))r_n^2\Delta_n^2$. In exactly the same way, we can bound $\mathbb{E}(B^2)$ by $\mathbb{E}(\sigma^2(X_0))r_n\Delta_n$. Then $\mathbb{E}\left((\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n})^2\right)$ is bounded by

$$2\mathbb{E}(\mu^2(X_0))r_n^2\Delta_n^2 + 2\mathbb{E}(\sigma^2(X_0))r_n\Delta_n \;\leq\; \beta r_n\Delta_n$$

for some constant β since $\mathbb{E}(\mu^2(X_0))$ and $\mathbb{E}(\sigma^2(X_0))$ are finite and $r_n\Delta_n \to 0$. This bound is uniform in j, so (iv) follows.

Next, (v) follows directly from the fact that the L^1 norm is bounded by the L^2 norm and that the square root function $f(x) = \sqrt{x}$ is monotonically increasing. Lastly, we can prove (vi) using (v), the argument of (ii) and the argument in the proof of (iv) to bound $\mathbb{E}(|X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}|)$ by $\sqrt{\gamma_n}$. ■

Now we are ready to present the proof of (2.20) and (2.21).

2.6.3 Proof of Equation (2.20)

Suppose Assumptions 2.1 to 2.5 and conditions (i), (ii), (iii) of Theorem 2.1 hold. Instead of proving (2.20), we prove the stronger statement, that

$$\mathcal{D}_{\bar{Y}}(x) - \mathcal{D}_X(x) = o_p\left(\frac{1}{\sqrt{n\Delta_n h_n}}\right).$$

First, the Lipschitz continuity of the kernel K implies

$$\begin{aligned} |\mathcal{D}_{\bar{Y}}(x) - \mathcal{D}_X(x)| &\leq \frac{1}{m_n - 2}\sum_{j=1}^{m_n-2}\frac{1}{h_n}\left|K\left(\frac{\bar{Y}_j^{r_n\Delta_n} - x}{h_n}\right) - K\left(\frac{X_{(j-1)r_n\Delta_n} - x}{h_n}\right)\right| \\ &\leq \frac{M}{m_n - 2}\sum_{j=1}^{m_n-2}\frac{|\bar{Y}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n}|}{h_n^2}, \end{aligned} \tag{2.26}$$

where M is as in Lemma 2.1. In addition, using the definition of $\bar{Y}_j^{r_n,\Delta_n}$ and (ii) of Lemma 2.3, we have

$$|\bar{Y}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n}| = |\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n} + \bar{\varepsilon}_j^{r_n,\Delta_n}| \leq \kappa_n + |\bar{\varepsilon}_j^{r_n,\Delta_n}|. \tag{2.27}$$

Combining (2.26) and (2.27), we can bound $|\mathcal{D}_{\bar{Y}}(x) - \mathcal{D}_X(x)|$ by

$$|\mathcal{D}_{\bar{Y}}(x) - \mathcal{D}_X(x)| \leq \frac{M\kappa_n}{h_n^2} + \frac{M}{m_n - 2}\sum_{j=1}^{m_n-2}\frac{|\bar{\varepsilon}_j^{r_n,\Delta_n}|}{h_n^2}.$$

We now study the orders of the two terms of this bound. We show that

$$\frac{\kappa_n}{h_n^2} = o_{a.s.}\left(\frac{1}{n\Delta_n h_n}\right) \qquad \text{and} \qquad \frac{1}{m_n - 2}\sum_{j=1}^{m_n-2}\frac{|\bar{\varepsilon}_j^{r_n,\Delta_n}|}{h_n^2} = o_p\left(\frac{1}{\sqrt{n\Delta_n h_n}}\right), \tag{2.28}$$

which will complete the proof since $n\Delta_n h_n \to \infty$ as in condition (ii) of Theorem 2.1. For the first claim of (2.28), we use condition (i) of Theorem 2.1, which says $(n\Delta_n/h_n) \times (r_n\Delta_n \ln(r_n\Delta_n))$ $\to 0$. This condition together with (i) of Lemma 2.3 yields $n\Delta_n h_n \times \kappa_n/h_n^2 = (n\Delta_n/h_n) \times \kappa_n = o_{a.s.}(1)$. Then this implies that the first term is $o_{a.s.}(1/(n\Delta_n h_n))$, since $n\Delta_n h_n \to \infty$ by condition (ii) of Theorem 2.1.

For the second claim of (2.28), since $\mathbb{E}(|\bar{\varepsilon}_j^{r_n,\Delta_n}|) \leq \sigma_\varepsilon/\sqrt{r_n}$ for all j by Lemma 2.2, we can

bound its expectation by $\sigma_\varepsilon/(h_n^2 r_n^{1/2})$, which is $o(1/\sqrt{n\Delta_n h_n})$ since

$$n\Delta_n h_n \times \frac{1}{h_n^4 r_n} = \frac{n}{h_n^3 r_n^2} \times r_n\Delta_n = o(1) \times o(1)$$

by condition (iii) of Theorem 2.1 and the assumption that $r_n\Delta_n \to 0$ (see Assumption 2.1). Therefore, the L^1 norm of the second term is $o(1/\sqrt{n\Delta_n h_n})$, implying that it is $o_p(1/\sqrt{n\Delta_n h_n})$.

■

2.6.4 Proof of Equation (2.21)

In order to prove (2.21) under the conditions, we first define the following 3 terms. Recall the definition of $\mathcal{N}_{\bar{Y}}(x)$ in Definition 2.2 and $\mathcal{N}_X$ in Definition 2.3.

$$\begin{aligned}
\mathcal{N}_{\bar{Y},\bar{X}}(j) &\equiv \left(\bar{Y}_{j+2}^{r_n,\Delta_n} - \bar{Y}_{j+1}^{r_n,\Delta_n}\right) K\left(\frac{\bar{Y}_j^{r_n,\Delta_n} - x}{h_n}\right) - \left(\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}\right) K\left(\frac{\bar{X}_j^{r_n,\Delta_n} - x}{h_n}\right), \\
\mathcal{N}_{\bar{X},X}(j) &\equiv \left(\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}\right) K\left(\frac{\bar{X}_j^{r_n,\Delta_n} - x}{h_n}\right) - \left(X_{(j+1)r_n\Delta_n} - X_{jr_n\Delta_n}\right) K\left(\frac{X_{(j-1)r_n\Delta_n} - x}{h_n}\right), \\
\mathcal{N}_{X,X}(j) &\equiv \left(X_{(j+1)r_n\Delta_n} - X_{jr_n\Delta_n}\right) K\left(\frac{X_{(j-1)r_n\Delta_n} - x}{h_n}\right) - \left(X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}\right) K\left(\frac{X_{(j-1)r_n\Delta_n} - x}{h_n}\right).
\end{aligned}$$

Then we have the following equality:

$$\sqrt{n\Delta_n h_n}\ \left\{\mathcal{N}_{\bar{Y}}(x) - \mathcal{N}_X(x)\right\} = \frac{\sqrt{n\Delta_n h_n}}{(m_n-2)r_n\Delta_n h_n} \sum_{j=1}^{m_n-2} \left\{\mathcal{N}_{\bar{Y},\bar{X}}(j) + \mathcal{N}_{\bar{X},X}(j) + \mathcal{N}_{X,X}(j)\right\}.$$

Then, to prove (2.21), we note that

$$\frac{\sqrt{n\Delta_n h_n}}{(m_n-2)r_n\Delta_n h_n} = \frac{m_n}{m_n-2} \times \frac{1}{\sqrt{n\Delta_n h_n}},$$

and we show that each of

$$\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X}}(j) \quad , \quad \sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{X},X}(j) \quad \text{and} \quad \sum_{j=1}^{m_n-2} \mathcal{N}_{X,X}(j) \tag{2.29}$$

is $o_p\left(\sqrt{n\Delta_n h_n}\right)$. Sometimes we treat the three sums in (2.29) all at once using the notation $\mathcal{N}_{Z,W}(j)$. Note that $\mathcal{N}_{Z,W}(j)$ can be rewritten in the form

$$\begin{aligned}\mathcal{N}_{Z,W}(j) &= Z_j^{inc} \times K\left(\frac{Z_j^{kern}-x}{h_n}\right) - W_j^{inc} \times K\left(\frac{W_j^{kern}-x}{h_n}\right) \\ &= Z_j^{inc} \times \left[K\left(\frac{Z_j^{kern}-x}{h_n}\right) - K\left(\frac{W_j^{kern}-x}{h_n}\right)\right] + (Z_j^{inc}-W_j^{inc}) \times K\left(\frac{W_j^{kern}-x}{h_n}\right) \\ &\equiv \mathcal{N}_{Z,W,kern}(j) + \mathcal{N}_{Z,W,inc}(j). \end{aligned} \tag{2.30}$$

We shall call $\mathcal{N}_{Z,W,kern}(j)$ the kernel difference term and $\mathcal{N}_{Z,W,inc}(j)$ the increment difference term. To be precise, the following table lists the Z's and W's for each $\mathcal{N}_{Z,W}(j)$.

Table 2.3: The list of Z's and W's for each $\mathcal{N}(j)$.

	Z_j^{inc}	W_j^{inc}	Z_j^{kern}	W_j^{kern}
$\mathcal{N}_{\bar{Y},\bar{X}}(j)$	$\bar{Y}_{j+2}^{r_n,\Delta_n} - \bar{Y}_{j+1}^{r_n,\Delta_n}$	$\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}$	$\bar{Y}_j^{r_n,\Delta_n}$	$\bar{X}_j^{r_n,\Delta_n}$
$\mathcal{N}_{\bar{X},X}(j)$	$\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}$	$X_{(j+1)r_n\Delta_n} - X_{jr_n\Delta_n}$	$\bar{X}_j^{r_n,\Delta_n}$	$X_{(j-1)r_n\Delta_n}$
$\mathcal{N}_{X,X}(j)$	$X_{(j+1)r_n\Delta_n} - X_{jr_n\Delta_n}$	$X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}$	$X_{(j-1)r_n\Delta_n}$	$X_{(j-1)r_n\Delta_n}$

Note that $Z_j^{kern} = W_j^{kern}$ for $\mathcal{N}_{X,X}(j)$, which means that $\mathcal{N}_{X,X,kern}(j)$ is equal to zero. We must show that the two kernel difference and the three increment difference terms, summed over j, are all $o_p\left(\sqrt{n\Delta_n h_n}\right)$. In what follows, we prove it. We first study the orders of the kernel difference terms, and then we study those of the increment difference terms. When studying each difference term, we study that of $\mathcal{N}_{\bar{Y},\bar{X}}$ first, and then we study that of $\mathcal{N}_{X,X}$ and that of $\mathcal{N}_{\bar{X},X}$. We proceed in this order because the study of the terms of $\mathcal{N}_{\bar{Y},\bar{X}}$ is simplest and the study of those of $\mathcal{N}_{\bar{X},X}$ is most delicate.

Study of the kernel difference terms

We first study the orders of the kernel difference terms, summed over j. First, we show that $\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X},kern}(j)$ is $o_p\left(\sqrt{n\Delta_n h_n}\right)$. We first use the Lipschitz continuity of K to bound its absolute value:

$$\left|\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X},kern}(j)\right| \leq \sum_{j=1}^{m_n-2} \frac{M}{h_n} |Z_j^{inc}| \, |Z_j^{kern} - W_j^{kern}|, \tag{2.31}$$

where the Z's and W's are as in Table 2.3 and M is as in Lemma 2.1. Then we take the expectation of the bound (2.31), and we use the independence of $\{X_t\}$ and $\{\varepsilon_t\}$ and the independence of ε_t's to bound the j^{th} summand of the right-hand side of (2.31) by

$$\begin{aligned}\frac{M}{h_n}\mathbb{E}\left(|\bar{Y}_{j+2}^{r_n,\Delta_n} - \bar{Y}_{j+1}^{r_n,\Delta_n}||\bar{\varepsilon}_j^{r_n,\Delta_n}|\right) &\leq \frac{M}{h_n}\left(\mathbb{E}(|\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}|) + \mathbb{E}(|\bar{\varepsilon}_{j+2}^{r_n,\Delta_n} - \bar{\varepsilon}_{j+1}^{r_n,\Delta_n}|)\right)\mathbb{E}(|\bar{\varepsilon}_j^{r_n,\Delta_n}|) \\ &\leq \frac{M}{h_n}\left(3\sqrt{\gamma_n} + \frac{2\sigma_\varepsilon}{\sqrt{r_n}}\right)\frac{\sigma_\varepsilon}{\sqrt{r_n}}, \qquad (2.32)\end{aligned}$$

where the last inequality used Lemma 2.1 and (vi) of Lemma 2.3. Then, since $\gamma_n \leq \beta r_n \Delta_n$ by (iv) of Lemma 2.3, the expression (2.32) is bounded further by a constant times

$$\frac{\sqrt{\Delta_n}}{h_n} + \frac{1}{r_n h_n}.$$

Therefore, $\mathbb{E}(|\sum_{j=1}^{m_n-2}\mathcal{N}_{\bar{Y},\bar{X},kern}(j)|)$ is bounded by a constant times

$$\frac{m_n\sqrt{\Delta_n}}{h_n} + \frac{m_n}{r_n h_n} = \frac{n\sqrt{\Delta_n}}{r_n h_n} + \frac{n}{r_n^2 h_n},$$

since $m_n = n/r_n$ as in Assumption 2.1. This bound is $o\left(\sqrt{n\Delta_n h_n}\right)$ since

$$\frac{1}{\sqrt{n\Delta_n h_n}} \times \frac{n\sqrt{\Delta_n}}{r_n h_n} = \sqrt{\frac{n}{r_n^2 h_n^3}} = o(1)$$

by condition (iii) of Theorem 2.1 and since

$$\frac{1}{\sqrt{n\Delta_n h_n}} \times \frac{n}{r_n^2 h_n} = \frac{n}{r_n^2 h_n^3} \times \frac{h_n^2}{\sqrt{n\Delta_n h_n}} = o(1) \times o(1)$$

by conditions (ii) and (iii) of Theorem 2.1 and since $h_n \to 0$ as in Assumption 2.1.

Next, we show that $\sum_{j=1}^{m_n-2}\mathcal{N}_{\bar{X},X,kern}(j)$, which is defined at the beginning of Section 2.6.4 and (2.30), is $o_p\left(\sqrt{n\Delta_n h_n}\right)$. We use the first order Taylor expansion of K and write $\mathcal{N}_{\bar{X},X,kern}(j)$

as follows:

$$\begin{aligned}\mathcal{N}_{\bar{X},X,kern}(j) &= Z_j^{inc} K'\left(\frac{\xi_j - x}{h_n}\right)\left(\frac{Z_j^{kern} - W_j^{kern}}{h_n}\right) \\ &= \left(\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}\right) K'\left(\frac{\xi_j - x}{h_n}\right)\left(\frac{\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n}}{h_n}\right), \end{aligned} \tag{2.33}$$

where ξ_j is a value between the values of $\bar{X}_j^{r_n,\Delta_n}$ and $X_{(j-1)r_n\Delta_n}$. We must show that the sum of (2.33) over j is $o_p\left(\sqrt{n\Delta_n h_n}\right)$. First, we use the definition of X_t in (2.2) that $X_t = X_0 + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dW_s$ a.s. to write the increments of (2.33) as

$$\begin{aligned}\bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n} &= \frac{1}{r_n}\sum_{i=1}^{r_n}\left(X_{[(j+1)r_n+i]\Delta_n} - X_{[jr_n+i]\Delta_n}\right) \\ &= \frac{1}{r_n}\sum_{i=1}^{r_n}\int_{[jr_n+i]\Delta_n}^{[(j+1)r_n+i]\Delta_n}\mu(X_s)ds + \frac{1}{r_n}\sum_{i=1}^{r_n}\int_{[jr_n+i]\Delta_n}^{[(j+1)r_n+i]\Delta_n}\sigma(X_s)dW_s \\ &= \frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{M}_{jr_n+i}^{(j+1)r_n+i} + \frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{W}_{jr_n+i}^{(j+1)r_n+i},\end{aligned}$$

and

$$\begin{aligned}\bar{X}_j^{r_n,\Delta_n} - X_{(j-1)r_n\Delta_n} &= \frac{1}{r_n}\sum_{k=1}^{r_n}\left(X_{[(j-1)r_n+k]\Delta_n} - X_{(j-1)r_n\Delta_n}\right) \\ &= \frac{1}{r_n}\sum_{k=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{[(j-1)r_n+k]\Delta_n}\mu(X_s)ds + \frac{1}{r_n}\sum_{k=1}^{r_n}\int_{(j-1)r_n\Delta_n}^{[(j-1)r_n+k]\Delta_n}\sigma(X_s)dW_s \\ &= \frac{1}{r_n}\sum_{k=1}^{r_n}\mathcal{M}_{(j-1)r_n}^{(j-1)r_n+k} + \frac{1}{r_n}\sum_{k=1}^{r_n}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}.\end{aligned}$$

Recall the definition of $\mathcal{M}$ and $\mathcal{W}$ in the Remark after Lemma 2.2. Using these, we expand (2.33) as follows:

$$(2.33) = \frac{1}{r_n^2}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\mathcal{M}_{jr_n+i}^{(j+1)r_n+i}\mathcal{M}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j - x}{h_n}\right) \tag{2.34}$$

$$+ \frac{1}{r_n^2}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\mathcal{M}_{jr_n+i}^{(j+1)r_n+i}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j - x}{h_n}\right) \tag{2.35}$$

$$+ \frac{1}{r_n^2}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\mathcal{M}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j - x}{h_n}\right) \tag{2.36}$$

$$+ \frac{1}{r_n^2}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j - x}{h_n}\right). \tag{2.37}$$

Throughout the calculations regarding (2.34) to (2.37), we will use the bounds based on the Remark after Lemma 2.2, that, for some constant C, the L^2 norms $\mathbb{E}\left(\left(\mathcal{M}_{jr_n+i}^{(j+1)r_n+i}\right)^2\right)$ and $\mathbb{E}\left(\left(\mathcal{M}_{(j-1)r_n}^{(j-1)r_n+k}\right)^2\right)$ are bounded by $Cr_n^2\Delta_n^2$ and that $\mathbb{E}\left(\left(\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right)^2\right)$ and $\mathbb{E}\left(\left(\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\right)^2\right)$ are bounded by $Cr_n\Delta_n$.

Note that, using boundedness of K' and the Cauchy-Schwarz inequality, we can bound the L^1 norm of the ik^{th} summand of (2.34) by

$$\sqrt{\mathbb{E}\left(\left[\mathcal{M}_{jr_n+i}^{(j+1)r_n+i}\right]^2\right)\mathbb{E}\left(\left[\mathcal{M}_{(j-1)r_n}^{(j-1)r_n+k}\right]^2\right)} \times \frac{M}{h_n} \;\leq\; CM\frac{r_n^2\Delta_n^2}{h_n}.$$

By the same reasoning, we can bound the L^1 norm of the ik^{th} summand of (2.35) and (2.36) by $CM(r_n\Delta_n)^{3/2}/h_n$. These bounds are uniform in i, j, k, so the L^1 norm of (2.34) to (2.36) summed over j is bounded by a constant times $m_n(r_n\Delta_n)^{3/2}/h_n = nr_n^{1/2}\Delta_n^{3/2}/h_n$ (recall $m_n = n/r_n$ in Assumption 2.1). In addition, by condition (i) of Theorem 2.1,

$$\frac{nr_n^{1/2}\Delta_n^{3/2}}{h_n} = \frac{n\Delta_n}{h_n}\sqrt{r_n\Delta_n} = o(1). \tag{2.38}$$

Therefore, the L^1 norm of (2.34) to (2.36) summed over j is $o(1)$. This implies that it is $o(\sqrt{n\Delta_n h_n})$ since $n\Delta_n h_n \to \infty$ by condition (ii) of Theorem 2.1.

Now it remains to show that (2.37) summed over j is $o_p\left(\sqrt{n\Delta_n h_n}\right)$. It requires a more delicate argument than (2.34) to (2.36). We first rearrange the sum of (2.37) over j as follows:

$$\frac{1}{r_n^2 h_n}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\sum_{j=1}^{m_n-2}\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\,K'\left(\frac{\xi_j - x}{h_n}\right). \tag{2.39}$$

We derive the bound of the squared L^2 norm of the ik^{th} summand of the above. We first prove that

$$\begin{aligned}
&\mathbb{E}\left(\left[\sum_{j=1}^{m_n-2}\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}K'\left(\frac{\xi_j - x}{h_n}\right)\right]^2\right)\\
=\;&\sum_{j=1}^{m_n-2}\mathbb{E}\left(\left[\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right]^2\left[\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\right]^2 K'^2\left(\frac{\xi_j - x}{h_n}\right)\right),
\end{aligned} \tag{2.40}$$

in other words, that

$$\mathbb{E}\left(\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j-x}{h_n}\right)\mathcal{W}_{lr_n+i}^{(l+1)r_n+i}\mathcal{W}_{(l-1)r_n}^{(l-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_l-x}{h_n}\right)\right)=0 \quad (2.41)$$

for all $j > l$. In order to prove this, we use the fact that $\int_{u_1}^{v_1}\sigma(X_s)dW_s$ is independent of $\int_{u_2}^{v_2}\sigma(X_s)dW_s$ and X_w whenever $u_2 \le v_2 \le u_1 \le v_1$ and $w \le u_1 \le v_1$. Since ξ_j is a number between $\bar{X}_j^{r_n,\Delta_n}$ and $X_{(j-1)r_n\Delta_n}$, the variable ξ_j depends on those X_s's such that $(j-1)r_n\Delta_n \le s \le jr_n\Delta_n$. Thus, by independence, the left-hand side of (2.41) equals to

$$\mathbb{E}\left(\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right)\mathbb{E}\left(\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_j-x}{h_n}\right)\mathcal{W}_{lr_n+i}^{(l+1)r_n+i}\mathcal{W}_{(l-1)r_n}^{(l-1)r_n+k}\frac{1}{h_n}K'\left(\frac{\xi_l-x}{h_n}\right)\right)=0,$$

because integrals with respect to Brownian motion, such as $\mathbb{E}\left(\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right)$, have mean zero. This proves (2.41).

Then, by the boundedness of K', (2.40) is bounded further by

$$M^2\sum_{j=1}^{m_n-2}\mathbb{E}\left(\left[\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right]^2\left[\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\right]^2\right)=M^2\sum_{j=1}^{m_n-2}\mathbb{E}\left(\left[\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}\right]^2\right)\mathbb{E}\left(\left[\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}\right]^2\right) \quad (2.42)$$

where we used the fact that $\mathcal{W}_{jr_n+i}^{(j+1)r_n+i}$ and $\mathcal{W}_{(j-1)r_n}^{(j-1)r_n+k}$ are independent. Applying the bounds based on the Remark after Lemma 2.2, we can bound (2.42) by a constant times $mr_n^2\Delta_n^2$. This is a bound of (2.40), so we can bound the L^1 norm of (2.39) by a constant times

$$\frac{1}{r_n^2h_n}\sum_{i=1}^{r_n}\sum_{k=1}^{r_n}\sqrt{m_nr_n^2\Delta_n^2}=\sqrt{\frac{m_nr_n^2\Delta_n^2}{h_n^2}}=\sqrt{\frac{nr_n\Delta_n^2}{h_n^2}}$$

(recall that $m_n = n/r_n$ as in Assumption 2.1). We can rewrite the term inside the square root as

$$\frac{nr_n\Delta_n^2}{h_n^2} \;=\; \frac{nr_n^{1/2}\Delta_n^{3/2}}{h_n}\times\frac{r_n^{1/2}\Delta_n^{1/2}}{h_n}.$$

Now we show that the right-hand side is $o(1)$. We have shown that the first component is $o(1)$ in (2.38). For the second component, by condition (i) of Theorem 2.1 and the assumption that

$n\Delta_n \to \infty$ (see Assumption 2.1),

$$\frac{\sqrt{r_n\Delta_n}}{h_n} = \frac{n\Delta_n}{h_n}\sqrt{r_n\Delta_n} \times \frac{1}{n\Delta_n} = o(1) \times o(1).$$

Therefore, the L^1 norm of (2.39) is $o(1)$ and thus $o(\sqrt{n\Delta_n h_n})$ as $n\Delta_n h_n \to \infty$ by condition (ii) of Theorem 2.1. This implies that (2.39) is $o_p(\sqrt{n\Delta_n h_n})$ as desired.

Study of the increment difference terms

We first show $\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X},inc}(j)$, which is defined at the beginning of Section 2.6.4 and (2.30), is $o_p(\sqrt{n\Delta_n h_n})$. We write $Z_j^{inc} - W_j^{inc} = D_{j+1} - D_j$ where $D_j = \bar{Y}_{j+1}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n} = \bar{\varepsilon}_{j+1}^{r_n,\Delta_n}$. Then we write

$$\begin{aligned}
\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X},inc}(j) &= \sum_{j=1}^{m_n-2} (D_{j+1} - D_j) K\left(\frac{W_j^{kern} - x}{h_n}\right) && (2.43)\\
&= \sum_{j=1}^{m_n-2} D_{j+1} K\left(\frac{W_j^{kern} - x}{h_n}\right) - \sum_{j=0}^{m_n-3} D_{j+1} K\left(\frac{W_{j+1}^{kern} - x}{h_n}\right)\\
&= D_{m_n-1} K\left(\frac{W_{m_n-2}^{kern} - x}{h_n}\right) - D_1 K\left(\frac{W_1^{kern} - x}{h_n}\right)\\
&\quad - \sum_{j=1}^{m_n-2} D_{j+1}\left[K\left(\frac{W_{j+1}^{kern} - x}{h_n}\right) - K\left(\frac{W_j^{kern} - x}{h_n}\right)\right]. && (2.44)
\end{aligned}$$

Then, the boundedness and the Lipschitz continuity of K and the boundedness of K' yield

$$\left|\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{Y},\bar{X},inc}(j)\right| \leq C\left\{|D_{m_n-1}| + |D_1| + \frac{1}{h_n}\sum_{j=1}^{m_n-2} |D_{j+1}||W_{j+1}^{kern} - W_j^{kern}|\right\} \quad (2.45)$$

for a suitable constant C. Recall that $W_{j+1}^{kern} - W_j^{kern} = \bar{\varepsilon}_j^{r_n,\Delta_n}$ for $\mathcal{N}_{\bar{Y},\bar{X},inc}(j)$. We use the independence of $\{X_t\}$ and $\{\varepsilon_t\}$ and the independence of ε_t's to bound the expectation of the right-hand side of (2.45) further by a constant times

$$\mathbb{E}(|\bar{\varepsilon}_{m_n}^{r_n,\Delta_n}|) + \mathbb{E}(|\bar{\varepsilon}_2^{r_n,\Delta_n}|) + \frac{1}{h_n}\sum_{j=1}^{m_n-3} \mathbb{E}(|\bar{\varepsilon}_{j+2}^{r_n,\Delta_n}|)\mathbb{E}(|\bar{X}_{j+1}^{r_n,\Delta_n} - \bar{X}_j^{r_n,\Delta_n}|).$$

We use Lemma 2.1 and (iv) and (vi) of Lemma 2.3 to bound it further by a constant times

$$\frac{1}{\sqrt{r_n}} + \frac{m_n\sqrt{\Delta_n}}{h_n}.$$

This bound is $o\left(\sqrt{n\Delta_n h_n}\right)$ by the following. First, $1/\sqrt{r_n} = o(1)$ by Assumption 2.1, which implies it is $o\left(\sqrt{n\Delta_n h_n}\right)$ as $n\Delta_n h_n \to \infty$ by condition (ii) of Theorem 2.1. Also,

$$\frac{1}{\sqrt{n\Delta_n h_n}} \times \frac{m_n\sqrt{\Delta_n}}{h_n} = \sqrt{\frac{n}{r_n^2 h_n^3}} = o(1)$$

by condition (iii) of Theorem 2.1.

Next, we show $\sum_{j=1}^{m_n-2} \mathcal{N}_{X,X,inc}(j)$, which is defined at the beginning of Section 2.6.4 and (2.30), is $o_p(\sqrt{n\Delta_n h_n})$. We first write $Z_j^{inc} - W_j^{inc} = D_{j+1} - D_j$ where $D_j = X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}$. Then, by (2.2), which is the definition of X_t that $X_t = X_0 + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dW_s$ a.s., we have the following equality almost surely:

$$D_j = \int_{(j-1)r_n\Delta_n}^{jr_n\Delta_n} \mu(X_s)ds + \int_{(j-1)r_n\Delta_n}^{jr_n\Delta_n} \sigma(X_s)dW_s = \mathcal{M}_{(j-1)r_n}^{jr_n\Delta_n} + \mathcal{W}_{(j-1)r_n}^{jr_n\Delta_n} \equiv E_j + F_j$$

(recall the definition of $\mathcal{M}$ and $\mathcal{W}$ in the Remark after Lemma 2.2). Therefore, we have $Z_j^{inc} - W_j^{inc} = D_{j+1} - D_j = (E_{j+1} - E_j) + (F_{j+1} - F_j)$ almost surely. We now write

$$\begin{aligned}\sum_{j=1}^{m_n-2} \mathcal{N}_{X,X,inc}(j) &= \sum_{j=1}^{m_n-2} (E_{j+1} - E_j)K\left(\frac{W_j^{kern} - x}{h_n}\right) + \sum_{j=1}^{m_n-2} (F_{j+1} - F_j)K\left(\frac{W_j^{kern} - x}{h_n}\right) \\ &\equiv \mathcal{N}_{X,X,e} + \mathcal{N}_{X,X,f}. \end{aligned} \tag{2.46}$$

Note that $\mathcal{N}_{X,X,e}$ and $\mathcal{N}_{X,X,f}$ are of the same forms as (2.43), except for having E_j's and F_j's instead of D_j's, respectively. Now we show $\mathcal{N}_{X,X,e}$ and $\mathcal{N}_{X,X,f}$ are $o_p(\sqrt{n\Delta_n h_n})$.

First, for $\mathcal{N}_{X,X,e}$, we use the bound (2.45) to bound $\mathcal{N}_{X,X,e}$ by a constant times

$$\left|\mathcal{M}_{(m_n-2)r_n}^{(m_n-1)r_n}\right| + \left|\mathcal{M}_0^{r_n}\right| + \frac{1}{h_n}\sum_{j=1}^{m_n-2} \left|\mathcal{M}_{jr_n}^{(j+1)r_n}\right| \left|X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}\right|. \tag{2.47}$$

Since $|X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n}| \leq \kappa_n$ by the definition of κ_n in Lemma 2.3, we can bound (2.47) by

$$\left|\mathcal{M}_{(m_n-2)r_n}^{(m_n-1)r_n}\right| + |\mathcal{M}_0^{r_n}| + \frac{\kappa_n}{h_n}\sum_{j=1}^{m_n-2}\left|\mathcal{M}_{jr_n}^{(j+1)r_n}\right|. \tag{2.48}$$

In addition, Lemma 2.2 and the Markov inequality imply that $\sum_{j=1}^{m_n-2}\left|\mathcal{M}_{jr_n}^{(j+1)r_n}\right| = O_p(n\Delta_n)$ and that $\mathcal{M}_j^{j+1} = O_p(r_n\Delta_n)$ for any j. Therefore, we can rewrite (2.48) as

$$O_p(r_n\Delta_n) + O_p(r_n\Delta_n) + \frac{\kappa_n}{h_n}O_p(n\Delta_n).$$

This bound is $o_p(1)$, which implies it is $o_p\left(\sqrt{n\Delta_n h_n}\right)$ as $n\Delta_n h_n \to \infty$ by condition (ii) of Theorem 2.1, by the following. First, $O_p(r_n\Delta_n) = o_p(1)$ as $r_n\Delta_n \to 0$ by Assumption 2.1. In addition, $(n\Delta_n/h_n) \times \kappa_n = o_{a.s.}(1)$ by condition (i) of Theorem 2.1 and (i) of Lemma 2.3, which implies $(\kappa_n/h_n) \times O_p(n\Delta_n) = (O_p(n\Delta_n)/h_n) \times \kappa_n = o_p(1)$.

For $\mathcal{N}_{X,X,f}$, defined in (2.46), we can bound the absolute value of $\mathcal{N}_{X,X,f}$, using the reasoning used from (2.43) to (2.44) and the boundedness of K, by a constant times

$$\left|\mathcal{W}_{m_n-2}^{m_n-1}\right| + \left|\mathcal{W}_0^1\right| + \left|\sum_{j=1}^{m_n-2}\mathcal{W}_{jr_n}^{(j+1)r_n}\left[K\left(\frac{X_{jr_n\Delta_n}-x}{h_n}\right) - K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)\right]\right|.$$

Lemma 2.2 and the Markov inequality imply that, for any j, $\mathcal{W}_j^{j+1} = O_p(\sqrt{r_n\Delta_n}) = o_p(1)$. Therefore, it remains to show that

$$\left|\sum_{j=1}^{m_n-2}\mathcal{W}_{jr_n}^{(j+1)r_n}\left[K\left(\frac{X_{jr_n\Delta_n}-x}{h_n}\right) - K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)\right]\right| = o_p\left(\sqrt{n\Delta_n h_n}\right).$$

To show this, we show that its L^2 norm is $o\left(\sqrt{n\Delta_n h_n}\right)$. Note that

$$\begin{aligned}
&\mathbb{E}\left(\left(\sum_{j=1}^{m_n-2}\mathcal{W}_{jr_n}^{(j+1)r_n}\left[K\left(\frac{X_{jr_n\Delta_n}-x}{h_n}\right) - K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)\right]\right)^2\right) \\
= &\sum_{j=1}^{m_n-2}\mathbb{E}\left(\left(\mathcal{W}_j^{j+1}\right)^2\right)\mathbb{E}\left(\left[K\left(\frac{X_{jr_n\Delta_n}-x}{h_n}\right) - K\left(\frac{X_{(j-1)r_n\Delta_n}-x}{h_n}\right)\right]^2\right)
\end{aligned}$$

by the same reasoning used to prove (2.41). In addition, using the Lipschitz continuity of K,

we can bound the above further by a constant times

$$\sum_{j=1}^{m_n-2} \mathbb{E}\left(\left(\mathcal{W}_j^{j+1}\right)^2\right) \frac{\mathbb{E}\left((X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n})^2\right)}{h_n^2} \;\leq\; \frac{\gamma_n}{h_n^2} \sum_{j=1}^{m_n-2} \mathbb{E}\left(\left(\mathcal{W}_j^{j+1}\right)^2\right), \tag{2.49}$$

where we used $\mathbb{E}\left((X_{jr_n\Delta_n} - X_{(j-1)r_n\Delta_n})^2\right) \leq \gamma_n$ for the inequality, which can be proved adapting the proof of (iv) of Lemma 2.3. Then, since $\mathbb{E}((\mathcal{W}_j^{j+1})^2) \leq \mathbb{E}(\sigma^2(X_0))r_n\Delta_n$ by Lemma 2.2 (and the Remark after that) and $\gamma_n \leq \beta r_n\Delta_n$ by (iv) of Lemma 2.3, we can bound the right-hand side of (2.49) further by a constant times

$$\frac{m_n r_n^2 \Delta_n^2}{h_n^2} \;=\; \frac{n r_n \Delta_n^2}{h_n^2}$$

($m_n = n/r_n$ as in Assumption 2.1). We must show that this bound is $o(n\Delta_n h_n)$, which proves that the L^2 norm is $o(\sqrt{n\Delta_n h_n})$. The following proves it is $o(n\Delta_n h_n)$:

$$\frac{1}{n\Delta_n h_n} \times \frac{n r_n \Delta_n^2}{h_n^2} = \left(\frac{n\Delta_n}{h_n}\right)^2 r_n\Delta_n \times \frac{1}{n\Delta_n h_n} \times \frac{1}{n\Delta_n} = o(1) \times o(1) \times o(1)$$

by conditions (i) and (ii) of Theorem 2.1 and the assumption that $n\Delta_n \to 0$ in Assumption 2.1.

Lastly, we show $\sum_{j=1}^{m_n-2} \mathcal{N}_{\bar{X},X,inc}(j)$, which is defined at the beginning of Section 2.6.4 and (2.30), is $o_p(\sqrt{n\Delta_n h_n})$. By (2.2), which is the definition of X_t that $X_t = X_0 + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dW_s$ a.s., the increment terms $Z_j^{inc} = \bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n}$ and $W_j^{inc} = X_{(j+1)r_n\Delta_n} + X_{jr_n\Delta_n}$ satisfy the following equations almost surely:

$$\begin{aligned} \bar{X}_{j+2}^{r_n,\Delta_n} - \bar{X}_{j+1}^{r_n,\Delta_n} &= \frac{1}{r_n}\sum_{i=1}^{r_n} \mathcal{M}_{jr_n+i}^{(j+1)r_n+i} + \frac{1}{r_n}\sum_{i=1}^{r_n} \mathcal{W}_{jr_n+i}^{(j+1)r_n+i}, \\ X_{(j+1)r_n\Delta_n} + X_{jr_n\Delta_n} &= \mathcal{M}_{jr_n}^{(j+1)r_n} + \mathcal{W}_{jr_n}^{(j+1)r_n} \end{aligned}$$

(recall the definition of $\mathcal{M}$'s and $\mathcal{W}$'s in the Remark after Lemma 2.2). Therefore, we can write

$$Z_j^{inc} - W_j^{inc} = \frac{1}{r_n}\sum_{i=1}^{r_n}\left(\mathcal{M}_{jr_n+i}^{(j+1)r_n+i} - \mathcal{M}_{jr_n}^{(j+1)r_n}\right) + \frac{1}{r_n}\sum_{i=1}^{r_n}\left(\mathcal{W}_{jr_n+i}^{(j+1)r_n+i} - \mathcal{W}_{jr_n}^{(j+1)r_n}\right).$$

In addition, as $\mathcal{M}$ and $\mathcal{W}$ are simplified notation for integrals, we have

$$\mathcal{M}_{jr_n+i}^{(j+1)r_n+i} - \mathcal{M}_{jr_n}^{(j+1)r_n} = \left(\mathcal{M}_{jr_n+i}^{(j+1)r_n} + \mathcal{M}_{(j+1)r_n}^{(j+1)r_n+i}\right) - \left(\mathcal{M}_{jr_n}^{jr_n+i} - \mathcal{M}_{jr_n+i}^{(j+1)r_n}\right) = \mathcal{M}_{(j+1)r_n}^{(j+1)r_n+i} - \mathcal{M}_{jr_n}^{jr_n+i}$$

and the same equation for $\mathcal{W}$'s. Then we can decompose $Z_j^{inc} - W_j^{inc}$ as

$$\begin{aligned} Z_j^{inc} - W_j^{inc} &= \left(\frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{M}_{(j+1)r_n}^{(j+1)r_n+i} - \frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{M}_{jr_n}^{jr_n+i}\right) + \left(\frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{W}_{(j+1)r_n}^{(j+1)r_n+i} - \frac{1}{r_n}\sum_{i=1}^{r_n}\mathcal{W}_{jr_n}^{jr_n+i}\right) \\ &\equiv \left(E_{j+1} - E_j\right) + \left(F_{j+1} - F_j\right). \end{aligned}$$

Then, similarly to $\sum_{j=1}^{m_n-2}\mathcal{N}_{X,X,inc}(j)$, we can decompose $\sum_{j=1}^{m_n-2}\mathcal{N}_{\bar{X},X,inc}(j)$ as sum of $\mathcal{N}_{\bar{X},X,e}$ and $\mathcal{N}_{\bar{X},X,f}$ and show that each is $o_p(\sqrt{n\Delta_n h_n})$. Briefly, $\mathcal{N}_{\bar{X},X,e}$ is bounded by a constant times

$$\frac{1}{r_n}\sum_{i=1}^{r_n}\left|\mathcal{M}_{(m_n-1)r_n}^{(m_n-1)r_n+i}\right| + \frac{1}{r_n}\sum_{i=1}^{r_n}\left|\mathcal{M}_{r_n}^{r_n+i}\right| + \frac{1}{r_n h_n}\sum_{i=1}^{r_n}\sum_{j=1}^{m_n-2}\left|\mathcal{M}_{(j+1)r_n}^{(j+1)r_n+i}\right|\left|X_{(j+1)r_n\Delta_n} - X_{jr_n\Delta_n}\right|,$$

and $\mathcal{N}_{\bar{X},X,f}$ is bounded by a constant times

$$\frac{1}{r_n}\sum_{i=1}^{r_n}\left|\mathcal{W}_{(m_n-1)r_n}^{(m_n-1)r_n+i}\right| + \frac{1}{r_n}\sum_{i=1}^{r_n}\left|\mathcal{W}_{r_n}^{r_n+i}\right| + \frac{1}{r_n}\sum_{i=1}^{r_n}\left|\sum_{j=1}^{m_n-2}\mathcal{W}_{(j+1)r_n}^{(j+1)r_n+i}\left[K\left(\frac{X_{(j+1)r_n\Delta_n} - x}{h_n}\right) - K\left(\frac{X_{jr_n\Delta_n} - x}{h_n}\right)\right]\right|.$$

Applying, again, the reasoning used to study $\mathcal{N}_{X,X,e}$ and $\mathcal{N}_{X,X,f}$ to the above completes the proof. ■

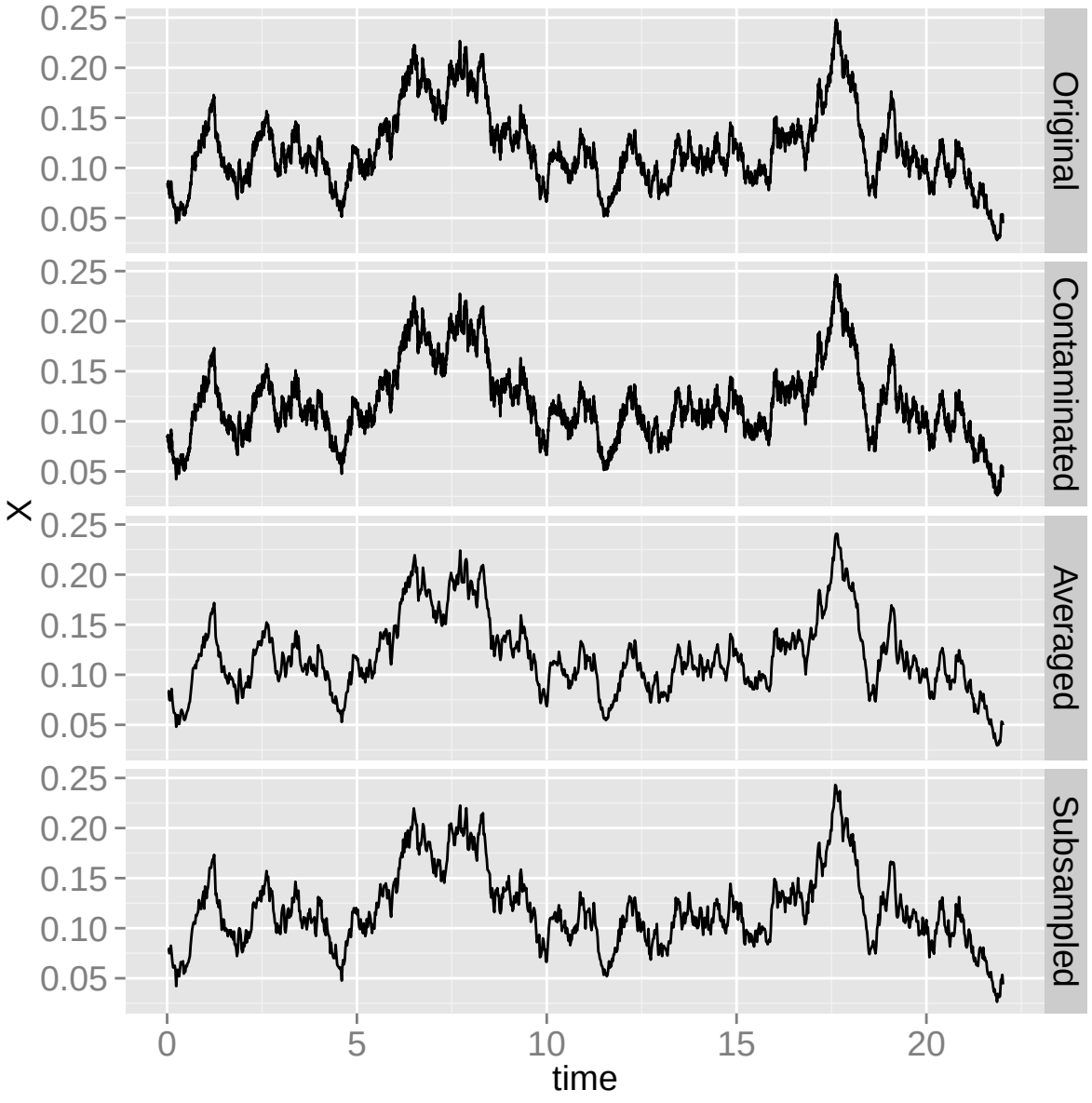

Figure 2.1: A sample path of the stochastic process defined by (2.18), with the linear drift coefficient. Label "Original" represents the process without measurement errors. Label "Contaminated" represents the process with independent $N(0, 0.002^2)$-distributed additive measurement errors. Label "Averaged" represents the averaged contaminated process with $r = 5$. Label "Subsampled" represents the subsampled process having 1/5 less sampling frequency than the original process.

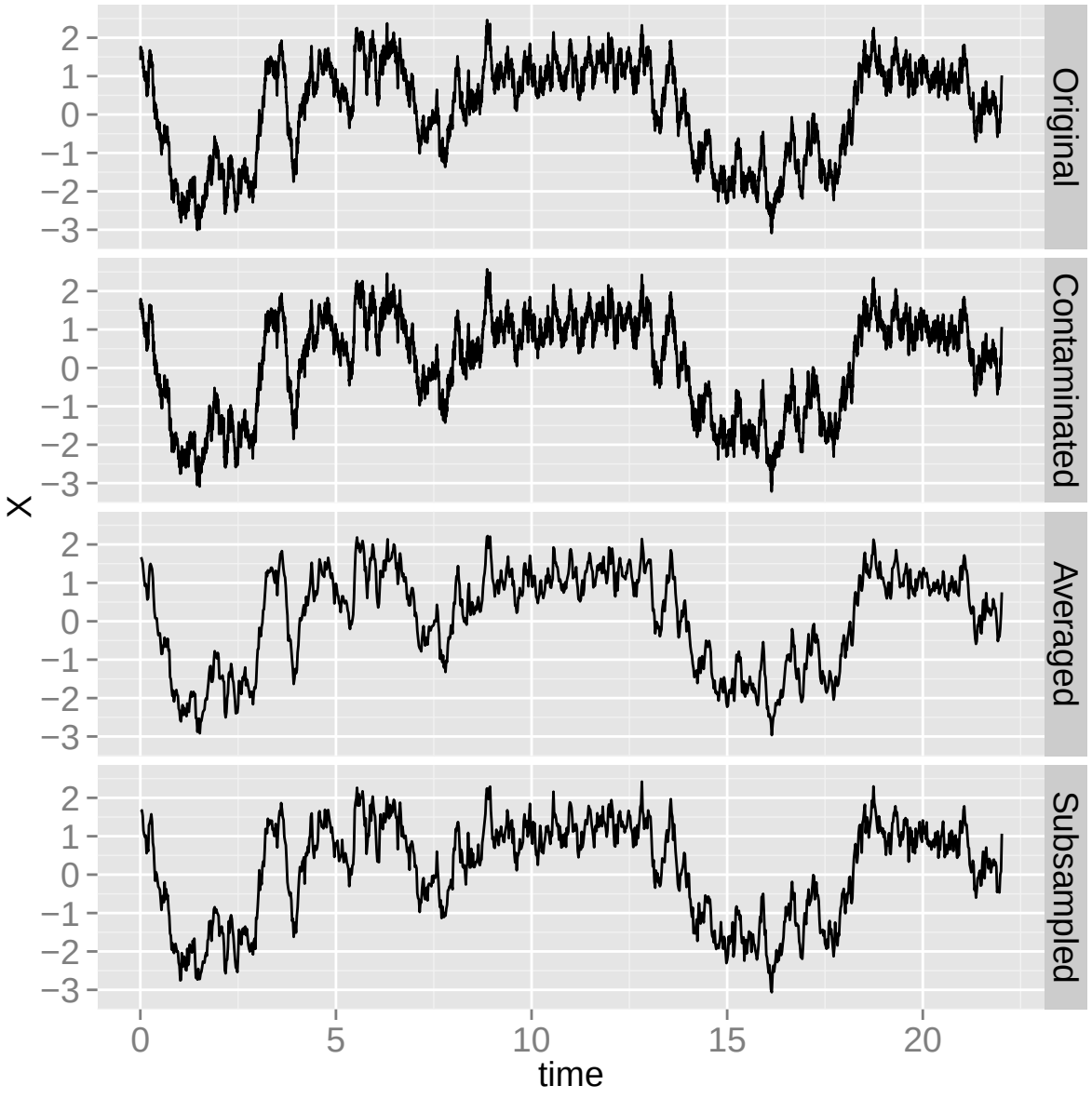

Figure 2.2: A sample path of the stochastic process defined by (2.19), with the nonlinear drift coefficient. Label "Original" represents the process without measurement errors. Label "Contaminated" represents the process with independent $N(0, 0.0661^2)$-distributed additive measurement errors. Label "Averaged" represents the averaged contaminated process with $r = 5$. Label "Subsampled" represents the subsampled process having 1/5 less sampling frequency than the original process.

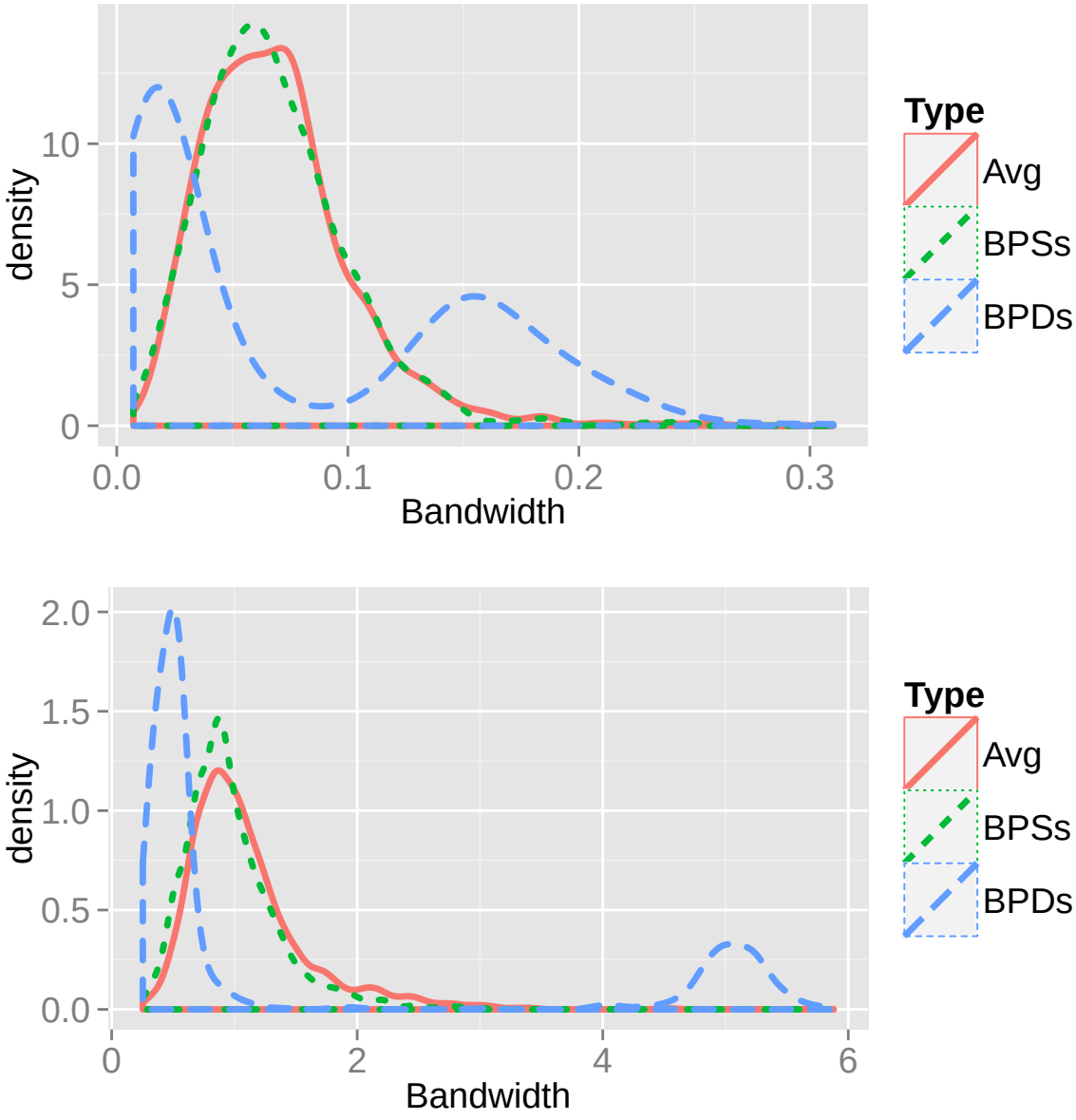

Figure 2.3: Density plot of cross-validation bandwidths of the BPSs, BPDs and Avg estimator. Labels "BPSs" and "BPDs" stand for the single-smoothing and the double-smoothing estimator of Bandi and Phillips (2003), respectively, both combined with the subsampling method. Label "Avg" stands for the pre-averaging estimator. The top panel corresponds to the model (2.18), and the bottom panel corresponds to the model (2.19).

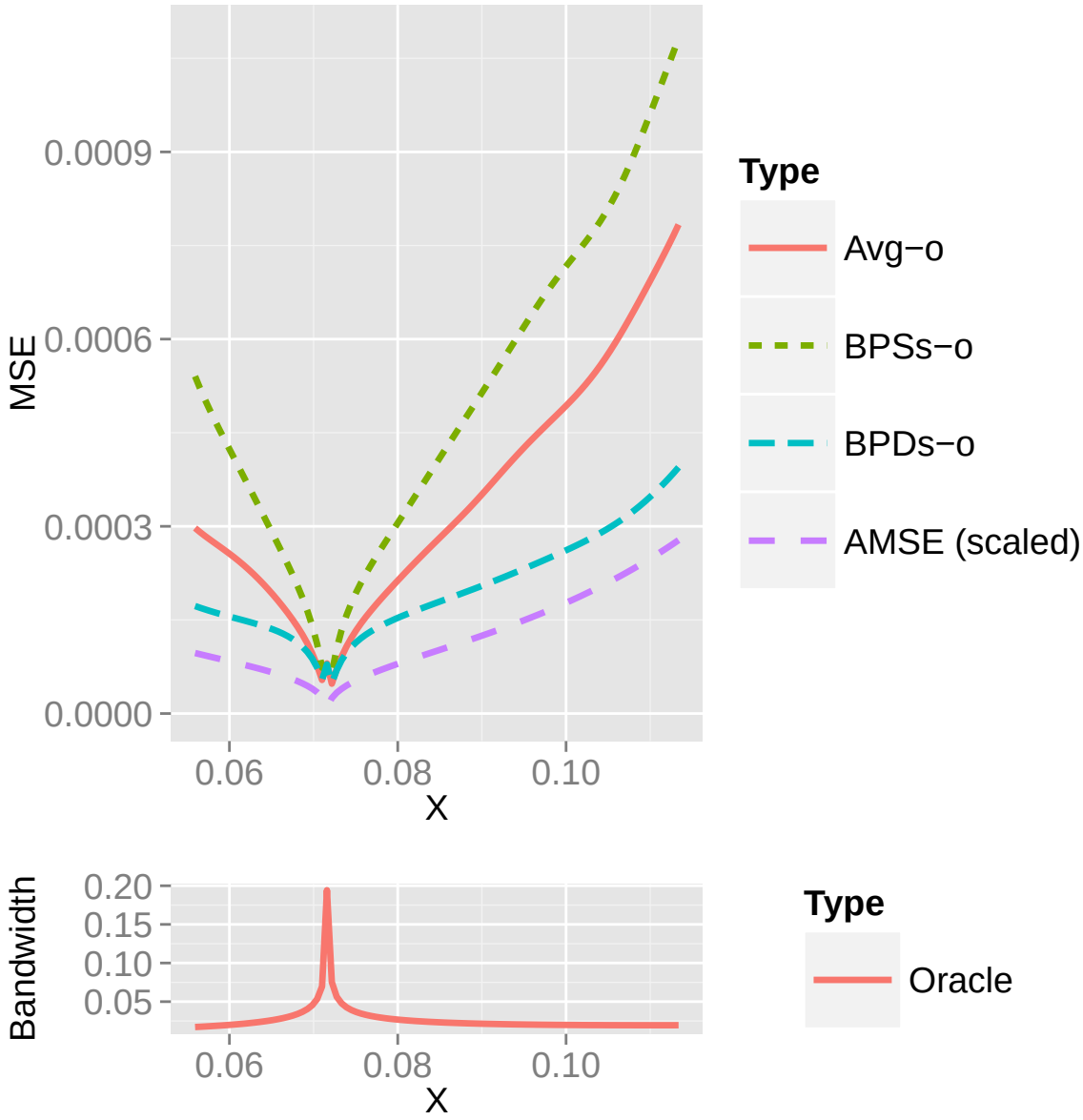

Figure 2.4: Pointwise mean squared errors (MSE) of the estimators for the model (2.18) with oracle bandwidths. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. Label "AMSE" represents the asymptotic mean squared error computed using the oracle bandwidth. The numbers of the vertical axis do not apply to the AMSE. The bottom panel depicts oracle bandwidths, $h_{opt}(x)$ defined in (2.14), according to the values of x.

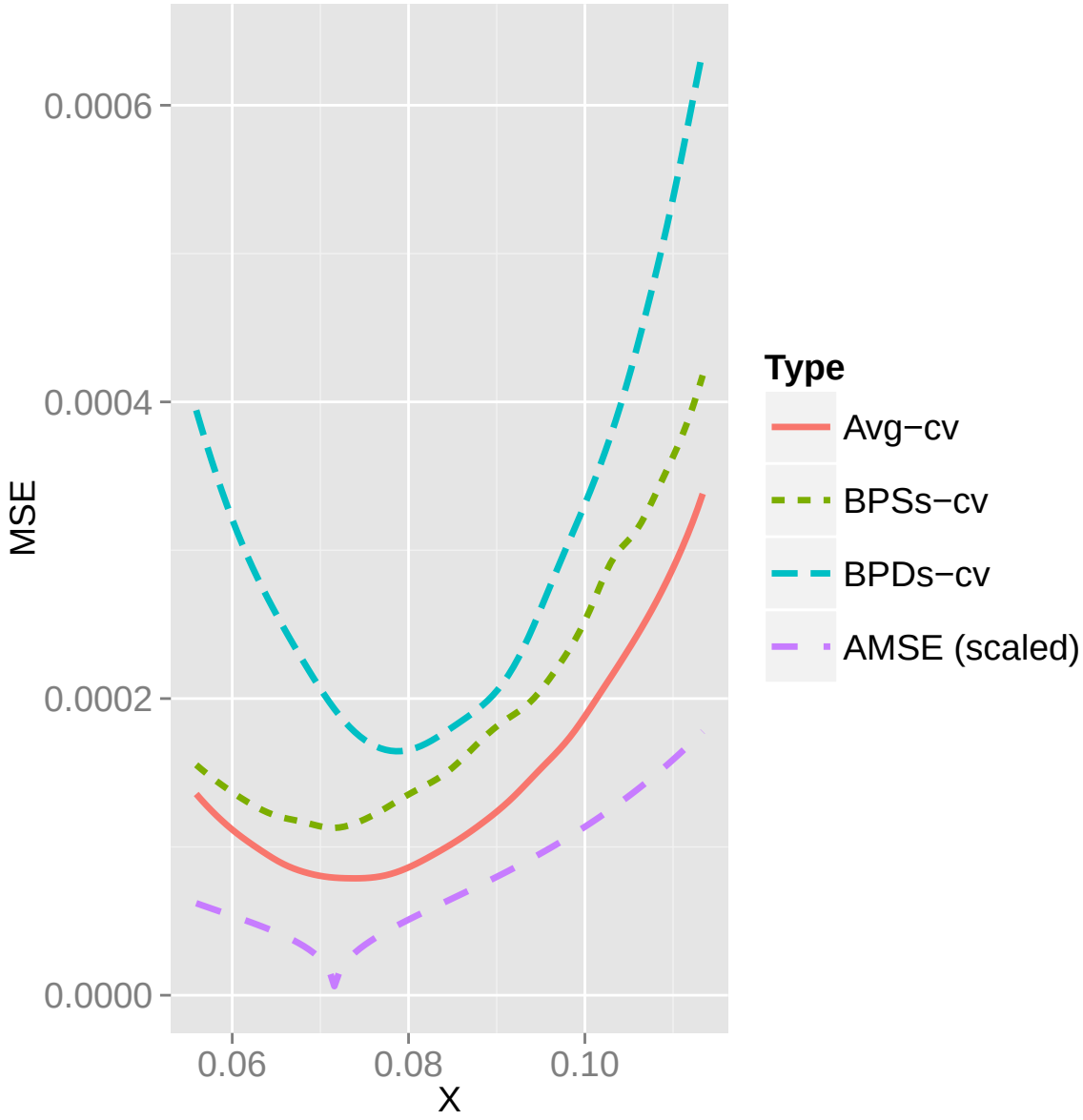

Figure 2.5: Pointwise mean squared errors (MSE) of the estimators for the model (2.18) with cross-validation bandwidths. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. Label "AMSE" represents the asymptotic mean squared error computed using the oracle bandwidth. The numbers of the vertical axis do not apply to the AMSE.

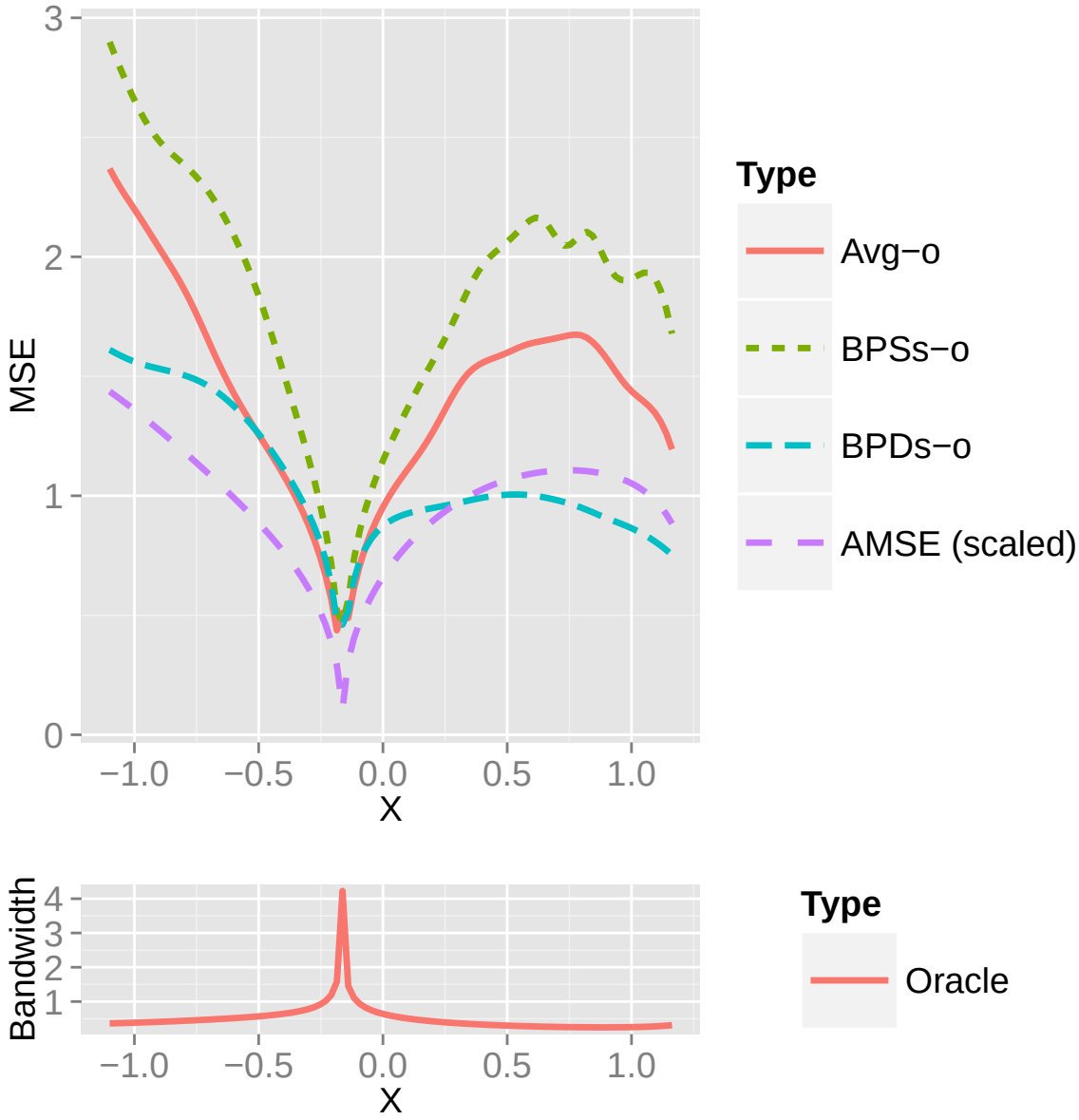

Figure 2.6: Pointwise mean squared errors (MSE) of the estimators for the model (2.19) with oracle bandwidths. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. Label "AMSE" represents the asymptotic mean squared error computed using the oracle bandwidth. The numbers of the vertical axis do not apply to the AMSE. The bottom panel depicts oracle bandwidths, $h_{opt}(x)$ defined in (2.14), according to the values of x.

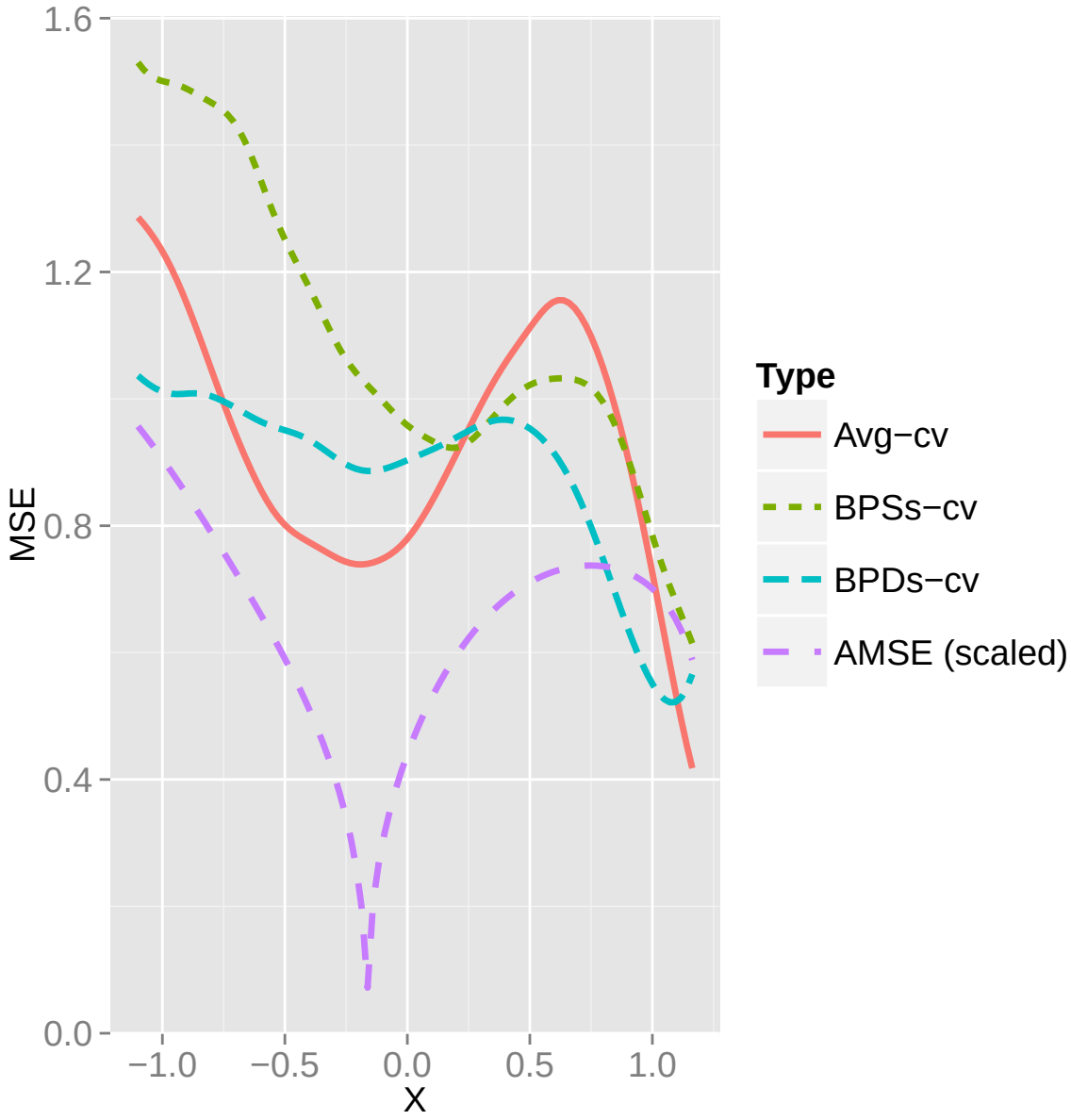

Figure 2.7: Pointwise mean squared errors (MSE) of the estimators for the model (2.19) with cross-validation bandwidths. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. Label "AMSE" represents the asymptotic mean squared error computed using the oracle bandwidth. The numbers of the vertical axis do not apply to the AMSE.

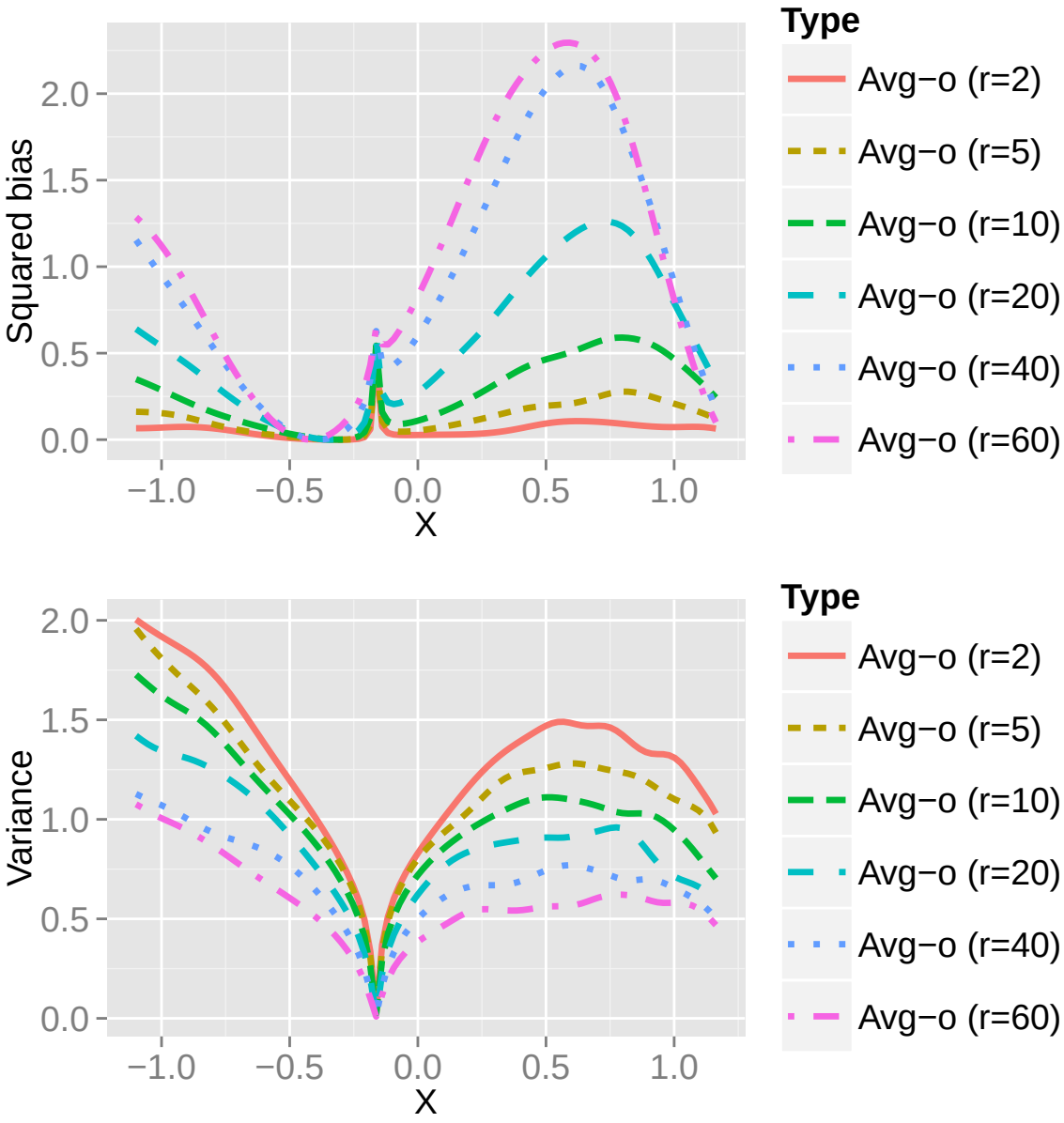

Figure 2.8: Pointwise squared biases (the top panel) and pointwise variances (the bottom panel) of the pre-averaging estimator with the oracle bandwidth (denoted by "Avg-o") under different values of the block size r for the model (2.19). The values of r are indicated in the legend.

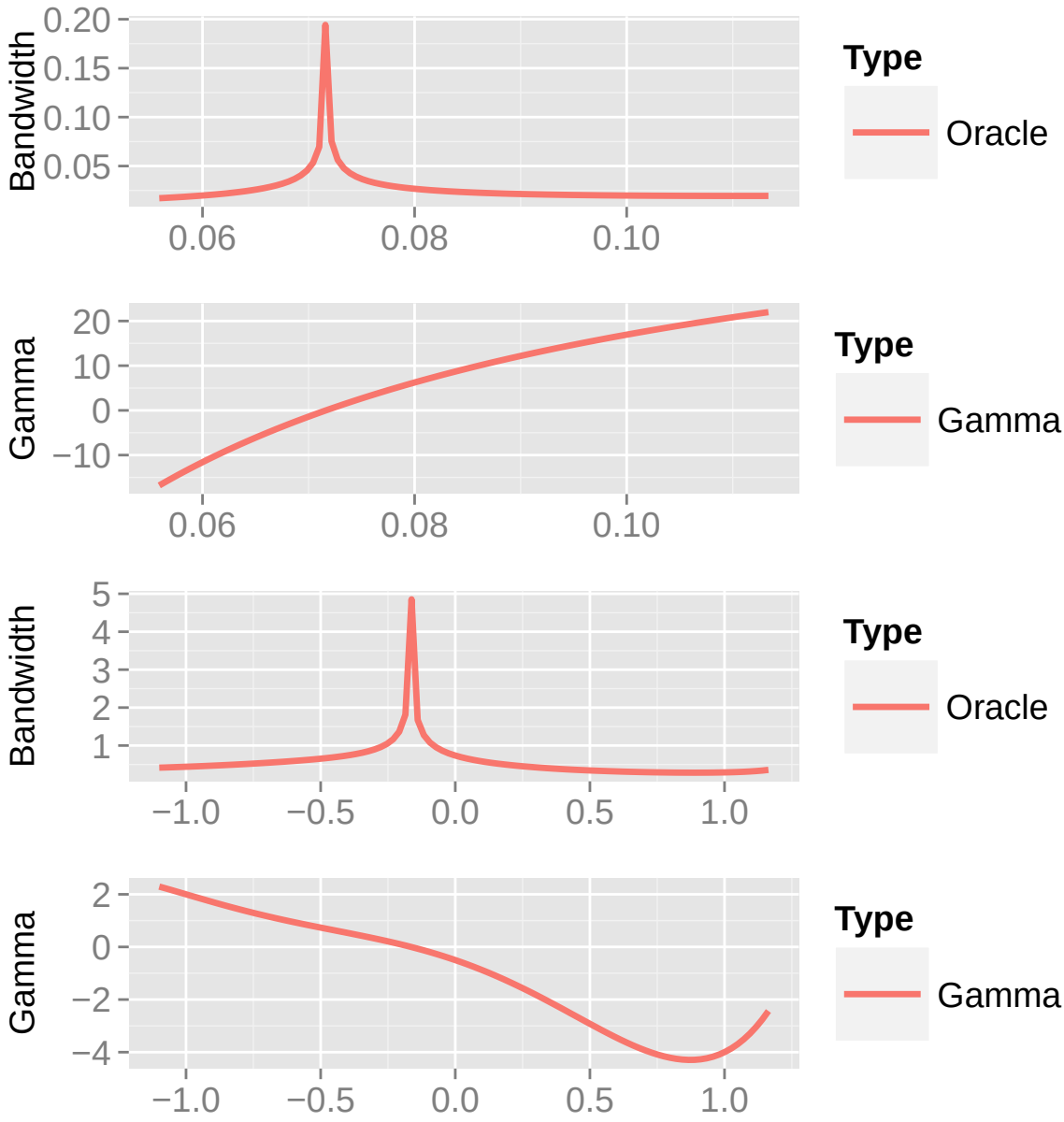

Figure 2.9: Values of the oracle bandwidth $h_{opt}(x)$ defined in (2.14) and the function $\Gamma_\mu(x)$ defined in Theorem 2.1. The two top panels depict values of $h_{opt}(x)$ and $\Gamma_\mu(x)$ according to the values of x for model (2.18). The two bottom panels depict the values for model (2.19).

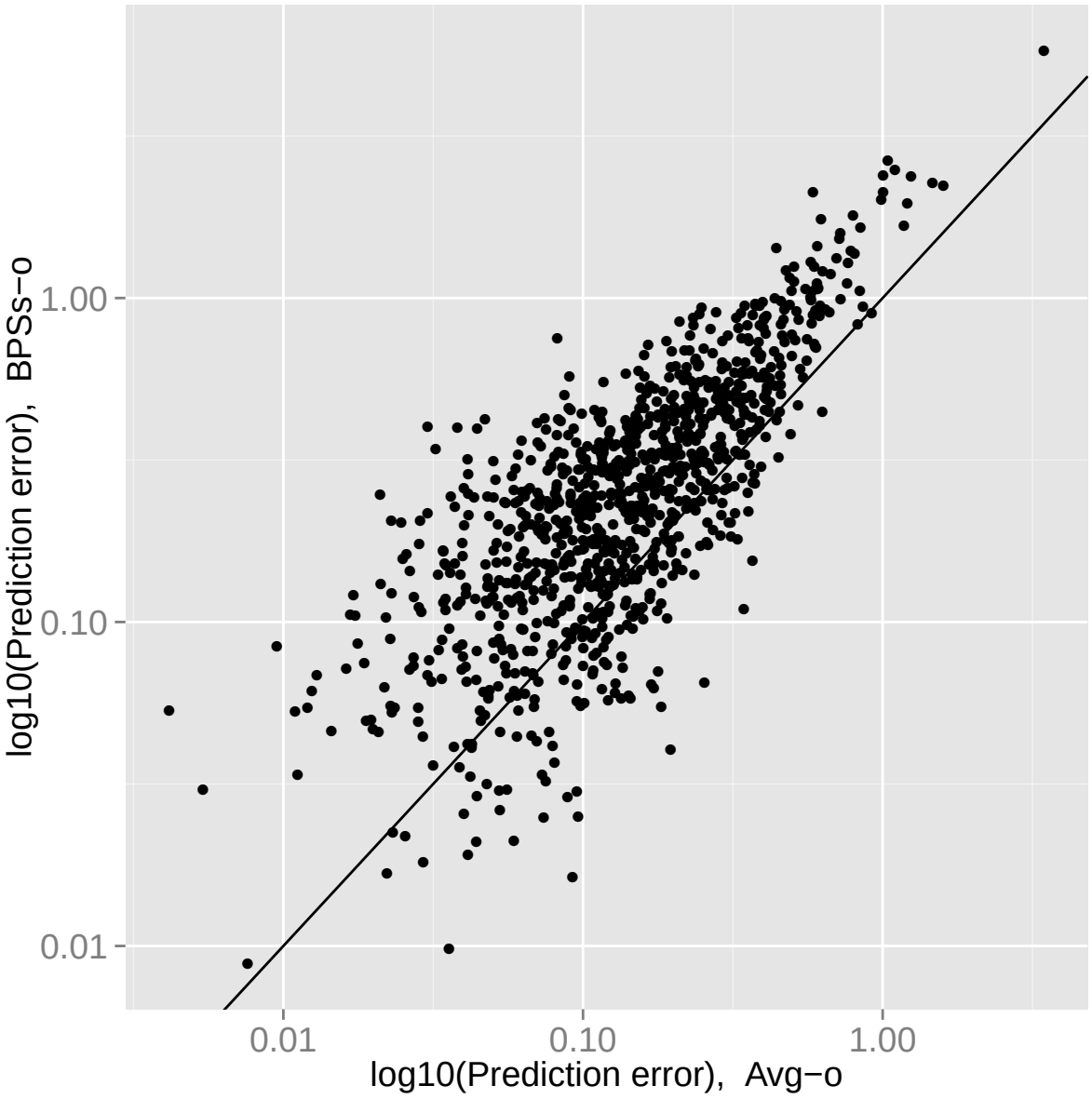

Figure 2.10: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPSs estimates for the model (2.18) with oracle bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. The black solid line is the 45 degrees line. 825 points out of 1,000 are above the line.

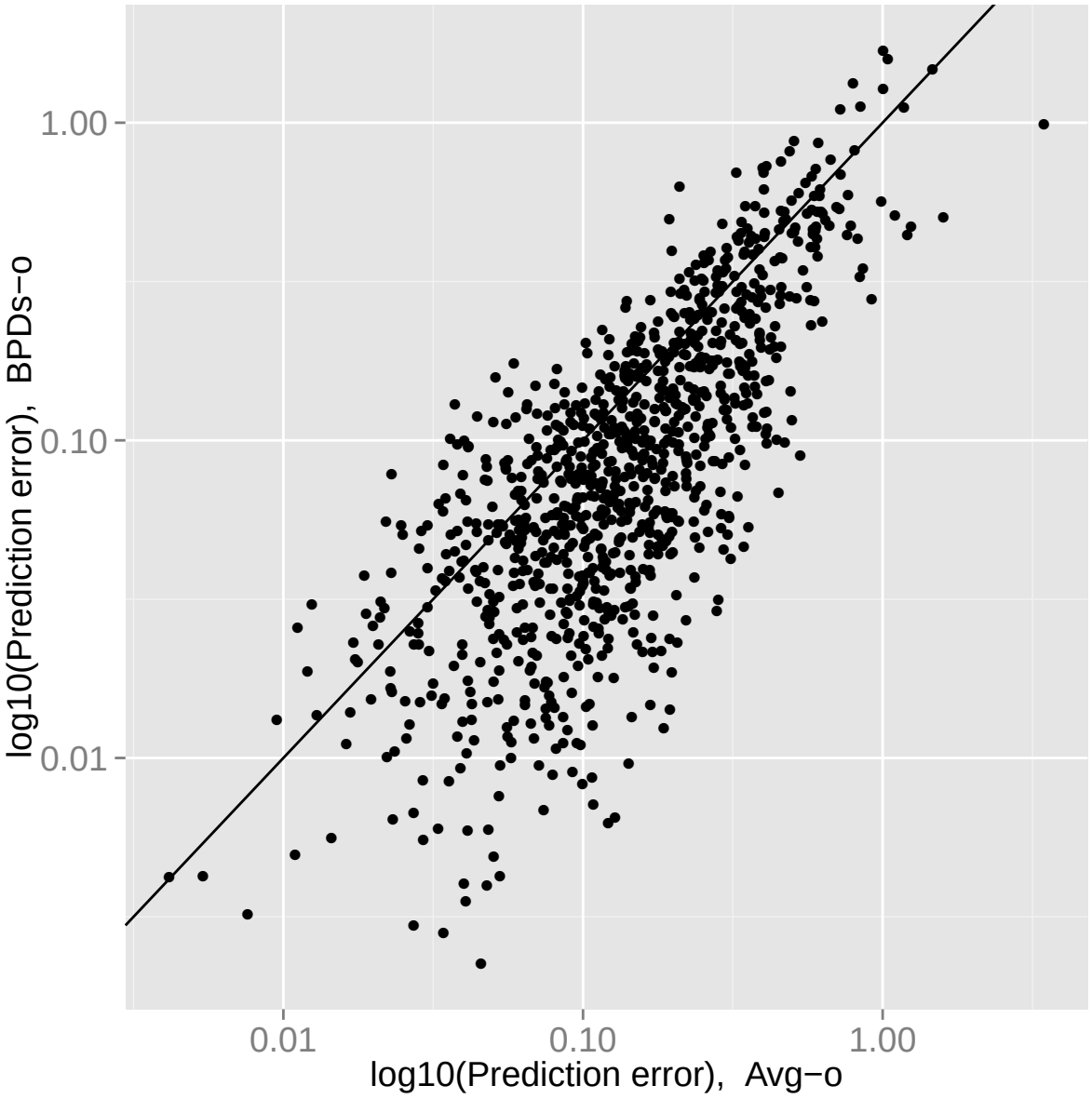

Figure 2.11: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPDs estimates for the model (2.18) with oracle bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. The black solid line is the 45 degrees line. 733 points out of 1,000 are below the line.

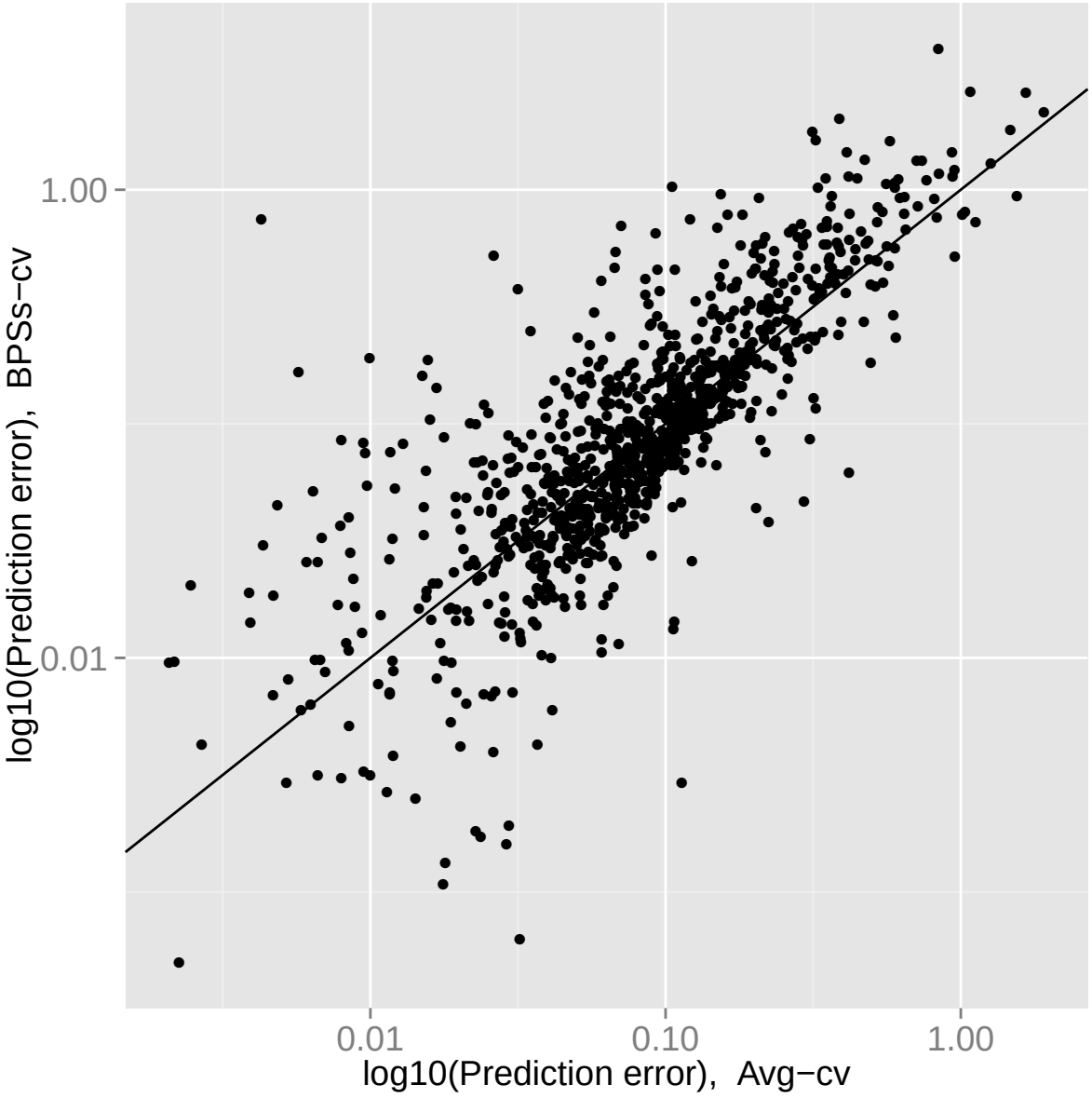

Figure 2.12: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPSs estimates for the model (2.18) with cross-validation bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. The black solid line is the 45 degrees line. 513 points out of 1,000 are above the line.

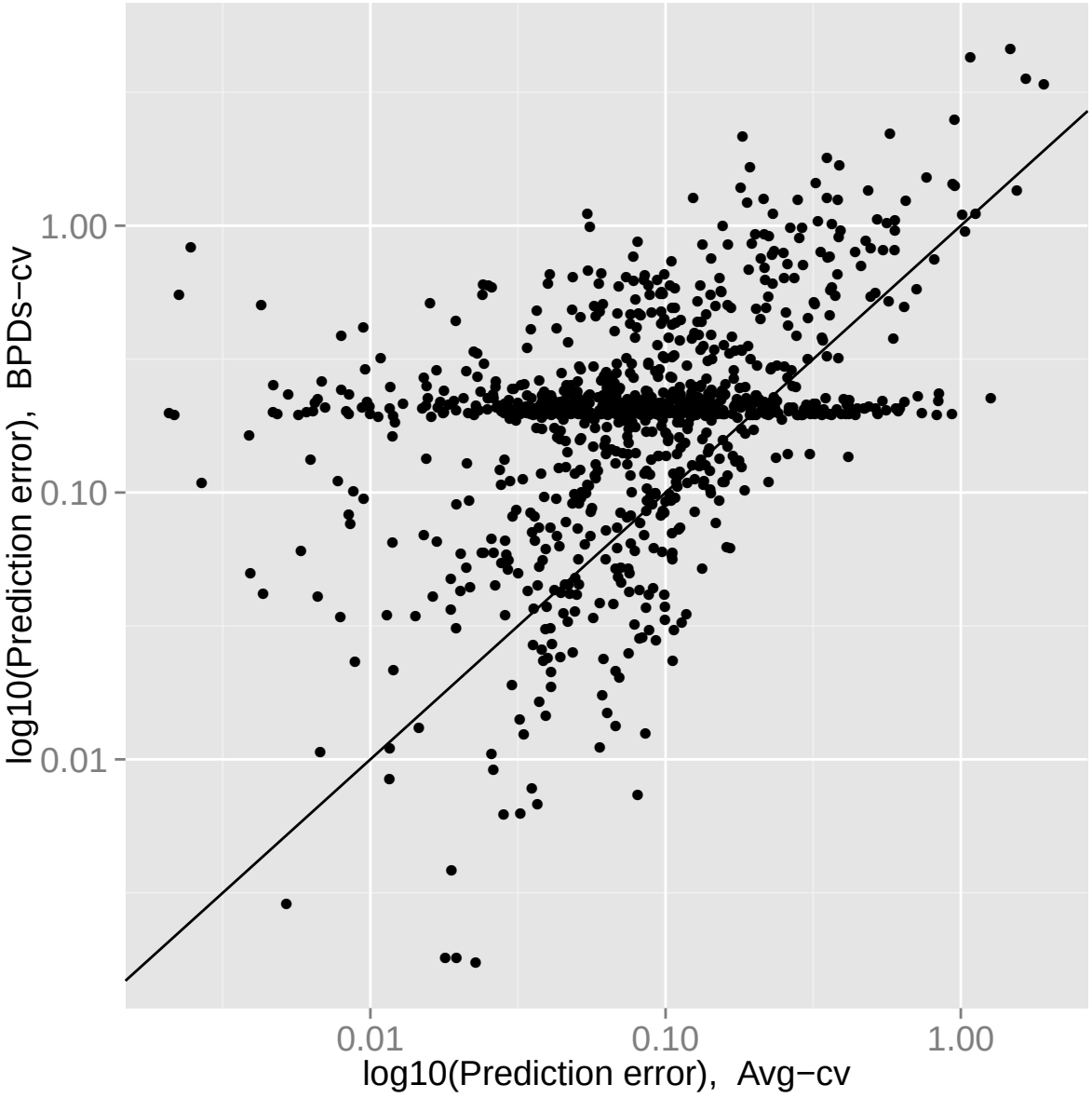

Figure 2.13: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPDs estimates for the model (2.18) with cross-validation bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. The black solid line is the 45 degrees line. 782 points out of 1,000 are above the line.

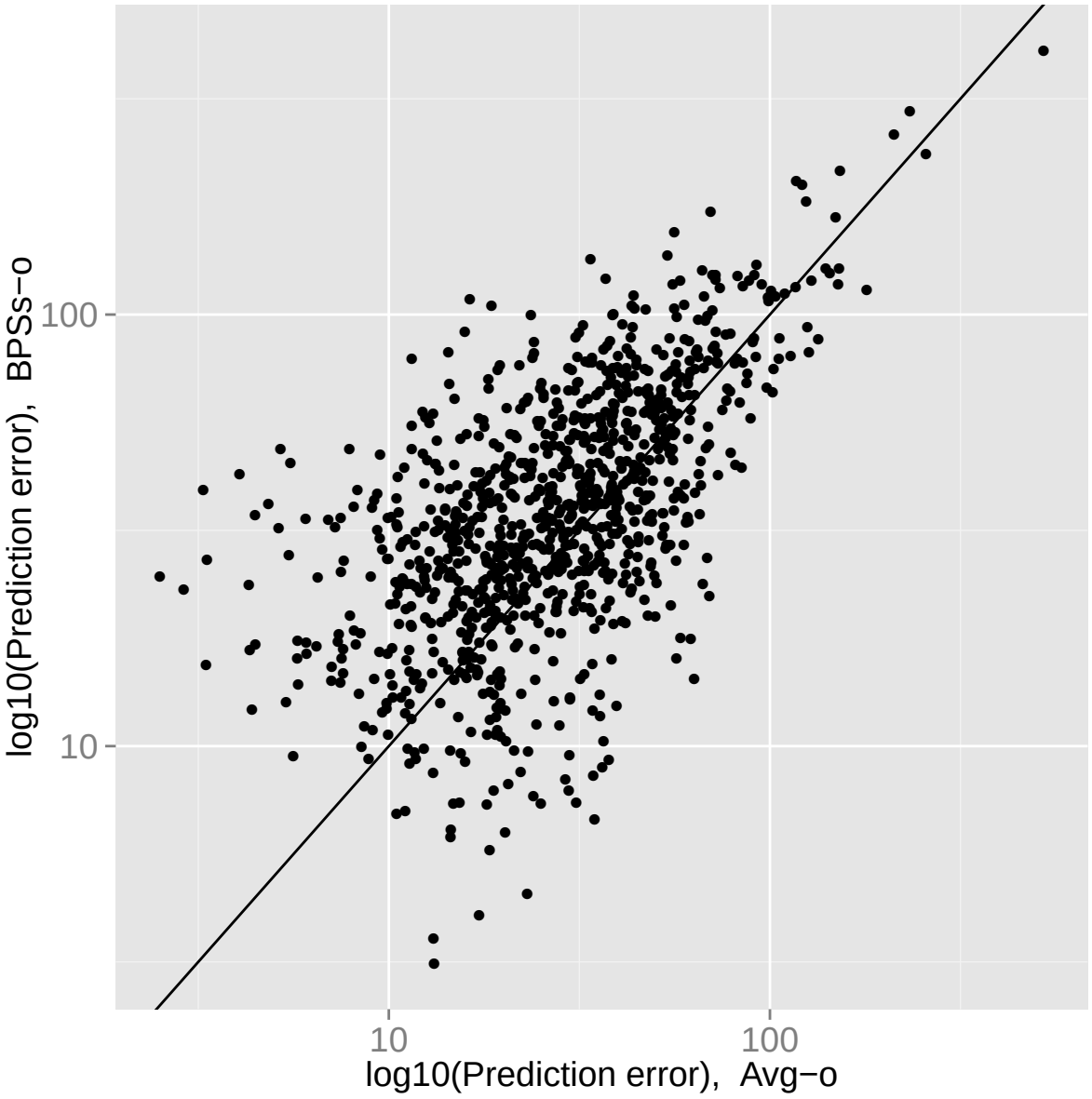

Figure 2.14: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPSs estimates for the model (2.19) with oracle bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. The black solid line is the 45 degrees line. 679 points out of 1,000 are above the line.

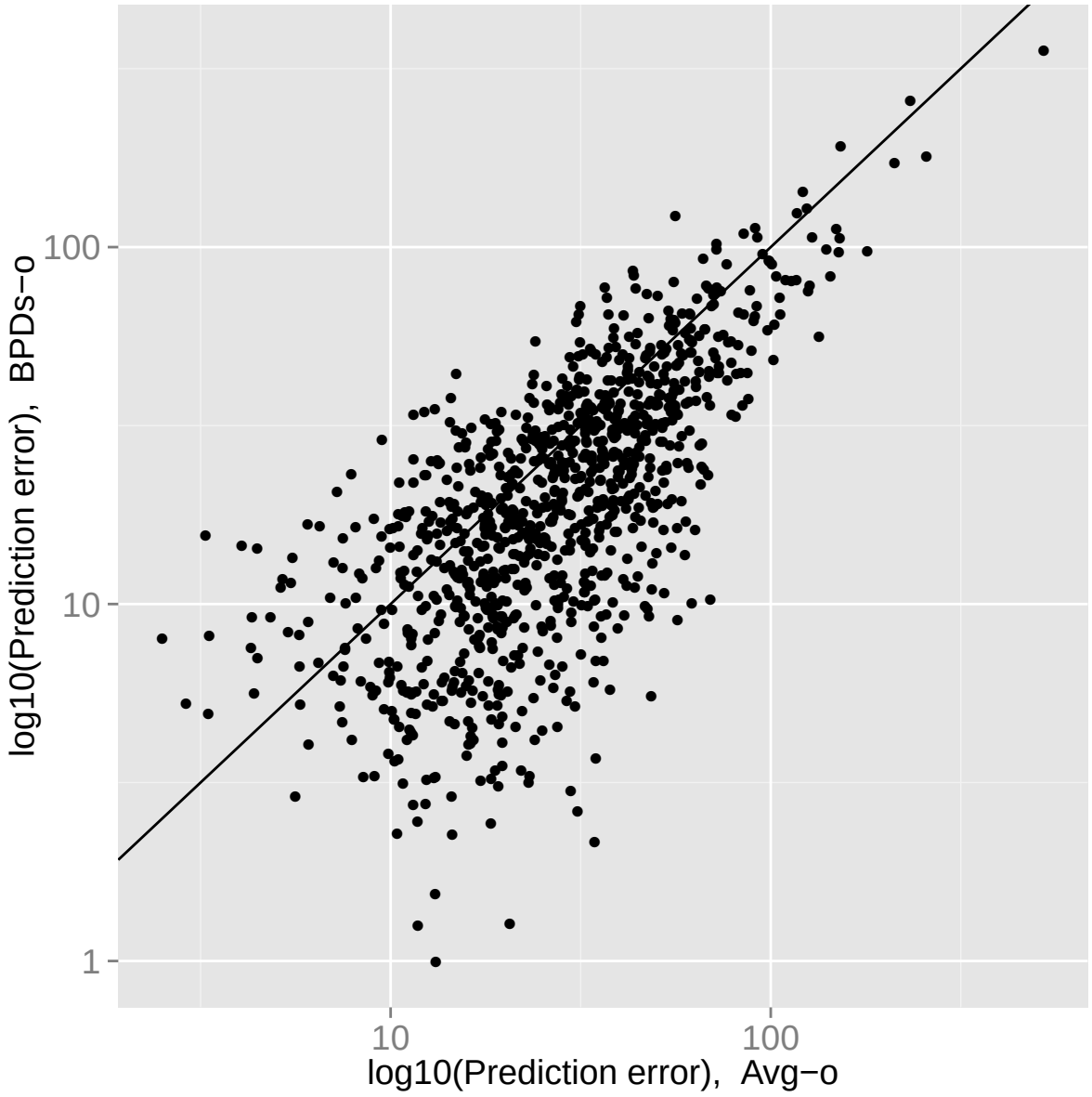

Figure 2.15: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPDs estimates for the model (2.19) with oracle bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-o" represents the oracle bandwidths are used. The black solid line is the 45 degrees line. 718 points out of 1,000 are below the line.

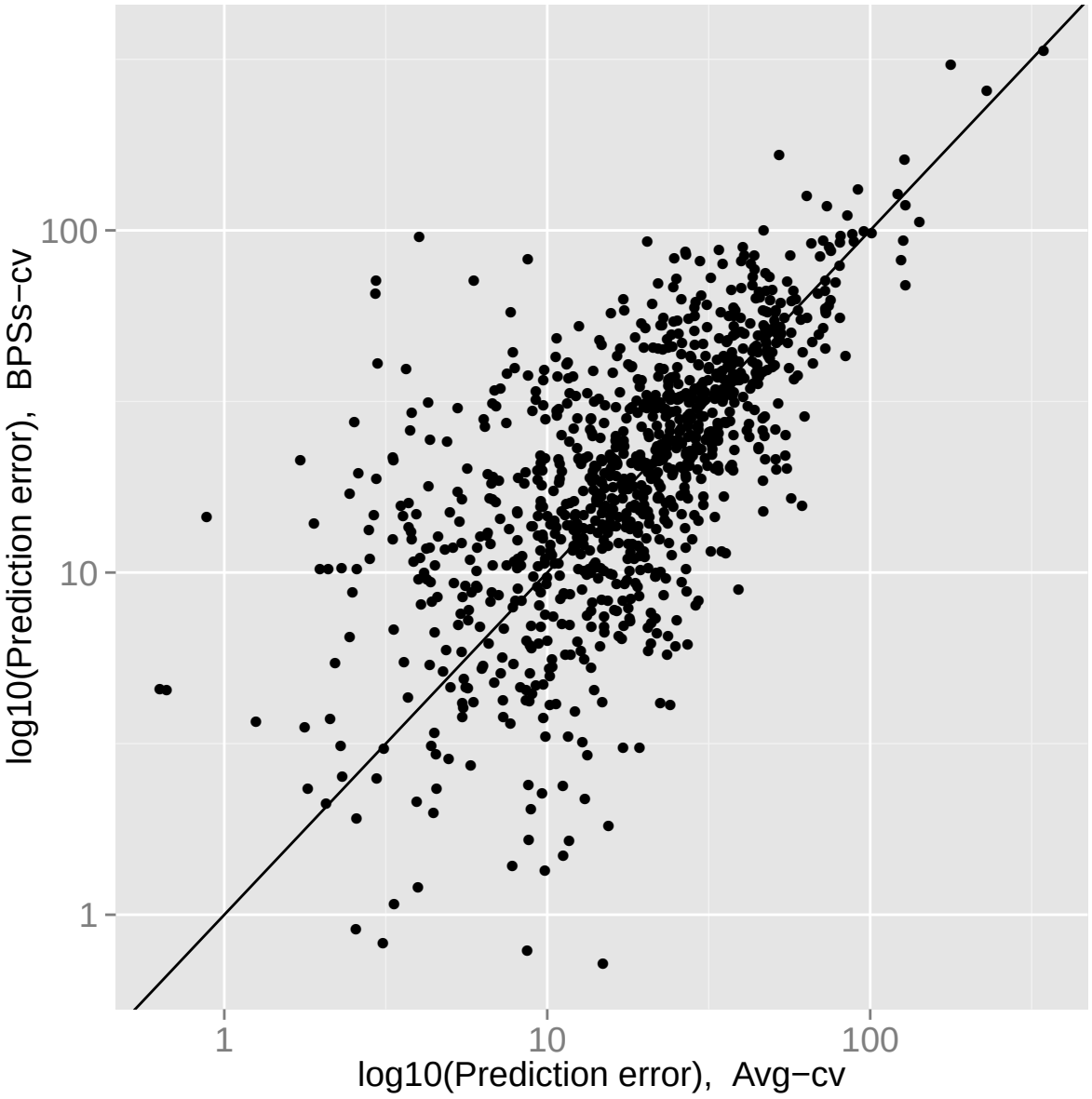

Figure 2.16: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPSs estimates for the model (2.19) with cross-validation bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. The black solid line is the 45 degrees line. 549 points out of 1,000 are above the line.

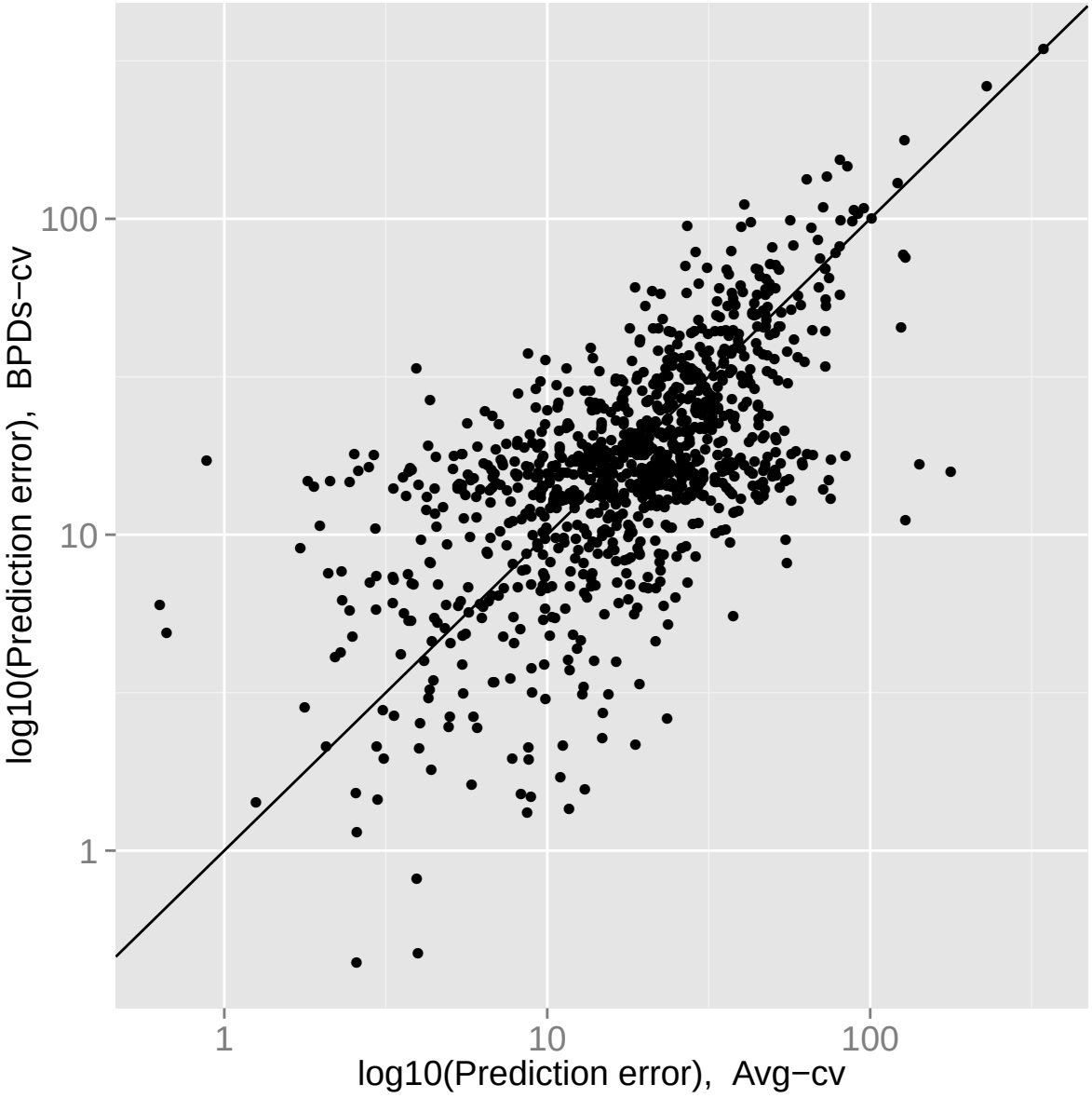

Figure 2.17: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of Avg and BPDs estimates for the model (2.19) with cross-validation bandwidths. The sum is computed along the grid of evaluation points described in Section 2.5. Refer to the caption of Table 2.2 for definition of the labels. The "-cv" represents the cross-validation bandwidths are used. The black solid line is the 45 degrees line. 536 points out of 1,000 are below the line.

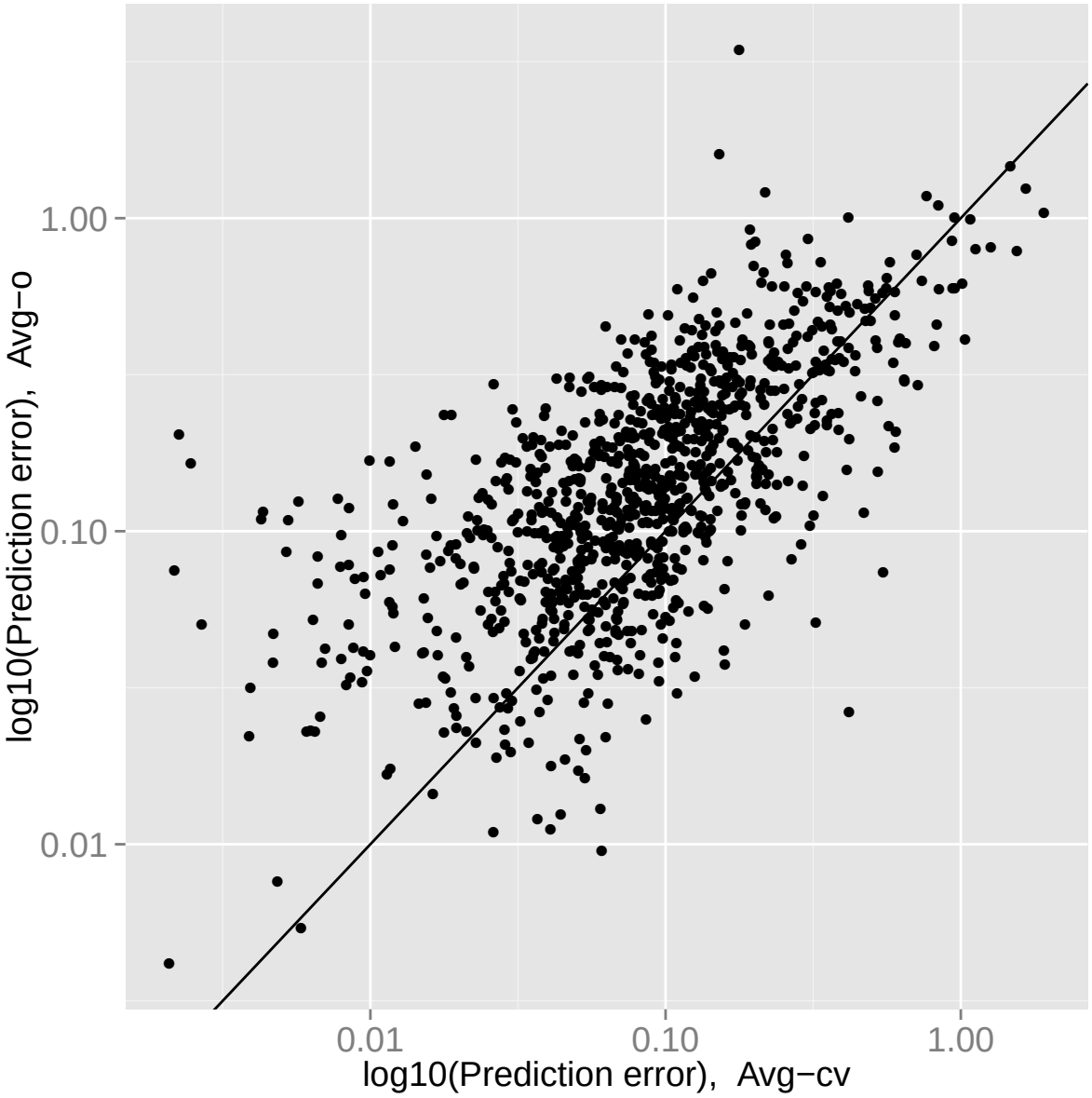

Figure 2.18: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of the pre-averaging estimator for the model (2.18). The sum is computed along the grid of evaluation points described in Section 2.5. The "-o" and "-cv" mean the oracle and the cross-validation bandwidths are used, respectively. The black solid line is the 45 degrees line. 741 points out of 1,000 are above the line.

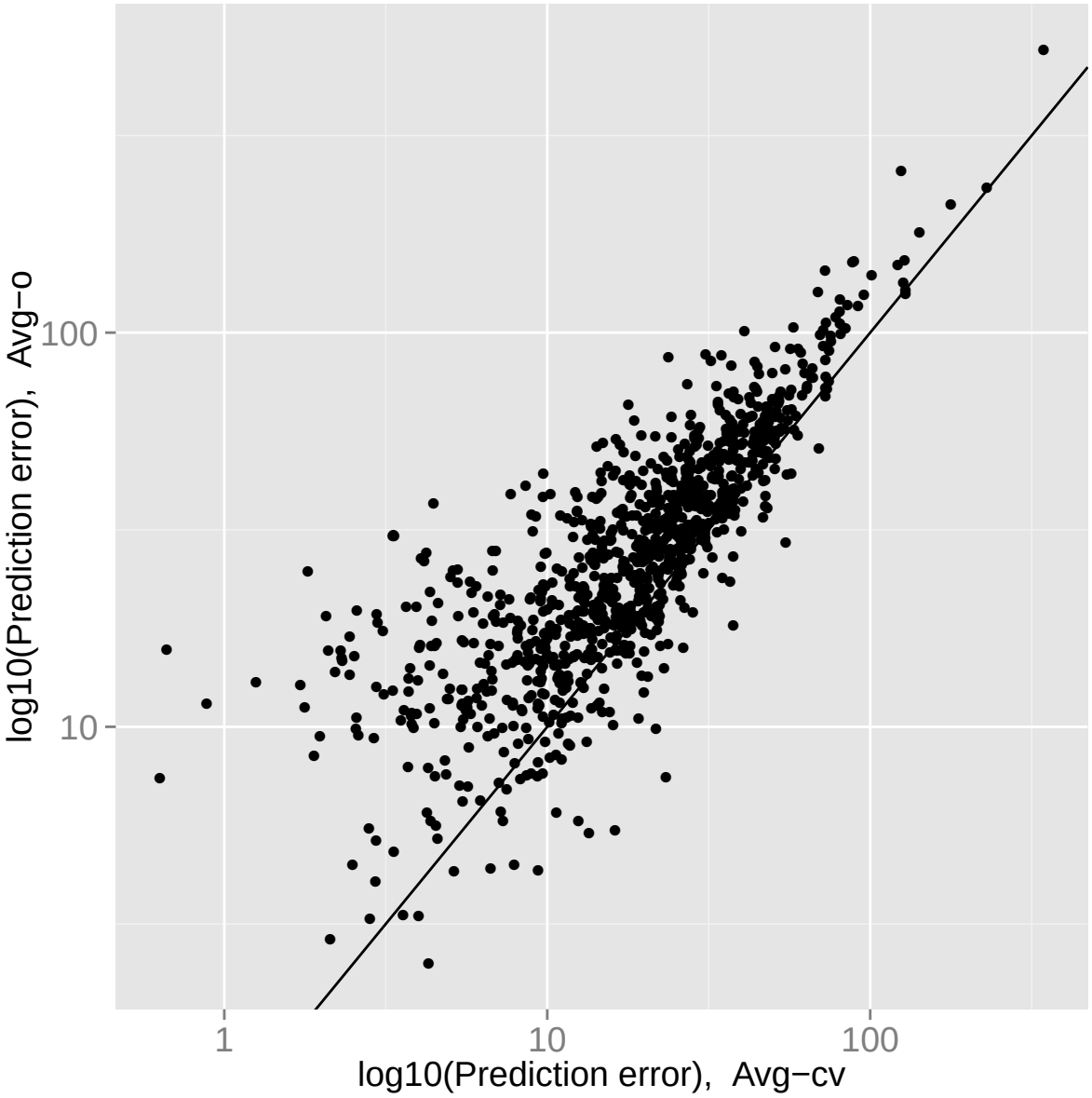

Figure 2.19: The log10-transformed pathwise invariant-density-weighted sum of squared pointwise prediction errors of the pre-averaging estimator for the model (2.19). The sum is computed along the grid of evaluation points described in Section 2.5. The "-o" and "-cv" mean the oracle and the cross-validation bandwidths are used, respectively. The black solid line is the 45 degrees line. 852 points out of 1,000 are above the line.

Chapter 3

Conclusion

In this thesis, we proposed a Nadaraya-Watson type kernel estimator of the drift coefficient of a diffusion process. Our estimator is consistent and asymptotically normal when the data are generated from a positive recurrent and strictly stationary diffusion process and are sampled discretely in time and with additive measurement errors. Our consistency and asymptotic normality result is built upon the result of Bandi and Phillips (2003), who proved consistency and asymptotic normality of the Nadaraya-Watson estimator of the drift coefficient when the data are generated from a recurrent diffusion process and are sampled discretely in time and without measurement error.

We recommended using the H-block cross-validation, proposed by Chu and Marron (1991) and Burman, Chow, and Nolan (1994), to choose the bandwidth h. Our simulation study in Section 2.5 indicates that, when the data are observed with independent and identically distributed additive measurement errors, our estimator with the H-block cross-validation bandwidth has smaller mean integrated squared error than our estimator with the oracle bandwidth, which has much smaller mean squared error than the estimators of Bandi and Phillips (2003) with the oracle bandwidth.

In our simulation study, as an alternative to our estimator, we also applied the subsampling method to the estimators of Bandi and Phillips (2003) for estimation of the drift coefficient. Our simulation study indicates that, when combined with the subsampling method, the estimators of Bandi and Phillips (2003) have mean squared errors that are as small as our estimator.

Because we have errors of observation in our model, which are not considered by Bandi

and Phillips (2003), we needed to reduce the noise caused by these errors in order to improve the accuracy of the estimate of the drift coefficient. Our approach was to construct a pre-averaged process $\{\bar{Y}_j^{r,\Delta}\}$, as in Definition 2.1. Alternatively, according to our simulation study, the subsampling method seems to be another effective way to reduce the noise.

Our estimator offers wider applicability compared to the estimators developed for the case of no measurement error, as we do often observe data with measurement errors. For example, Zhou (1996) reported the presence of measurement error in foreign exchange rates data, and Jones (2003) argued the presence of measurement error in the dataset of Aït-Sahalia (1996), the seven-day Eurodollar rates dataset.

Despite advantages of our proposed approach, the choice of the block size r and the choice of the subsampling rate are largely unsolved issues. In practice, we rely on an ad-hoc choice of r because of difficulties in using existing methods to choose r, as discussed in Section 2.4.2. In addition, our estimator involves shifts of the time-indices, as discussed right after Definition 2.2, and the effect of the shifts on the performance of the estimator is not clear. In another simulation study using the oracle bandwidth, which is not included in Section 2.5, we saw that the shifts of the time-indices increase the mean squared error of our pre-averaging estimator. However, the shifts of the time-indices do not seem necessarily to increase the mean squared error. The simulation study not included in Section 2.5 indicates that applying the shifts of the time-indices to the single-smoothing and the double-smoothing estimators, with subsampling, of Bandi and Phillips (2003) decreases their mean squared errors. The investigation of these issues could be a possible future research topic.

Another possible future research topic is estimation of $\mu(x)$ from time-irregularly observed data. An advantage of using a continuous-time process over a discrete-time process is that a continuous-time process allows us to consider time-irregularly observed data. For asymptotics, we can consider the situation where the time-difference between each pair of time-adjacent observations is a random number between 0 and infinity.

Lastly, the H-block cross-validation is a bandwidth choice method for a finite-sample, i.e. for a fixed n, and the asymptotic behavior of the H-block cross-validation bandwidth as n tends to infinity is not studied yet. As n tends to infinity, we require $\{h_n\}$ to satisfy the conditions (i), (ii) and (iii) of Theorem 2.1. The study of the asymptotic behavior of the sequence

of H-block cross-validation bandwidths in relation to the conditions of Theorem 2.1 is another possible future research topic.

Bibliography

Aït-Sahalia, Yacine. 1996. "Nonparametric pricing of interest rate derivative securities." *Econometrica* 64 (3):527–560.

Andersen, Torben G, Tim Bollerslev, Francis X Diebold, and Paul Labys. 2001. "The distribution of realized exchange rate volatility." *Journal of the American Statistical Association* 96 (453):42–55.

———. 2009. "Parametric and nonparametric volatility measurement." *Handbook of Financial Econometrics* 1:67–138.

Bandi, Federico, Valentina Corradi, and Guillermo Moloche. 2009. "Bandwidth selection for continuous-time Markov processes." *Working paper* .

Bandi, Federico M and Peter CB Phillips. 2003. "Fully nonparametric estimation of scalar diffusion models." *Econometrica* 71 (1):241–283.

Barndorff-Nielsen, Ole E, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. 2008. "Designing realized kernels to measure the ex post variation of equity prices in the presence of noise." *Econometrica* 76 (6):1481–1536.

Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. "A cross-validatory method for dependent data." *Biometrika* 81 (2):351–358.

Chapman, David A and Neil D Pearson. 2000. "Is the short rate drift actually nonlinear?" *The Journal of Finance* 55 (1):355–388.

Chu, C-K and James S Marron. 1991. "Comparison of two bandwidth selectors with dependent errors." *The Annals of Statistics* 19 (3):1906–1918.

Cleveland, William S. 1979. "Robust locally weighted regression and smoothing scatterplots." *Journal of the American Statistical Association* 74 (368):829–836.

Felsenstein, Joseph. 1985. "Phylogenies and the comparative method." *American Naturalist* 125 (1):1–15.

Florens-Zmirou, Danielle. 1993. "On estimating the diffusion coefficient from discrete observations." *Journal of Applied Probability* 30 (4):790–804.

Hall, Peter and Jeffrey D Hart. 1990. "Nonparametric regression with long-range dependence." *Stochastic Processes and Their Applications* 36 (2):339–351.

Hardle, Wolfgang. 1990. *Applied Nonparametric Regression*, vol. 27. Cambridge Univ Press.

Hart, Jeffrey D. 1994. "Automated kernel smoothing of dependent data by using time series cross-validation." *Journal of the Royal Statistical Society. Series B (Methodological)* 56 (3):529–542.

Iacus, Stefano Maria. 2008. *Simulation and Inference for Stochastic Differential Equations: with R Examples*. Springer.

———. 2009. *sde: Simulation and Inference for Stochastic Differential Equations*. URL `http://CRAN.R-project.org/package=sde`. R package version 2.0.10.

Jacod, Jean, Yingying Li, Per A Mykland, Mark Podolskij, and Mathias Vetter. 2009. "Microstructure noise in the continuous case: the pre-averaging approach." *Stochastic Processes and their Applications* 119 (7):2249–2276.

Jones, Christopher S. 2003. "Nonlinear mean reversion in the short-term interest rate." *Review of Financial Studies* 16 (3):793–843.

Jones, M Chris, James S Marron, and Simon J Sheather. 1996. "A brief survey of bandwidth selection for density estimation." *Journal of the American Statistical Association* 91 (433):401–407.

Karatzas, Ioannis Autor and Steven Eugene Shreve. 1991. *Brownian Motion and Stochastic Calculus*, vol. 113. Springer.

Kutoyants, Yu A. 2004. *Statistical Inference for Ergodic Diffusion Processes*. Springer.

Nadaraya, Elizbar A. 1964. "On estimating regression." *Theory of Probability & Its Applications* 9 (1):141–142.

Øksendal, Bernt. 1992. *Stochastic Differential Equations*. Springer.

Parzen, Emanuel. 1962. "On estimation of a probability density function and mode." *Annals of Mathematical Statistics* 33 (3):1065–1076.

Robinson, Peter M. 1983. "Nonparametric estimators for time series." *Journal of Time Series Analysis* 4 (3):185–207.

Rosenblatt, Murray. 1956. "Remarks on some nonparametric estimates of a density function." *Annals of Mathematical Statistics* 27 (3):832–837.

Ruppert, David and Matthew P Wand. 1994. "Multivariate locally weighted least squares regression." *The Annals of Statistics* 22 (3):1346–1370.

Simonoff, Jeffrey S. 1996. *Smoothing Methods in Statistics*. Springer.

Stanton, Richard. 1997. "A nonparametric model of term structure dynamics and the market price of interest rate risk." *The Journal of Finance* 52 (5):1973–2002.

Stone, Charles J. 1977. "Consistent nonparametric regression." *The Annals of Statistics* 5 (4):595–620.

Stone, Mervyn. 1974. "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (2):111–147.

Watson, Geoffrey S. 1964. "Smooth regression analysis." *Sankhyā: The Indian Journal of Statistics, Series A* :359–372.

Zhang, Lan, Per A Mykland, and Yacine Aït-Sahalia. 2005. "A tale of two time scales." *Journal of the American Statistical Association* 100 (472):1394–1411.

Zhou, Bin. 1996. "High-frequency data and volatility in foreign-exchange rates." *Journal of Business & Economic Statistics* 14 (1):45–52.