Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Kernel Smoothing for Nested Estimation with Application to Portfolio ~~Risk~~ Measurement

L. Jeff Hong, Sandeep Juneja, Guangwu Liu

To cite this article:
L. Jeff Hong, Sandeep Juneja, Guangwu Liu (2017) Kernel Smoothing for Nested Estimation with Application to Portfolio Risk Measurement. Operations Research 65(3):657-673. https://doi.org/10.1287/opre.2017.1591

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 65, No. 3, May–June 2017, pp. 657–673
ISSN 0030-364X (print), ISSN 1526-5463 (online)

CONTEXTUAL AREAS

Kernel Smoothing for Nested Estimation with Application to Portfolio Risk Measurement

L. Jeff Hong,[a] Sandeep Juneja,[b] Guangwu Liu[c]

[a] Department of Economics and Finance and Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong; [b] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai 400005, India; [c] Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong

Contact: jeffhong@cityu.edu.hk (LJH); juneja@tifr.res.in (SJ); msgw.liu@cityu.edu.hk (GL)

Received: October 5, 2012
Revised: December 13, 2013; October 21, 2014; May 14, 2016
Accepted: December 12, 2016
Published Online in Articles in Advance: April 12, 2017

Subject Classifications: finance: portfolio; simulation: applications, statistical analysis
Area of Review: Financial Engineering

https://doi.org/10.1287/opre.2017.1591

Abstract. Nested estimation involves estimating an expectation of a function of a conditional expectation via simulation. This problem has of late received increasing attention amongst researchers due to its broad applicability particularly in portfolio risk measurement and in pricing complex derivatives. In this paper, we study a kernel smoothing approach. We analyze its asymptotic properties, and present efficient algorithms for practical implementation. While asymptotic results suggest that the kernel smoothing approach is preferable over nested simulation only for low-dimensional problems, we propose a decomposition technique for portfolio risk measurement, through which a high-dimensional problem may be decomposed into low-dimensional ones that allow an efficient use of the kernel smoothing approach. Numerical studies show that, with the decomposition technique, the kernel smoothing approach works well for a reasonably large portfolio with 200 risk factors. This suggests that the proposed methodology may serve as a viable tool for risk measurement practice.

Funding: This research was partially supported by the Hong Kong Research Grants Council [GRF 145610, 613012, 16203214, 613213; N_HKUST 626/10; and ECS 155312].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2017.1591.

Keywords: nested estimation; kernel estimation; portfolio risk measurement

1. Introduction

Many simulation applications require estimating an expectation of a function of a conditional expectation. This class of problems is referred to as *nested estimation*. Here, we consider the problem of estimating

$$\alpha = \mathrm{E}[g(\mathrm{E}[Y|X])], \tag{1}$$

via Monte Carlo simulation, where X is a random vector in $\mathcal{R}^d$ with $d \geqslant 1$, Y is a random variable in $\mathcal{R}$, and $g(\cdot)$ may be a nonlinear function.

This problem has important applications in financial risk management, especially in portfolio risk measurement. Consider risk measurement of a portfolio that consists of a number of financial instruments, including many sophisticated financial derivatives that need Monte Carlo simulation for accurate pricing. Suppose that the random vector X denotes risk factors at a future time horizon up to which we measure the risk, and $\mathrm{E}[Y|X]$ is the portfolio loss given an outcome of X. This is a conditional expectation as it requires knowing the cash flows in the portfolio that may be a deterministic function of X, as well as conditional mark-to-market values of the derivatives in the portfolio. The latter typically are conditional expectations under a suitable pricing martingale measure (see, e.g., Duffie 1996, Shreve 2004). Typically, $\mathrm{E}[Y|X]$ is not known in a closed form as a function of X and has to be estimated via simulation. Then, the inner expectation in (1) reprices the portfolio given outcomes of the risk factors, and the outer expectation calculates risk of the portfolio.[1] Specification of the function $g(\cdot)$ depends on risk measures being used. For instance, when the risk is measured by the second moment of the portfolio loss, $g(\cdot)$ is a quadratic function; when the risk is measured by the probability that the portfolio loss is larger than some threshold value, $g(\cdot)$ is an indicator function; and when the risk is measured by average loss given that the portfolio loss is larger than some threshold value, $g(\cdot)$ is a hockey stick function. Note that the last two examples are directly linked to value at risks and tail conditional expectations (also known as expected shortfalls) that are two very important risk measures; see, e.g., Baysal and Staum (2008), Bank for International Settlements (2012), and Broadie et al. (2015).

Due to high computational costs, portfolio risk measurement is often performed by using highly simplified models for portfolio losses (see, e.g., Rouvinez 1997, Britten-Jones and Schaefer 1999, and Duffie and Pan 2001 for delta-gamma approximations), so that the inner expectation may be evaluated analytically for

different outcomes of the risk factors. However, these simplified models may fail to capture important features of market dynamics and thus perform poorly. To avoid drawbacks of highly simplified models, more realistic models of market dynamics need to be used for portfolio repricing. Under these models, the inner expectation may not have a closed-form expression and estimation via simulation is often necessary. In this paper, we consider situations where evaluation of the inner and outer expectations requires Monte Carlo simulation.

Besides applications to portfolio risk measurement, other applications of (1) include pricing of complex derivatives such as collateralized debt obligations (CDOs). For instance, X may represent underlying variables such as credit intensities and default event indicators, and Y may represent portfolio loss of defaultable assets referencing the CDO. Then, α may be the price of a tranche of the CDO with an attachment point a and a detachment point b ($a < b$), which can be written as (see, e.g., Gordy and Juneja 2006)

$$\begin{aligned}\alpha &= \mathrm{E}[\min(b, \max(\mathrm{E}[Y|X], a)) - a] \\ &= \mathrm{E}[(\mathrm{E}[Y|X] - a)^+ - (\mathrm{E}[Y|X] - b)^+],\end{aligned}$$

where $x^+ \equiv \max(x, 0)$ is known as a hockey-stick function. Another application corresponds to pricing compound options (see Glasserman 2004). For instance, consider a call option expiring at time T_1 to buy a call option expiring at time $T_2 > T_1$. Assuming for notational simplicity that the discount rate is zero, the price of this compound option may be expressed as

$$\mathrm{E}[(\mathrm{E}[(S(T_2) - K_2)^+|S(T_1)] - K_1)^+],$$

where $S(t)$ denotes the price of the underlying stock at time t, K_2 denotes the strike price of the call option on this stock that expires at time T_2, and K_1 denotes the strike price of the option on the call option that expires at time T_1.

A typical approach to nested estimation is to use *nested simulation*. Specifically, nested simulation refers to a two-level simulation procedure. In the outer level, one simulates a number of scenarios of X. Then, in the inner level, one simulates a number of samples of Y for each generated X to estimate $\mathrm{E}[Y|X]$. Traditionally, it is perceived that nested simulation imposes an unacceptable computational burden. However, Gordy and Juneja (2010) show that this is not necessarily the case and that a relatively small number of samples in the inner level may yield reasonably accurate estimates for α in (1), particularly for large portfolios. Along the line of nested simulation, recently some work has been done to improve its efficiency by exploiting structural information of the function $g(\cdot)$. For instance, Lan et al. (2010) use ranking-and-selection techniques to improve the efficiency of the inner estimation, Liu et al. (2010) study how to adaptively allocate computational effort based on ranking-and-selection techniques for nested estimation of expected shortfalls, and Broadie et al. (2011) propose algorithms to sequentially allocate computational effort to inner-level simulations for estimating probabilities of large portfolio losses.

To explore other approaches to nested estimation, some researchers (see, e.g., Liu and Staum 2010, Broadie et al. 2015) have observed That, in practice, the inner expectation $\mathrm{E}[Y|X = x]$ is typically *a continuous function of* x. This suggests that the value of the inner expectation can be inferred statistically based on its properties as a function of x, in a parametric or a nonparametric manner. This leads to a second class of estimation techniques that we refer to as *smoothing approaches*, via which the values of the inner expectation on the space of risk factors may be inferred by sampling a few realizations of these factors.

Smoothing approaches include the least-squares method (LSM) and the stochastic kriging approach. LSM was proposed by Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001) for American option pricing; see also the previous work of Carriere (1996) that includes general discussions of nonparametric regression techniques. Recently, it was used by Broadie et al. (2015) for portfolio risk measurement. The key idea of the LSM is to posit that the inner expectation $\mathrm{E}[Y|X = x]$ may be represented as a linear combination of a number of user-specified basis functions for each x, and then use a small number of inner-level samples to estimate these linear coefficients. The key idea of the stochastic kriging approach is to build a spatial metamodel of $\mathrm{E}[Y|X = x]$ based on the samples Y taken from a few x values, and then infer the values of the inner expectation using this metamodel (see, e.g., Liu and Staum 2010). Unlike the LSM, the stochastic kriging approach does not assume a parametric relationship between $\mathrm{E}[Y|X = x]$ and x, and is therefore more flexible and robust. See Ankenman et al. (2010) for details of stochastic kriging techniques.

Even though results in the literature show that the smoothing Approaches, in general, work well for portfolio risk measurement problems, the problems considered in these papers are often of small dimensions. Intuitively, however, smoothing approaches may be affected by the dimension of the risk factors and may be subject to the curse of dimensionality. Moreover, many practical portfolio risk measurement problems are of high dimensions and may have more than a hundred risk factors. Therefore, in this paper, we consider the following two important research questions on smoothing approaches and their applications in portfolio risk measurement:

1. How is the performance of a smoothing approach affected by the dimension of the risk factors?

2. How can we design an efficient smoothing approach that can be applied to portfolio risk measurement problems with a large number of risk factors?

To answer the first question, we study a simple smoothing approach that estimates $\mathrm{E}[Y|X=x]$ by the average of all the inner samples whose X values are within a hypercube centered at x. This is a special case of the Nadaraya-Watson (NW) kernel estimators for nonparametric regression (see, e.g., the monograph of Bosq 1998), where the kernel function is chosen to be a d-dimensional uniform distribution (hence a hypercube centered at x). Hence we call it the *kernel smoothing approach*. We analyze the asymptotic properties of the approach for different types of $g(\cdot)$, including smooth, nondifferentiable, and discontinuous cases. Our main finding is that, at the optimal choice of the side length of the hypercube and for all three cases of $g(\cdot)$, the rate of convergence of the mean squared error (MSE) of the kernel-smoothing estimator is $\Gamma^{-\min(1,4/(d+2))}$ as $\Gamma \to \infty$, where Γ denotes the available sampling budget and d is the dimension of X. Therefore the performance of the estimator deteriorates as the dimension d increases. Moreover, compared to the nested simulation approach whose optimal rate of convergence of the MSE is $\Gamma^{-2/3}$ regardless of the dimension of X (see Lee 1998 and Gordy and Juneja 2010), the kernel smoothing approach has the same or a better rate of convergence when $d \leqslant 4$, and a slower rate of convergence when $d > 4$.

To answer the second question, we note that for many practical portfolio risk measurement problems, the number of risk factors is often very large, e.g., over 100, which renders the direct use of the kernel smoothing approach impractical. However, in portfolio risk measurement, we find that a high-dimensional problem may often be decomposed into low-dimensional ones, which allow an efficient use of the kernel smoothing approach. Such a decomposition technique relies on an observation that a portfolio loss is typically a summation of losses of individual financial instruments that form the portfolio, while individual financial instruments may depend on only a small number of risk factors. In other words,

$$\mathrm{E}[Y|X] = \sum_{i=1}^{l} \mathrm{E}[Y^i|X] = \sum_{i=1}^{l} \mathrm{E}[Y^i|X^j,\ j \in \mathcal{B}_i], \quad (2)$$

where l is the number of groups, $\mathcal{B}_i$ is a subset of $\{1,\dots,d\}$, Y^i denotes the total loss of the financial instruments in group i, which depends on only the risk factors $\{X^j,\ j \in \mathcal{B}_i\}$, and Y^i satisfies $\sum_{i=1}^{l} Y^i = Y$.

With the structure in (2), one may apply the kernel smoothing approach to each $\mathrm{E}[Y^i|X^j,\ j \in \mathcal{B}_i]$ for $i = 1,\dots,l$. The approach may work very well when the numbers of elements in $\mathcal{B}_i$'s are small, i.e., individual financial instrument depends on only a few risk factors. This is indeed the case for many financial instruments commonly traded in financial markets. According to the statistics of the World Federation of Exchanges (2014), in terms of the number of contracts traded, more than 90% of options traded on exchanges worldwide in 2013 are written on a single underlying asset that may be a stock, a stock index, an ETF, an exchange rate, or a commodity. Price dynamics of these underlying assets can often be modeled with fewer than three risk factors, such as the Black-Scholes model and Heston's stochastic volatility (SV) model. Therefore the idea of decomposition is particularly attractive. To illustrate the practical value of the proposed decomposition technique, we show in Section 6 that it works well for a representative example with 200 risk factors.

On the implementation side, we present algorithms for fast computation of the NW estimator. These algorithms work well for a small d. However, their benefit is not substantial for a large d. To circumvent this difficulty, we suggest using a variant of the kernel smoothing estimators, the so-called k-nearest neighbor (kNN) estimator for large dimensions. The kNN estimator shares the same spirit with the NW estimator in that it also approximates the expectation $\mathrm{E}[Y|X=x]$ by taking the average of all inner samples whose X values are within a neighborhood of x. The difference between these two estimators is that the kNN estimator chooses the set of kNNs for some fixed integer k as the neighborhood, while the NW estimator chooses the hypercube with side length h and centered at x as the neighborhood. We find that the kNN estimator appears to have a better performance than the nested simulation estimator even for problems with dimensions as large as 20 and it can be calculated efficiently. It should be pointed out that in this paper, asymptotic results are provided only for the NW estimator. While it is expected that the kNN estimator may have similar asymptotic properties, the proofs are technically more challenging and are left for future research.

The rest of the paper is organized as follows. Section 2 reviews the background of nonparametric kernel regression and proposes the kernel smoothing approach. Asymptotic results of the NW estimators are provided in Section 3, and implementation issues are discussed in Section 4. In Section 5, we propose a decomposition technique when the kernel smoothing approach is applied to portfolio risk measurement. Numerical results are presented in Section 6, followed by conclusions in Section 7. Some lengthy discussions and technical proofs are provided in the appendix.

2. A Kernel Approach

2.1. Background

For notational ease, we let $m(X)$ denote $\mathrm{E}[Y|X]$, where the conditional expectation $m(x) \equiv \mathrm{E}[Y|X=x]$ is also known as a regression function in statistics. Then, a

key issue of estimating $\alpha = \mathrm{E}[g(m(X))]$ is how to estimate $m(x)$. Estimation of $m(x)$ has been studied extensively in the literature, see, e.g., the monograph of Bosq (1998). Among various estimation methods for $m(x)$, in this paper, we consider the well-known NW kernel estimator, which originates from Nadaraya (1964) and Watson (1964). Specifically, given n=independent and identically distributed (i.i.d.) observations of (X, Y), denoted by $\{(X_i, Y_i), 1 \leqslant i \leqslant n\}$, the NW estimator of $m(x)$ is

$$\tilde{m}_n(x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}, \tag{3}$$

where $K_h(x) = (1/h^d)K(x/h)$, K is a kernel function that is a symmetric density function on $\mathcal{R}^d$, and h is a bandwidth parameter satisfying $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

It is well known (see, e.g., Section 4.5 of Härdle 1990) that the choice of the kernel function K is not critical. Therefore, in the rest of the paper, we let

$$K(u) = \prod_{i=1}^d 1_{\{-1/2 \leqslant u^i \leqslant 1/2\}} \tag{4}$$

for $u = (u^1, \ldots, u^d)$, which is called a d-dimensional uniform kernel.[2]

With the uniform kernel, an intuitive explanation of the kernel method is straightforward: for an outer-level scenario x, $m(x)$ is estimated by the average of inner-level samples taken from those outer-level scenarios lying in a neighborhood of x, where the neighborhood is chosen to be a hypercube centered at x with a length of h for each side.

Throughout the paper, we use the notation of $o(\cdot)$ and $O(\cdot)$, where for two deterministic sequences a_n and b_n, $a_n = o(b_n)$ if $\lim_{n\to\infty}(a_n/b_n) = 0$, and $a_n = O(b_n)$ if $\limsup_n a_n/b_n < \infty$. We also say that a_n is *asymptotically smaller* than b_n if $\lim_{n\to\infty}(a_n/b_n) = 0$.

Let $B_n(x)$ and $V_n(x)$ denote the bias and variance of $\tilde{m}_n(x)$, respectively. We summarize the asymptotic properties of $\tilde{m}_n(x)$ in the following lemma.

Lemma 2.1. *Let* $\sigma^2(x) = \mathrm{E}[Y^2|X = x] - m^2(x)$, $\mu_K = \int_{\mathcal{R}^d} u^T u K(u)\,du$, *and* $c_K = \int_{\mathcal{R}^d} K^2(u)\,du$. *Then, under regularity conditions discussed in Remark 1,*

$$B_n(x) = B(x)h^2 + o(h^2), \tag{5}$$

$$V_n(x) = \frac{V(x)}{nh^d} + \frac{o(1)}{nh^d}, \tag{6}$$

where

$$B(x) = \frac{\mu_K}{2f(x)}\left[\mathrm{tr}\left\{\frac{\partial}{\partial x}\frac{\partial}{\partial x^T}[m(x)f(x)]\right\} - m(x)\mathrm{tr}\left\{\frac{\partial}{\partial x}\frac{\partial}{\partial x^T}f(x)\right\}\right], \quad V(x) = \frac{c_K\sigma^2(x)}{f(x)},$$

with $f(x)$ *being the density function of* X, *and* A^T *and* $\mathrm{tr}(A)$ *denoting transpose and trace of a matrix* A, *respectively.*

Moreover, $\tilde{m}_n(x)$ *satisfies the following central limit theorem*:

$$\sqrt{nh^d}(\tilde{m}_n(x) - m(x)) \Rightarrow \sqrt{c_K\sigma^2(x)/f(x)}\cdot Z, \tag{7}$$

if $nh^{d+4} \to 0$ *as* $n \to \infty$, *where "*$\Rightarrow$*" denotes convergence in distribution and* Z *denotes a standard normal random variable.*

Remark 1. Key regularity conditions of Lemma 2.1 include *existence and continuity conditions*: $f(x)$ and $\mathrm{E}[Y^2|X = x]$ exist, $f(x) > 0$, $m(x)$ and $f(x)$ are thrice continuously differentiable at x and have bounded third-order derivatives; and *bandwidth conditions*: $h \to 0$, $nh^d \to \infty$ and $nh^{d+4} \to 0$ as $n \to \infty$. Moreover, for Lemma 2.1 to hold, one also needs some other regularity conditions that are imposed for mathematical convenience. Different sets of such conditions have been proposed in the literature of nonparametric statistics. For ease of exposition, we omit the listing of these regularity conditions but refer interested readers to Bosq (1998), Bierens (1985), Härdle (1990), Jennen-Steinmetz and Gasser (1988), and Gasser and Engel (1990) for detailed discussions.

In this paper, we assume that the results of Lemma 2.1 hold. These results serve as the basis for subsequent analysis.

2.2. A Kernel Estimator

Given the NW estimator, we may estimate $\alpha = \mathrm{E}[g(m(X))]$ by using either

$$\bar{M}_n = \frac{1}{n}\sum_{k=1}^n g(\tilde{m}_n(X_{0,k}))$$

or

$$\tilde{M}_n = \frac{1}{n}\sum_{k=1}^n g(\tilde{m}_n(X_k)),$$

where $\{X_{0,k}, 1 \leqslant k \leqslant n\}$ are i.i.d. samples of X that are independent of $\{(X_k, Y_k), 1 \leqslant k \leqslant n\}$.

The estimators $\bar{M}_n$ and $\tilde{M}_n$ are slightly different, as $\bar{M}_n$ requires an additional set of samples of X. Both of them are important in practical applications. For instance, one may use $\bar{M}_n$ in a two-phase procedure (see, e.g., construction of low-biased estimators in Broadie and Glasserman 2004) for pricing American options, where in the first phase, a set of samples is used to approximate the optimal exercise policy, and in the second phase, a different set of samples is used to evaluate the option price. But in applications such as portfolio risk measurement, $\tilde{M}_n$ may be a more natural estimator, because arguably an additional set of samples may not be necessary.

Note that $\bar{M}_n$ focuses mainly on the setting where only one inner-level sample is generated for each outer-level sample. In a more general setting, one may set the number of outer-level samples to be l, and for each

outer-level sample, the number of inner-level samples is n/l, where l and n/l are integers. For $i = 1, \ldots, l$, let $\{Y_{ij}, 1 \leqslant j \leqslant n/l\}$ denote the inner-level samples for X_i. Define

$$\bar{Y}_i = \frac{l}{n} \sum_{j=1}^{n/l} Y_{ij}, \quad i = 1, \ldots, l,$$

and the kernel estimator

$$\tilde{m}_{nl}(x) = \frac{\sum_{i=1}^{l} \bar{Y}_i K_{h_l}(x - X_i)}{\sum_{i=1}^{l} K_{h_l}(x - X_i)},$$

where the bandwidth is a function of l, and it is made explicit by the notation h_l.

Suppose that another set of N samples of X, denoted by $\{X_{0,j}, 1 \leqslant j \leqslant N\}$, are used to estimate α. Then, a more general estimator of α is

$$\bar{M}_{N,n,l} = \frac{1}{N} \sum_{j=1}^{N} g(\tilde{m}_{nl}(X_{0,j})). \tag{8}$$

In what follows, we focus our discussion mainly for the estimator $\bar{M}_n$, while asymptotic properties of $\bar{M}_{N,n,l}$ are also discussed at the end of Section 3. The key idea of the kernel approach is that once we have generated a set of samples, $\{(X_i, Y_i), 1 \leqslant i \leqslant n\}$, then for any $x \in \mathcal{R}^d$, sampled or not, we may infer the value of $g(m(x))$ by using this set of samples. Therefore, no nested simulation is required, because only one inner-level sample is simulated for each outer-level scenario. Compared to the nested simulation approach, the kernel approach may provide a saving in computational cost, especially when simulating Y is computationally expensive, which is often the case in practical applications.

To measure the performance of the kernel approach, one needs to know the rates of convergence of the estimators. As far as we are aware of, results on rates of convergence of estimators such as $\bar{M}_n$ and $\tilde{M}_n$ are still unknown in the literature. One of our objectives is to fill this gap. A convergence result will also show how the dimension of X affects the performances of the kernel estimators.

To analyze the rates of convergence of the kernel estimators, in this paper, we focus mainly on $\bar{M}_n$, while arguing that $\tilde{M}_n$ has the same rate of convergence as $\bar{M}_n$. We do so for the sake of conciseness of the presentation, because a rigorous analysis of $\tilde{M}_n$ may require additional technicality without adding much value to the central ideas of the paper. Without going deep into the required technical conditions, we argue in Section EC.2.1 of the e-companion that MSEs of $\bar{M}_n$ and $\tilde{M}_n$ have the same rate of convergence.

3. Asymptotic Properties

This section is devoted to the asymptotic analysis of $\bar{M}_n$. To be meaningful, we let $g(\cdot)$ be a nonlinear function, because the problem is trivial when $g(\cdot)$ is linear. Specifically, if $g(t) = at + b$ for some constants a and b, then it can be easily seen that

$$\alpha = \mathrm{E}(g(\mathrm{E}[Y|X])) = \mathrm{E}(a\mathrm{E}[Y|X] + b) = \mathrm{E}[aY + b],$$

and the estimation of α becomes a simple problem.

In the latter analysis, we consider three types of function $g(\cdot)$: a *smooth function*, a *hockey-stick function*, i.e., $g(t) = t^+$, and an *indicator function*, i.e., $g(t) = 1_{\{t \geqslant 0\}}$. The latter two functions differ from the first one in smoothness. Specifically, the hockey-stick function has a nondifferentiable but continuous point, while the indicator function has a discontinuous point. We focus on these three types of functions for two reasons. Firstly, they are of great practical importance. For instance, when we estimate expected shortfalls or probabilities of large portfolio losses, the functions $g(\cdot)$ of interest typically have the forms of a hockey-stick function or an indicator function, respectively. Secondly, for a more general function with a finite number of nondifferentiable and/or discontinuous points, we may decompose it as a linear combination of these three types of functions, leading to a straightforward conclusion that its corresponding kernel estimator shares the same rate of convergence as that of the three types of functions. An intuition behind the decomposition is that adding an appropriately chosen linear combination of hockey-stick and indicator functions to a function with a finite number of nondifferentiable and/or discontinuous points makes it differentiable everywhere. A detailed discussion on the decomposition is given in Section EC.2.2 of the e-companion.

Henceforth, we let $\mathcal{D} \subseteq \mathcal{R}^d$ denote the support of X, and assume that X is a continuous random vector with a density function $f(\cdot)$ on the support. We note from Lemma 2.1 that $\sqrt{nh^d}(\tilde{m}_n(x) - m(x))$ has a nontrivial limit for $x \in \mathcal{D}$. For notational ease, we define

$$Z_n(x) = \sqrt{nh^d}(\tilde{m}_n(x) - m(x)). \tag{9}$$

Under mild conditions, $Z_n(x)$ converges to a normal distribution for each x. Furthermore, we make the following assumption.

Assumption 1. $\sup_n \mathrm{E}[|Z_n(X)|^4] < \infty$.

3.1. Analysis for a Smooth Function

We first consider the relatively simple case where $g(\cdot)$ is a smooth function. Specifically, we assume that $g(\cdot)$ is thrice differentiable, and has bounded second- and third-order derivatives. We study the asymptotic MSE of $\bar{M}_n$. To do so, we consider its bias and variance separately.

3.1.1. Asymptotic Bias. As the asymptotic bias of $\tilde{m}_n(X)$ can be easily analyzed based on the known results of kernel estimation, it is natural to link the bias of $g(\tilde{m}_n(X))$ to that of $\tilde{m}_n(X)$. Because $g(\cdot)$ is thrice differentiable, using Taylor expansion, we have

$$\begin{aligned}&\mathrm{E}[g(\tilde{m}_n(X)) - g(m(X))]\\ &\quad = \mathrm{E}[g'(m(X))(\tilde{m}_n(X) - m(X))]\\ &\qquad + \mathrm{E}\left[\frac{g''(m(X))}{2}(\tilde{m}_n(X) - m(X))^2\right]\\ &\qquad + \mathrm{E}\left[\frac{g'''(\Xi)}{6}(\tilde{m}_n(X) - m(X))^3\right], \end{aligned} \tag{10}$$

where Ξ is a random variable that lies between $\tilde{m}_n(X)$ and $m(X)$.

The basic idea of the bias analysis is based on an argument that the third term on the right-hand side (RHS) of Equation (10) is asymptotically smaller than the first two terms. In particular, we show in Section A.2.1 of the appendix that if Assumption 1 holds and $|g'''(t)|$ is bounded for all t, then the third term is of order $(nh^d)^{-3/2}$, which shall be seen to be negligible when compared to the first two terms. More precisely, by Lemma 2.1, the first term can be written as

$$\begin{aligned}&\mathrm{E}[g'(m(X))(\tilde{m}_n(X) - m(X))]\\ &\quad = \mathrm{E}(g'(m(X))\mathrm{E}[\tilde{m}_n(X) - m(X)|X])\\ &\quad = \int_{\mathscr{D}} g'(m(x))B_n(x)f(x)\,dx\\ &\quad = \int_{\mathscr{D}} g'(m(x))B(x)h^2(1+o_x(1))f(x)\,dx\\ &\quad = h^2\mathrm{E}[g'(m(X))B(X)](1+o(1)), \end{aligned} \tag{11}$$

where the small term $o_x(1)$ depends on x, and the second to the last equality follows from (5) of Lemma 2.1.[3]

In a similar manner, we can establish the asymptotic order of the second term:

$$\begin{aligned}&\mathrm{E}[g''(m(X))(\tilde{m}_n(X) - m(X))^2]\\ &= h^4\mathrm{E}[g''(m(X))B^2(X)] + \frac{1+o(1)}{nh^d}\mathrm{E}[g''(m(X))V(X)]. \end{aligned} \tag{12}$$

Recall that the third term is of order $(nh^d)^{-3/2}$. Because $nh^d \to \infty$ as $n \to \infty$, it can be easily seen that the third term is negligible when compared to the second term.

Then, by Equations (11) and (12), we establish the following result on the asymptotic bias of $\bar{M}_n$.

Proposition 3.1. *Suppose that Assumption 1 is satisfied, and the results of Lemma 2.1 hold for all $x \in \mathscr{D}$. Assume that the following conditions are satisfied*:

(a) *The function $g(\cdot)$ is thrice differentiable with a bounded third-order derivative.*

(b) *The random variables $g'(m(X))B(X)$, $g''(m(X)) \cdot B^2(X)$, and $g''(m(X))V(X)$ have finite expectations.*

Then,

$$\begin{aligned}\mathrm{E}(\bar{M}_n) - \alpha &= \mathrm{E}[g'(m(X))B(X)]h^2(1+o(1))\\ &\quad + \frac{1}{2}\mathrm{E}[g''(m(X))V(X)]\frac{1+o(1)}{nh^d}. \end{aligned} \tag{13}$$

Proposition 3.1 reveals an important insight on how the errors of the inner estimator affect the bias of $\bar{M}_n$. Recall that

$$\bar{M}_n = \frac{1}{n}\sum_{k=1}^{n} g(\tilde{m}_n(X_{0,k})),$$

and by Lemma 2.1, the inner estimator $\tilde{m}_n(X_{0,k})$ has a bias of order h^2 and a variance of order $(nh^d)^{-1}$. Proposition 3.1 shows that the bias and the variance of $\tilde{m}_n(X_{0,k})$ affect the bias of $\bar{M}_n$. Specifically, the first term on the RHS of (13) is contributed by the bias of the inner estimator, while the second term is contributed by its variance. Note that nh^d in the denominator is proportional to the expected number of samples in the hypercube.

3.1.2. Asymptotic Variance. To analyze the variance of $\bar{M}_n$, we first note that

$$\begin{aligned}\mathrm{Var}(\bar{M}_n) &= \frac{1}{n}\mathrm{Var}[g(\tilde{m}_n(X))]\\ &\quad + \left(1 - \frac{1}{n}\right)\mathrm{Cov}(g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))). \end{aligned} \tag{14}$$

The variance term and the covariance term on the RHS of (14) will be analyzed separately. To analyze the variance term, we use the following lemma, whose proof is provided in Section A.1 of the appendix.

Lemma 3.1. *Suppose that the discontinuity set of g, denoted by D_g, satisfies $\Pr\{m(X) \in D_g\} = 0$, and there exist a constant C and an integer $p \geqslant 0$ such that $|g(t)| \leqslant C|t|^p$. Assume that $\tilde{m}_n(X)$ converges to $m(X)$ in probability as $n \to \infty$, and there exists some $\delta > 0$ such that*

$$\sup_n \mathrm{E}[|\tilde{m}_n(X)|^{2p+\delta}] < \infty.$$

Then,

$$\mathrm{Var}[g(\tilde{m}_n(X))] = \mathrm{Var}[g(m(X))] + o(1).$$

Lemma 3.1 shows that under mild regularity conditions, the variance of $\mathrm{Var}[g(\tilde{m}_n(X))]$ comes mainly from the variance of $g(m(X))$. The lemma can be directly applied to the variance term in (14). Next, we consider the covariance term. It turns out the covariance term is always nonnegative, i.e.,

$$\mathrm{Cov}[g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))] \geqslant 0, \tag{15}$$

whose proof is provided in Section A.2.2 of the appendix.

Moreover, as discussed below we also find an upper bound of the covariance term. Since $X_{0,1}$ and $X_{0,2}$

are independent, and are independent of $\{(X_k, Y_k), 1 \leqslant k \leqslant n\}$, it can be verified that

$$n\mathrm{Cov}[g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))] = n\left\{\int_{\mathscr{D}}\int_{\mathscr{D}} \mathrm{Cov}(g(\tilde{m}_n(x_1)), g(\tilde{m}_n(x_2)))f(x_1)f(x_2)\,dx_1\,dx_2\right\}.$$

By the definition of the NW estimator, $\tilde{m}_n(x)$ is essentially the sample average of Y_k's whose associated X_k's fall into a hypercube centered at x with a side length h. Therefore $\tilde{m}_n(x_1)$ and $\tilde{m}_n(x_2)$ are dependent only if the hypercubes, with the same side length h and centered at x_1 and x_2, respectively, have an intersection. It can be shown that the volume of the intersection is of order h^d. We show in Section A.2.3 of the appendix that if $f(x)$ is bounded by a constant C_1 for all $x \in \mathscr{D}$, then

$$n\mathrm{Cov}[g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))] \leqslant C_1 nh^d \int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x))]f(x)\,dx. \tag{16}$$

Furthermore, we show in Section A.2.4 of the appendix that if $|g''(t)|$ is bounded by a constant C_2 and $\mathrm{E}[(g'(m(X)))^4] < \infty$, then

$$nh^d \int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x))]f(x)\,dx \leqslant 2\sqrt{\mathrm{E}[(g'(m(X)))^4]\mathrm{E}[Z_n^4(X)]} + \frac{C_2^2}{2nh^d}\mathrm{E}[Z_n^4(X)], \tag{17}$$

where Z_n was defined in (9).

Because $\sup_n \mathrm{E}[Z_n^4(X)] < \infty$ by Assumption 1 and $nh^d \to \infty$ as $n \to \infty$, combining Equations (16) and (17), we can see that when n is large enough,

$$n\mathrm{Cov}(g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))) \leqslant 2C_1(\sqrt{\mathrm{E}[(g'(m(X)))^4]\mathrm{E}[Z_n^4(X)]} + 1).$$

Then, $\mathrm{Var}(\bar{M}_n)$ is of order n^{-1} by (14). This result is summarized in the following proposition.

Proposition 3.2. *Suppose that Assumption 1 and conditions in Lemma 3.1 are satisfied, and the results of Lemma 2.1 hold for all $x \in \mathscr{D}$. Assume that there exist constants C_1 and C_2 such that $f(x) \leqslant C_1$ for all $x \in \mathscr{D}$, and $|g''(t)| \leqslant C_2$ for all t, and $\mathrm{E}[(g'(m(X)))^4] < \infty$. Then,*

$$\mathrm{Var}(\bar{M}_n) = \frac{\mathrm{Var}[g(m(X))] + c_n + o(1)}{n}, \tag{18}$$

where c_n satisfies $0 \leqslant c_n \leqslant 2C_1(\sqrt{\mathrm{E}[(g'(m(X)))^4]\mathrm{E}[Z_n^4(X)]} + 1) < \infty$ when n is large enough.

Proposition 3.2 shows that the variance of $\bar{M}_n$ is of order n^{-1}. It turns out that the covariance term on the RHS of (14) does not affect the order of the variance, although it may inflate the variance by a factor that is bounded by a constant.

3.1.3. Asymptotic Mean Squared Error.

Combining the results of Propositions 3.1 and 3.2, we immediately establish the asymptotic MSE of $\bar{M}_n$, which is summarized in the following theorem.

Theorem 3.1. *Suppose that Assumption 1 and conditions in Lemma 3.1 are satisfied, and the results in Lemma 2.1 hold for all $x \in \mathscr{D}$. Assume that there exist constants C_1 and C_2 such that $f(x) \leqslant C_1$ for all $x \in \mathscr{D}$, and $|g''(t)| \leqslant C_2$ for all t. Further assume that the following conditions are satisfied:*

(a) The function $g(\cdot)$ is thrice differentiable with a bounded third-order derivative.

(b) The random variables $g'(m(X))B(X)$, $g''(m(X))B^2(X)$, $g''(m(X))V(X)$ and $(g'(m(X)))^4$ have finite expectations.

Then, the MSE of $\bar{M}_n$ is

$$\mathrm{MSE}(\bar{M}_n) = \left(\mathrm{E}[g'(m(X))B(X)]h^2 + \frac{\mathrm{E}[g''(m(X))V(X)]}{2nh^d}\right)^2 + \frac{\mathrm{Var}[g(m(X))] + c_n}{n} + o\left(\left(h^2 + \frac{1}{nh^d}\right)^2 + \frac{1}{n}\right), \tag{19}$$

where c_n satisfies $0 \leqslant c_n \leqslant 2C_1(\sqrt{\mathrm{E}[(g'(m(X)))^4]\mathrm{E}[Z_n^4(X)]} + 1) < \infty$ when n is large enough.

3.2. Analysis for Hockey Stick and Indicator Functions

Theorem 3.1 establishes the asymptotic MSE of $\bar{M}_n$ when g is a smooth function. In this subsection, we consider an extension to the cases of hockey-stick functions and indicator functions.

We note that such an extension is not straightforward, because the presence of nondifferentiable or discontinuous points in g violates the critical conditions required by Theorem 3.1. To revolve these technical issues, more elaborate analysis and additional regularity conditions are required. In particular, we summarize in Theorems 3.2 and 3.3 the asymptotic MSEs of the kernel estimators corresponding to the hockey-stick function and the indicator function, respectively. Proofs of these theorems and lengthy discussions of the additional regularity conditions are provided in Section EC.1 of the e-companion.

Theorem 3.2. *Let $g(t) = t^+$ and $\bar{M}_n$ be the corresponding kernel estimator. Suppose that $m(X)$ has a continuous and bounded density in a neighborhood of 0, Assumption 1 is satisfied, the results in Lemma 2.1 hold for all $x \in \mathscr{D}$, and $\mathrm{E}(|B(X)|) < \infty$. Assume that $f(x)$ is bounded from above by a constant C_1 for all $x \in \mathscr{D}$, and there exists some $\delta > 0$ such that*

$$\sup_n \mathrm{E}[|\tilde{m}_n(X)|^{2+\delta}] < \infty.$$

Then, under some additional regularity conditions,[4] *the MSE of* $\bar{M}_n$ *is*

$$\begin{aligned}&\mathrm{MSE}(\bar{M}_n)\\&=\left(\mathrm{E}[B(X)\cdot 1_{\{m(X)\geqslant 0\}}]h^2+f_m(0)\mathrm{E}[V(X)\mid m(X)=0]\frac{1}{2nh^d}\right)^2\\&\quad+\frac{\mathrm{Var}[m(X)\cdot 1_{\{m(X)\geqslant 0\}}]+c_n}{n}+o\left(\left(h^2+\frac{1}{nh^d}\right)^2+\frac{1}{n}\right),\end{aligned}$$

where f_m *denotes the density of* $m(X)$*, and* c_n *satisfies* $0\leqslant c_n\leqslant C_1\mathrm{E}[Z_n^2(X)]<\infty$.

Theorem 3.2 shows that even in the presence of a nondifferentiable point, the asymptotic MSE is of the same order as that for a smooth function.

Theorem 3.3. *Let* $g(t)=1_{\{t\geqslant 0\}}$ *and* $\bar{M}_n$ *be the corresponding kernel estimator. Suppose that* $m(X)$ *has a continuous and bounded density in a neighborhood of* 0*, the results of Lemma 2.1 hold for all* $x\in\mathcal{D}$*. Assume that* $f(x)$ *is bounded from above by a constant* C_1 *for all* $x\in\mathcal{D}$*, and there exists some* $\delta>0$ *such that*

$$\sup_n \mathrm{E}[|\tilde{m}_n(X)|^{\delta}]<\infty.$$

Then, under the same additional regularity conditions as in Theorem 3.2, the MSE of $\bar{M}_n$ *is*

$$\begin{aligned}\mathrm{MSE}(\bar{M}_n)=&\left(f_m(0)\mathrm{E}[B(X)\mid m(X)=0]h^2\right.\\&\left.-\frac{1}{2}\int_{-\infty}^{\infty}y_1\partial_{y_2}p_2(y_1,0)\,dy_1\frac{1}{nh^d}\right)^2\\&+\frac{\mathrm{Var}[1_{\{m(X)\geqslant 0\}}]}{n}\\&+o\left(\left(h^2+\frac{1}{nh^d}\right)^2+\frac{1+nh^{d+2}}{n}\right),\end{aligned}$$

where p_2 *denotes the joint density of* $(V(X),m(X))$.

Theorem 3.3 establishes both the asymptotic bias and variance of the kernel estimator when g is an indicator function. The asymptotic bias has the same order as that for a smooth function. However, compared to the cases when g is a smooth or a hockey stick function, its asymptotic variance is slightly different. In this case, the contribution of the covariance term to the variance of the estimator is $o(1+nh^{d+2})$, while it is $O(1)$ when g is a smooth or a hockey-stick function.

If we let $h=O(n^{-1/(d+2)})$, it can be seen that the contribution of the covariance term is asymptotically smaller than $O(1)$. In other words, with appropriate selection of h, the covariance term does not inflate the asymptotic variance of the estimator.

3.3. Optimized Mean Squared Error

With Theorems 3.1–3.3, we can analyze the optimal rate of convergence of $\bar{M}_n$ for a given computational budget. In many practical problems, for instance, risk measurement of financial portfolios (Gordy and Juneja 2010, Liu and Staum 2010), computational effort is primarily spent in generating samples of (X,Y) (especially Y). We refer to this computational effort as sampling effort, and the corresponding computational budget as sampling budget. In this section, we ignore the effort required for computing the estimator given the samples of (X,Y), and mainly focus on the relationship between the rate of convergence and the sampling budget. The issue of computing the estimator shall be discussed in detail in Section 4.

For all three different types of functions g, we can see that the dominant term of the MSE of $\bar{M}_n$ is

$$\left(ah^2+\frac{b}{nh^d}\right)^2+\frac{c_n}{n},$$

where $c_l\leqslant c_n\leqslant c_u$; a, b, $c_l\geqslant 0$, and $c_u\geqslant 0$ are constants.

Without loss of generality, let the computational effort required to generate a sample of (X,Y) be 1. Then, we have $\Gamma=n$, where Γ denotes the total sampling budget. We let $\Gamma\to\infty$ to analyze the asymptotic rate of convergence of the estimator $\bar{M}_n$ with respect to the sampling budget. Then, the dominant term of the MSE of $\bar{M}_n$ can be written as

$$a^2h^4+\frac{2ab}{\Gamma h^{d-2}}+\frac{b^2}{\Gamma^2h^{2d}}+\frac{c_n}{\Gamma}.$$

It can be verified that (see Section EC.2.3 of the e-companion for a detailed derivation) the optimal rate of convergence of the MSE then equals

$$\Gamma^{-\min(1,4/(d+2))},$$

and the optimal h is of order $\Gamma^{-1/(d+2)}$.

An observation from the above analysis is that when the dimension of X is one or two, i.e., $d=1,2$, the optimal rate of convergence is actually Γ^{-1}, which is the same as that for a typical sample mean estimator of an ordinary expectation. However, when the dimension becomes higher, the rate of convergence of the kernel estimator becomes slower, in contrast to the nested simulation approach where the rate of convergence is insensitive to the dimension of X.

A major difference between the nested simulation approach and the kernel approach is on the bias, while their variances are both inversely proportional to the number of outer-level observations. The order of the asymptotic bias of the nested simulation approach depends solely on the number of samples used in the inner-level simulation, no matter how large the dimension d is. However, for the kernel approach, the bias

depends on both the bias and variance of the inner estimators. When d becomes larger, either the bias or the variance of the inner estimator becomes larger. Therefore the rate of convergence of the kernel approach deteriorates as d increases.

Compared to nested simulation estimators that converge at a rate of $\Gamma^{-2/3}$ (see Lee 1998, Gordy and Juneja 2010), the kernel estimators have the same or faster rates of convergence when $d \leqslant 4$, and slower rates of convergence when $d > 4$.

In a more general setting where different numbers of inter- and outer-samples are used in the kernel estimator, asymptotic result on rate of convergence follows in a similar manner. In particular, for the estimator $\bar{M}_{N,n,l}$ in (8), the dominant term of its asymptotic MSE is

$$\left(ah_l^2 + \frac{b}{nh_l^d}\right)^2 + \frac{v_1}{n} + \frac{v_2}{N},$$

for some constants a, b, v_1, and v_2. Given a sampling budget $\Gamma = (N + l)\gamma_1 + n\gamma_2$, where γ_1 and γ_2 denote the sampling costs for X and Y, respectively, it can be verified that the optimal rate of convergence of the MSE is again

$$\Gamma^{-\min(1,4/(d+2))}.$$

However, it should be noted that the optimal h_l is different from that in the previous setting.

4. Implementation Issues and Practical Algorithms

In the asymptotic analysis, we have focused mainly on computational budget for sampling, while ignoring the effort required to compute the NW kernel estimator. For some other approaches such as the nested simulation approach, the effort required to compute estimators may not be an issue and can be ignored. For instance, given n_1 outer-level samples, for each of which n_2 inner samples are generated, computing a nested estimator typically requires $O(n_1 n_2)$ operations, which can be done very quickly, and thus the computational effort is usually ignored. However, with the same sampling budget, one generates $n = n_1 n_2$ samples of (X, Y) in the kernel smoothing approach. Because computing $\tilde{m}_n(x)$ requires $O(n)$ operations for each x, it can be checked that directly calculating $\tilde{M}_n$ requires $O(n^2)$ operations in total. When sample size n is large, which is often the case in practice, computation of the NW kernel estimator may be a bottleneck in the implementation.

To address this computational issue, one direction is to propose fast algorithms for computing the NW kernel estimator. Indeed, when $d = 1$, fast algorithms have been proposed in the literature of computational statistics based on sorting the samples; see Seifert et al. (1994). For completeness, we briefly describe one of these algorithms with complexity of $O(n \log n)$ in Section E.C.2.6 of the e-companion. However, $O(n \log n)$ algorithms are not available for $d \geqslant 2$. When $d \geqslant 2$, by sorting the samples along one of the d dimensions and applying similar ideas as in Section EC.2.6, we can develop an algorithm that requires $O(n^2 h)$ operations, where h is the bandwidth. When the bandwidth h is optimally chosen as $O(n^{-1/(d+2)})$, the complexity of the algorithm is $O(n^{(2d+3)/(d+2)})$.

When d becomes larger, the complexity of the algorithm becomes closer to $O(n^2)$, which makes the implementation practically infeasible. We found that, on a personal computer, computing the NW estimator is hardly affordable when $d = 5$ and $n = 10^5$. To circumvent this difficulty, a practical solution may be developing an approximate algorithm that enables fast implementation. To this end, we consider a variant of kernel smoothing, the so-called k-nearest neighbor (kNN), which can be implemented more efficiently.

k-Nearest Neighbor Estimator

The kNN estimator can be viewed as a special case of the NW estimator with a variable bandwidth. Intuitively, when estimating $m(x) = \mathrm{E}[Y|X = x]$, both the NW estimator and the kNN estimator take the average of the samples of Y, the associated X values of which lie within a neighborhood of x. These two estimators are different in terms of choosing the "neighborhood." The NW estimator chooses a hypercube with side length of h and centered at x as the neighborhood, while the kNN estimator chooses the set of k nearest neighbors as the neighborhood where the term "nearest" is defined under certain distance metric, e.g., the Euclidean distance.

Suppose that the distance between two samples of X is measured by the Euclidean distance metric. Then, a kNN estimator of $m(x)$ is

$$\tilde{m}_n^{knn}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ki}(x) Y_i,$$

where the weight sequence $\{W_{ki}, 1 \leqslant i \leqslant n\}$ is defined by the set of indices

$$J_x = \{i\colon X_i \text{ is one of the } k\text{-nearest observations to } x\}$$

such that for $i = 1, \ldots, n$,

$$W_{ki} = \begin{cases} n/k & \text{if } i \in J_x \\ 0 & \text{otherwise.} \end{cases}$$

Then, a kNN estimator of the quantity of interest, α, is

$$\tilde{M}_n^{knn} = \frac{1}{n}\sum_{k=1}^{n} g(\tilde{m}_n^{knn}(X_k)).$$

One of the most attractive features of the kNN estimator is that it offers computational advantages. Typically, k is chosen to be an integer that is several orders

of magnitude smaller than n. Compared to the NW estimator that requires almost $O(n^2)$ operations when d is large, the kNN estimator can be computed with $O(n \log n)$ operations using fast search algorithms; see, e.g., Remark 3.3 of Belomestny et al. (2010). With the help of fast kNN search algorithms, the kNN estimator is computationally tractable for reasonably large sample sizes; see more discussion in Härdle (1990). Implementable packages of kNN search algorithms have been provided by several scientific computing software packages such as Matlab and R.

5. Decomposition Portfolio Risk Measurement

As an important application of nested estimation, portfolio risk measurement has received increasing attention recently. To illustrate how portfolio risk measurement fits into the framework of nested estimation, we first consider an option portfolio as an example in Section 5.1.

5.1. A Motivating Example

Consider a portfolio of options that are written on d different stocks and have the same maturity date T. Let $\{X_t^1, \dots, X_t^q\}$ denote price dynamics of the underlying stocks at time $t \geqslant 0$. For each stock X^i, $i = 1, \dots, q$, the portfolio includes a vanilla European call option with a terminal payoff $(X_T^i - K^i)^+$, and an exotic option (in particular, a barrier option) with a terminal payoff $(X_T^i - S^i)^+ 1_{\{\max_{0\leqslant t\leqslant T} X_t^i \leqslant U^i\}}$, where K^i, S^i, and U^i are constants specified in the option contracts.

A risk manager is interested in the probability that the portfolio loss at a future time τ is larger than or equal to a given threshold y_0, where $\tau < T$. Note that the portfolio loss at time τ can be written as $L = V_0 - V_\tau$, where V_0 and V_τ denote the portfolio values at time 0 and time τ, respectively, and V_0 is a known constant.

Suppose that stock prices evolve according to the Black-Scholes model, i.e.,

$$X_t^i = X_0^i \exp((\mu_i - \sigma_i^2/2)t + \sigma_i B_t^i), \quad i = 1, \dots, q,$$

where μ_i is set as the rate of return of the ith stock if $t \leqslant \tau$ and a constant risk-free rate r if $t > \tau$,[5] and $\{B_t^i, \dots, B_t^q\}$ are correlated standard Brownian motions. Let $\mathcal{F}_\tau^i$ denote the filtration generated by $\{X_t^i, 0 \leqslant t \leqslant \tau\}$. By the derivative pricing theory (Duffie 1996), V_τ is given by

$$\begin{aligned} V_\tau &= e^{-r\Delta} \sum_{i=1}^{d} (\mathrm{E}[(X_T^i - K^i)^+ | \mathcal{F}_\tau^i] \\ &\quad + \mathrm{E}[(X_T^i - S^i)^+ 1_{\{\max_{0\leqslant t\leqslant T} X_t^i \leqslant U^i\}} | \mathcal{F}_\tau^i]) \\ &= e^{-r\Delta} \sum_{i=1}^{d} (\mathrm{E}[(X_T^i - K^i)^+ | X_\tau^i] \\ &\quad + \mathrm{E}[(X_T^i - S^i)^+ 1_{\{\max_{0\leqslant t\leqslant T} X_t^i \leqslant U^i\}} | X_\tau^i, \max_{0\leqslant t\leqslant \tau} X_t^i]), \end{aligned} \tag{20}$$

where $\Delta = T - \tau$, and the second equality follows from the Markovian property of $\{X_t^i, t \geqslant 0\}$.

We define a vector of risk factors by

$$X = \left(X_\tau^1, \dots, X_\tau^q, \max_{0\leqslant t\leqslant \tau} X_t^1, \dots, \max_{0\leqslant t\leqslant \tau} X_t^q\right),$$

and a random variable

$$Y = V_0 - e^{-r\Delta} \sum_{i=1}^{q} [(X_T^i - K)^+ - (X_T^i - K)^+ 1_{\{\max_{0\leqslant t\leqslant T} X_t^i \leqslant U\}}].$$

Then, the portfolio loss $L = V_0 - V_\tau$ can be written as

$$L = \mathrm{E}[Y | X],$$

and the quantity of interest is

$$\Pr(L \geqslant y_0) = \mathrm{E}[g(\mathrm{E}(Y | X))],$$

where $g(x) = 1_{\{x \geqslant y_0\}}$. The kernel smoothing estimator proposed earlier in this paper can then be applied to estimate $\Pr(L \geqslant y_0)$.

In this example, the dimension of the risk factors X is $d = 2q$. For large-scale portfolios held by investors, in practice, q could be very large, e.g., $q = 100$ or above. Based on the asymptotic results derived in Section 3, a direct application of the kernel smoothing approach to these cases is not practical, because the dimensions are much higher than the range where the kernel estimators are competitive.

We now examine the example in greater detail. By (20), we note that the quantity of interest can also be written as

$$\begin{aligned} \Pr(L \geqslant y_0) = \mathrm{E}\Bigg[g\Bigg(V_0 &- \sum_{i=1}^{d} \mathrm{E}(Y_{1,i} | X_\tau^i) \\ &- \sum_{k=1}^{d} \mathrm{E}(Y_{2,k} | X_\tau^k, \max_{0\leqslant t\leqslant T} X_t^k) \Bigg) \Bigg], \end{aligned} \tag{21}$$

where $Y_{1,i} = e^{-r\Delta}(X_T^i - K^i)^+$ and $Y_{2,k} = e^{-r\Delta}(X_T^k - S^k)^+ \cdot 1_{\{\max_{0\leqslant t\leqslant T} X_t^k \leqslant U^k\}}$.

Then, the kernel smoothing approach can be applied to estimate *each* of the conditional expectations on the RHS of (21). Note that the dimensions of the risk factors associated with these conditional expectations are low, either one for the case of $\mathrm{E}(Y_{1,i} | X_\tau^i)$ or two for the case of $\mathrm{E}(Y_{2,k} | X_\tau^k, \max_{0\leqslant t\leqslant T} X_t^k)$. Therefore the kernel smoothing approach is expected to perform well based on the asymptotic analysis in previous sections.

In other words, while the dimension of all risk factors associated with the entire portfolio is typically high, the dimensions of the risk factors associated with individual options are often low. This observation motivates us to apply the kernel smoothing approach to individual options. By doing so, the effective dimension of the problem is reduced. In the following subsection, we describe this decomposition technique for more general settings.

5.2. Decomposition

A risk manager is often interested in measuring the downside risk of a portfolio, which can be represented by $\mathrm{E}[g(\mathrm{E}[Y|X])]$ in many cases, where Y is the random portfolio loss at a maturity date, g depends on the risk measure being used, and $X = (X^1, \ldots, X^d)$ is a d-dimensional random vector representing risk factors at a future time horizon up to which we measure the risk. The problem then fits into the framework of nested estimation naturally.

While the total number of risk factors associated with the entire portfolio is usually large, the value of each financial instrument in the portfolio often depends on only a small number of risk factors. For instance, statistics of the World Federation of Exchanges (2014) show that more than 90% of options traded on exchanges worldwide in 2013 are written on a single underlying asset that may be a stock, a stock index, an ETF, an exchange rate, or a commodity. Typically, the loss of a portfolio is a linear combination of losses of individual financial instruments. Therefore we may divide these instruments into groups, each depending on only a small number of common risk factors. In particular, we may rewrite $\mathrm{E}[Y|X]$ as

$$\mathrm{E}[Y|X] = \sum_{i=1}^{l} \mathrm{E}[Y^i|X] = \sum_{i=1}^{l} \mathrm{E}[Y^i|X^j,\ j \in \mathscr{B}_i], \qquad (22)$$

where l is the number of groups, $\mathscr{B}_i$ is a subset of $\{1, \ldots, d\}$, Y^i denotes the total loss of the financial instruments in group i, which depends on only the risk factors $\{X^j, j \in \mathscr{B}_i\}$, and Y^i satisfies $\sum_{i=1}^{l} Y^i = Y$. It should be pointed out that $\mathscr{B}_i$'s are not necessarily disjoint. For instance, the first group may depend on (X^1, X^2, X^3), while the second may depend on (X^3, X^4, X^5). Here, we want to emphasize that the decomposition of risk factors in (22) is possible not because of specific models postulated for X, but because of the unique structure of financial portfolios, and this structure is independent of models of X.

Based on the representation in (22), we apply the kernel smoothing approach to estimate $\mathrm{E}[Y^i|X^j,\ j \in \mathscr{B}_i]$ for $i = 1, \ldots, l$. Denoted by $\tilde{m}_{ni}(x^j,\ j \in \mathscr{B}_i)$, the kernel estimator of $\mathrm{E}[Y^i|X^j,\ j \in \mathscr{B}_i]$ evaluated at $X^j = x^j$, $j \in \mathscr{B}_i$. As recommended in Section 4, $\tilde{m}_{ni}$ is chosen to be the NW estimator when the number of elements in $\mathscr{B}_i$ is equal to 1, while it is chosen to be the kNN estimator otherwise. Then, an estimator of $\alpha = \mathrm{E}[g(\mathrm{E}[Y|X])]$ is

$$\tilde{M}_n^r = \frac{1}{n} \sum_{k=1}^{n} g\left(\sum_{i=1}^{l} \tilde{m}_{ni}(X_k^j,\ j \in \mathscr{B}_i) \right),$$

where $\{X_k^j, j \in \mathscr{B}_i\}$ denotes the kth sample of $\{X^j, j \in \mathscr{B}_i\}$.

Define $d_e = \max\{|\mathscr{B}_1|, \ldots, |\mathscr{B}_l|\}$, where $|\mathscr{B}_i|$ denotes the number of elements in $\mathscr{B}_i$. We call d_e the effective dimension. For many equity portfolios in practice, effective dimension d_e is often much smaller than the actual dimension d.

In summary, by decomposing a high-dimensional problem into a number of low-dimensional ones, the kernel smoothing approach may be a viable tool for portfolio risk measurement problems. The decomposition allows for general dependence structures of risk factors and works for portfolios that are comprised of a wide range of financial instruments; see Remark 2. Furthermore, the decomposition idea may apply to other smoothing approaches as well. For instance, when using stochastic kriging, instead of building a metamodel for a high-dimensional response surface, one may build metamodels for a number of low-dimensional surfaces. Whether such a decomposition offers benefits to stochastic kriging is yet to be confirmed, and opens up a direction for future research.

Remark 2. As shown in Section 5.1, the decomposition technique applies to portfolios with both vanilla and exotic options. In general, if the payoff of a financial instrument in the portfolio is path-dependent, augmented risk factors may be appropriately defined to enable the decomposition. For instance, consider a portfolio that includes an Asian call option written on an asset X_t within the Black-Scholes model. Let $((1/T)\int_0^T X_t\,dt - \kappa)^+$ be the option payoff at maturity with κ being the strike price. Then, at time τ, the option value depends on $X^{\mathrm{aug}} \triangleq \int_0^\tau X_t\,dt$ and X_τ. By adding X^{aug} as an augmented risk factor, the decomposition proceeds in the same manner as the vanilla case.

The decomposition also adapts to complex pricing models, such as SV models that are popular, in practice; see, e.g., Fouque et al. (2000) for details. When SV models are used, closed-form pricing formulas may not exist and simulation is often a preferred pricing tool. In such cases, the volatility (or variance) processes may be added as augmented risk factors, and the decomposition then proceeds with the augmented set of risk factors.

Remark 3. Although not common, in practice, it is possible that a portfolio may include high-dimensional instruments such as European swaptions with dimensions larger than 20. In such cases, one may use nested simulation in estimating the values of high-dimensional instruments, while using the kernel smoothing approach for other low-dimensional ones. This may perform better than a pure nested simulation procedure or a pure kernel smoothing approach. A thorough understanding of this hybrid method is of practical interest and deserves further investigation.

6. Numerical Experiments

We consider two examples. The first one is designed to examine the impact of dimensionality on performances

of different estimators, while the second one is a reasonably large and representative example with 200 risk factors, aiming to demonstrate the practical value of the proposed kernel smoothing approach in conjunction with the decomposition technique.

6.1. Impact of Dimensionality

Consider a portfolio that consists of options written on d stocks whose price dynamics follow the Black-Scholes model. In the portfolio, there are three call options written on each stock with different strike prices. For simplicity, assume that stock returns are the same, denoted by μ, while risk-free interest rate is r. Price dynamics of the stocks $\mathbf{X}_t = (X_t^1, \ldots, X_t^q)$ evolve according to

$$dX_t^i = \mu' X_t^i \, dt + \sum_{j=1}^{d} \sigma_{ij} X_t^i \, dB_t^i, \quad i = 1, \ldots, q, \tag{23}$$

where μ' is chosen to be μ under the real-world probability measure, while it is chosen to be r under the risk-neutral probability measure. Here, $\mathbf{B}_t = (B_t^1, \ldots, B_t^q)$ is a standard d-dimensional Brownian motion, and without loss of generality, we let $\Sigma = (\sigma_{ij})$ be a subtriangular matrix. Then,

$$X_t^i = X_0^i \exp\left(\left(\mu' - \frac{1}{2}\sum_{j=1}^{i} \sigma_{ij}^2\right)t + \sum_{j=1}^{i} \sigma_{ij} B_t^j\right), \quad i = 1, \ldots, q.$$

We assume that the maturities of all the options in the portfolio are the same, denoted by T. We want to measure the portfolio risk at a future time τ ($\tau < T$). In the simulation, we first simulate $\mathbf{X}_\tau$ under the real-world probability measure and then simulate $\mathbf{X}_T$ under the risk-neutral probability measure. Note that the (discounted) payoff of the portfolio at time T is a known function of $\mathbf{X}_T$, denoted by $V_T(\mathbf{X}_T)$. We let a constant V_0 denote the value of the portfolio at time 0, which can be calculated by the Black-Scholes formula. At time τ, the portfolio loss given $\mathbf{X}_\tau$ is

$$L(\mathbf{X}_\tau) = \mathrm{E}[V_0 - V_T(\mathbf{X}_T) | \mathbf{X}_\tau].$$

Note that the dimension of the problem is $d = q$ for this example. We want to measure the portfolio risk that is represented by

$$\alpha = \mathrm{E}[g(L(\mathbf{X}_\tau))],$$

where we consider three cases: a quadratic function $g(t) = t^2$, a hockey stick function $g(t) = (t - y_0) \cdot 1_{\{t > y_0\}}$, and an indicator function $g(t) = 1_{\{t > y_0\}}$, respectively. Here, y_0 is a prespecified threshold. Parameters of the model are set as follows: $X_0^1 = \cdots = X_0^q = 100$, $T = 1$, $\tau = 1/50$, $\mu = 8\%$, $r = 5\%$, and $y_0 = 20\% V_0$. Strike prices of the three call options are $K = 90$, 100, 110, respectively, and the elements of the volatility matrix Σ are randomly selected.

The kernel smoothing approach can be applied to estimate α. To measure its performance, we need the true value of α as a benchmark. Note that in this example, the analytical expression of $L(\mathbf{X}_\tau)$ can be derived using the Black-Scholes formula, and thus we can generate a large amount (10^9) of samples of $\mathbf{X}_\tau$ and use sample mean of $g(L(\mathbf{X}_\tau))$ to accurately approximate α. We then use this accurate estimate as a benchmark to measure the performance of an estimators by its relative root mean squared error (RRMSE), defined as the percentage of the root MSE to the benchmark. All RRMSEs reported are estimated based on 1,000 independent replications.

We compare the kernel smoothing approach to the nested simulation approach of Gordy and Juneja (2010). For both approaches, the sampling budget is fixed by letting the total number of inner-level samples be 10^6. Note that the performance of the nested simulation approach depends on the number of inner-level samples per each outer-level sample. We try different numbers of inner-level samples and choose the one that yields the best performance for the nested simulation approach. We then compare this "best-possible" performance to the kernel smoothing approach. It should be emphasized that for practical problems, the optimal number of inner-level sample per each outer-level sample is unknown, and thus the best-possible performance of the nested simulation estimator may be difficult to achieve.

When implementing the NW and kNN estimators for $d \geqslant 2$, we fix the sampling budget as 10^6 inner-level samples, and set the number of outer-level samples to be $l = 10^5$ as in (8), each with 10 inner-level samples. Here, 10^5, rather than 10^6, outer-level samples are preferred mainly due to the computational advantages it offers, because computational burden is affected mainly by the number of outer-level samples, and thus fewer outer-level samples may accelerate the implementation considerably. When $d = 1$, we set $n_1 = 10^6$ and $n_2 = 1$, because a very fast algorithm as in Section EC.2.6 of the e-companion.

Numerical results for the NW and kNN estimators are presented in Sections 6.1.1 and 6.1.2, respectively. Section 6.1.3 reports the numerical results for the kernel smoothing approach when the decomposition technique in Section 5 is applied.

6.1.1. Nadaraya-Watson Estimator. When using the NW estimator, we apply the well-known leave-one-out cross-validation procedure to select a good bandwidth among several prespecified candidates; see, e.g., Li and Racine (2007). Numerical results confirm that the MSEs of the NW estimators decay in rates that are consistent with the theoretical results proved in Section 3 for $d = 1$, 2, and 4. Details are provided in Section EC.2.7 of the e-companion.

Table 1. Comparison of RRMSE (%) for the NW Estimator and Nested Simulation Estimator

	NW estimator	Nested
$d = 1$		
Quadratic	2.8	7.3
Hockey-stick	6.2	17.4
Indicator	6.0	12.7
$d = 2$		
Quadratic	3.3	7.5
Hockey-stick	7.9	16.3
Indicator	5.7	11.4
$d = 4$		
Quadratic	2.4	7.3
Hockey-stick	5.7	20.9
Indicator	5.9	13.9

Comparison results for the NW estimators and the nested simulation estimators are summarized in Table 1 for $d = 1$, 2, and 4. From the table, it can be seen that the NW estimator outperforms the nested simulation estimator, which coincides with the theoretical results that the MSE of the NW estimator may have a faster rate of convergence when d is relatively low. For instance, when $d = 4$ and g is a hockey-stick function, the RRMSE ratio of the nested estimator to the NW estimator is 3.6. It implies that to achieve the same level of accuracy, sampling budget of the nested simulation approach has to be larger than $3.6^3 \approx 46$ times of that of the NW estimator, because RRMSEs of the nested simulation estimator converge to zero in a rate of $n^{-1/3}$.

While numerical results indicate that the NW estimator may have better accuracy than the nested simulation estimator when $d \leqslant 4$, we found it difficult to compare their performances for larger d, because the computational burden of the NW estimator increases in d, and it is hardly affordable to run 1,000 replications when $d \geqslant 5$. The computational time of a single replication increases from 28 minutes to more than 1 hour when the dimension increases from 4 to 5 using Matlab running on a PC with 2.4 GHz Intel Xeon CPU.

6.1.2. k-Nearest Neighbor Estimator. Comparison results for the kNN estimator and the nested simulation estimator are summarized in Table 2, where we vary the smoothing parameter k to illustrate its impact on the performance of the kNN estimator. From the table, we can see that the kNN estimator may outperform the nested simulation estimator for dimensions as large as 20. For instance, when g is a quadratic function, $d = 10$ and $k = 100$, the RRMSE ratio of the nested simulation estimator to the kernel estimator is 4. Taking into account that the convergence rate of the RRMSE of the nested simulation approach is $n^{-1/3}$, the sample size of the nested simulation approach has to be as large as $4^3 = 64$ times of that of the kNN estimator to achieve the same level of accuracy.

Table 2. Comparison of RRMSE (%) for the kNN Estimator and Nested Simulation Estimator

		kNN				
	k	50	100	200	300	Nested
$d = 2$	Quadratic	14.1	6.7	3.0	2.0	7.5
	Hockey-stick	27.7	13.7	7.3	5.6	16.3
	Indicator	15.9	8.8	6.3	5.5	11.4
$d = 4$	Quadratic	11.3	4.2	1.9	3.3	7.3
	Hockey-stick	27.0	11.1	6.2	8.0	20.9
	Indicator	17.1	8.5	5.4	5.2	13.9
$d = 8$	Quadratic	8.6	2.0	4.1	6.2	7.2
	Hockey-stick	22.4	7.3	9.1	12.9	21.2
	Indicator	15.3	6.5	5.3	6.4	14.6
$d = 10$	Quadratic	7.3	1.8	5.2	7.2	7.2
	Hockey-stick	20.3	6.6	11.5	15.8	22.1
	Indicator	14.4	5.8	6.1	8.0	15.3
$d = 20$	Quadratic	5.7	2.7	7.0	9.1	7.0
	Hockey-stick	18.6	8.3	18.0	23.7	25.3
	Indicator	15.0	5.8	9.1	12.0	17.8

From Table 2, it can also be seen that while the performance of the kNN estimator is affected by the smoothing parameter k, it is reasonably good for a number of k's ranging from 100 to 300. To select a reasonably good k, we use the leave-one-out cross-validation procedure. However, it should be noted that such cross-validation procedures lead to the same smoothing parameter for different functions g, while from numerical results, we note that different k's may be more appropriate for different functions g. How to take into account the functions g in the selection of smoothing parameters may deserve further investigation and is left as a topic for future research.

We note that computing the kNN estimator for large d may require longer time than the nested simulation estimator (roughly a few minutes versus a few seconds using Matlab running on a PC with 2.4 GHz Intel Xeon CPU, 24 GB RAM). However, this is typically not an issue, in practice, where sampling effort is typically of major concern and dominates the effort required for computing estimators, especially when the maturities of financial instruments are long and risk factors follow complex pricing models.

6.1.3. Effectiveness of Decomposition. For this example, the problem of interest can be decomposed into several one-dimensional ones after applying the decomposition technique in Section 5. We compare the k-nearest neighbor (kNN) estimator in conjunction with decomposition to the nested simulation estimator, and the recent regression approach in Broadie et al. (2015). To set basis functions for the regression method, we use the popular weighted Laguerre polynomials (see, e.g., Longstaff and Schwartz 2001) on individual risk factors up to an order of 4. Numerical results are summarized in Table 3, where the three columns for each approach represent the relative errors

Table 3. Comparison of RRMSE (%) for the Kernel Smoothing Approach, the Nested Simulation Approach, and the Regression Approach

	Kernel			Nested			Regression		
$d=2$	5.0	5.4	3.2	7.5	16.3	11.4	1.9	4.5	2.8
$d=4$	2.9	5.4	2.4	7.3	20.9	13.9	1.8	5.3	3.3
$d=8$	1.8	4.1	2.2	7.2	21.2	14.6	1.7	5.4	3.4
$d=10$	1.6	3.9	2.3	7.2	22.1	15.3	1.7	5.5	3.5
$d=20$	1.2	2.7	2.5	7.0	25.3	17.8	1.8	7.0	4.4

Note. For the kernel smoothing approach, the NW estimator is applied with decomposition.

for quadratic, hockey-stick and indicator functions, respectively. From the table, we can see that the kernel smoothing approach in conjunction with decomposition works very well. Its relative errors range roughly from 1% to6% for $d = 2, 4, 8, 10, 20$. It outperforms the nested simulation approach significantly in terms of accuracy, while having comparable computational speed.

Compared to the regression approach, the kernel smoothing approach has slightly higher relative errors when the dimension is low, e.g., $d = 2, 4$. It turns out, however, the kernel smoothing approach outperforms the regression approach when the dimension is higher, e.g., $d = 20$, especially when g is a hockey stick function.

6.2. A Representative Example

Consider a portfolio that is comprised of derivative contracts written on $q = 100$ assets. We divide these 100 underlying assets into four groups with $q' = 25$ assets in each group. Suppose that underlying assets from different groups are independent.

We assume that the assets in Groups 1–3 follow multidimensional geometric Brownian motions (GBMs) in the form of (23), while the assets in Group 4 follow a multidimensional Heston's SV model as in Dimitroff et al. (2011). In particular, under this SV model for q' assets, the price of the ith asset is governed by

$$\begin{aligned} dX_i(t) &= \mu_i X_i(t)\,dt + \sqrt{v_i(t)}X_i(t)\,dB_i(t),\\ dv_i(t) &= \kappa_i(\theta_i - v_i(t))\,dt + \eta_i\sqrt{v_i(t)}\\ &\qquad \cdot(\rho_i\,dB_i(t) + \sqrt{1-\rho_i^2}\,d\tilde{B}_i(t)), \quad i = 1,\ldots,q', \end{aligned}$$

where $v_i(t)$ describes a mean-reverting variance process with reversion rate κ_i, mean-level θ_i, volatility η_i, and initial value $v_i(0)$, $B(t) = (B_1(t),\ldots,B_{q'}(t))$, and $\tilde{B}(t) = (\tilde{B}_1(t),\ldots,\tilde{B}_{q'}(t))$ are q'-dimensional Brownian motions, and have correlation matrices $\Sigma = (\rho_{ij})_{1\leqslant i,j\leqslant q'}$ and $I_{q'}$, respectively, where $I_{q'}$ is an identity matrix with size q'. In this model, $B(t)$ and $\tilde{B}(t)$ are independent, and ρ_i is used to capture the correlation between the price process and the variance process.

The portfolio is comprised of different types of derivative contracts. Details are summarized as follows.

- Group 1 (GBM Model) and Group 4 (SV Model): Written on each asset, there are European vanilla call options.
- Group 2 (GBM Model): Written on each asset, there are European vanilla call options and geometric Asian options with payoffs $((\prod_{k=1}^{m} X(t_k))^{1/m} - K)^+$, where K is the strike price.
- Group 3 (GBM Model): Written on each asset, there are up-and-out call options with payoffs $(X(T)-K)^+1_{\{\max_{0\leqslant t\leqslant T} X(t)\leqslant U\}}$ and down-and-out call options with payoffs $(X(T)-K)^+1_{\{\min_{0\leqslant t\leqslant T} X(t)\geqslant H\}}$.

Because closed-form pricing formulas for these derivative contracts are available, we can obtain an accurate estimate of the portfolio risk, and use this estimate as a benchmark to examine the performances of different estimators.

This portfolio involves $d = 2q = 200$ risk factors in total, including underlying asset prices, stochastic volatilities, running geometric averages, running maxima and minima up to the risk horizon. Our objective is to measure the portfolio risk up to a time horizon $\tau = 3/50$, where the maturities of all derivative contracts are set to be 1. Parameters of the model are specified as follows:

- For all assets under either GBM or SV models, returns of the assets are set to be 8%, while risk-free rate is 5%. Initial asset prices are set to be $X_i(0) = 100$. Volatilities for assets in Groups 1–3 are set to be 15%, 40%, and 30%, respectively.
- For SV model, $\rho_i = -0.3$, $\kappa = 3$, $\eta_i = 0.5$, $\theta_i = 0.09$, and $v_i(0) = 0.09$ for all i. Correlation between $B_i(t)$ and $B_j(t)$ is 0.3 if $i \neq j$.
- For each type of derivatives, we consider three options with different strikes $K = 90, 100, 110$. For Asian options, $m = 50$ and t_k's are evenly spaced in $[0, T]$.

During the implementation, the SV model is simulated with the Euler scheme with 200 time-steps; see Dimitroff et al. (2011). When simulating the continuously monitoring maximum and minimum for barrier options, 200 time-steps are used and Brownian bridge approximation is applied for any two adjacent time-points; see Glasserman (2004, pp. 367–368) for details of Brownian bridge approximations.

We compare the kernel smoothing estimator to the nested simulation estimator with different settings, and the regression estimator as in Broadie et al. (2015) with weighted Laguerre polynomials up to an order of 4 as basis functions. The decomposition technique and the leave-one-out cross-validation procedure for selecting smoothing parameters are applied when using the kernel smoothing approach. From the comparison results in Table 4, we find that the kernel smoothing estimator (kNN) significantly outperforms

Table 4. Comparison of RRMSE (%) for the Kernel Smoothing Estimator, Nested Simulation Estimators, and the Regression Estimator

Nested simulation					kNN	Regression
$n = 10^3$	10×100	20×50	40×25	50×20		
Quadratic	53.6	47.2	60.8	68.4	9.2	266
Hockey-stick	143.3	126.5	141.5	162.5	27.3	572
Indicator	99.9	83.6	74.8	81.4	22.4	155
$n = 10^4$	50×200	100×100	200×50	400×25		
Quadratic	24.4	21.6	27.6	48.6	2.4	27.2
Hockey-stick	63.0	53.4	60.1	107.1	9.1	58.2
Indicator	46.6	34.5	34.9	54.0	7.2	24.4
$n = 10^5$	200×500	400×250	$1{,}000 \times 100$	$2{,}000 \times 50$		
Quadratic	12.1	9.2	12.9	24.0	1.0	3.2
Hockey-stick	30.3	23.7	28.3	50.2	2.9	6.0
Indicator	21.5	16.7	17.1	28.1	2.3	3.1

Notes. For hockey-stick and indicator functions, y_0 is set to be the 90% quantile of the portfolio loss distribution. The setting $k \times j$ for nested simulation estimators means that k outer samples with inner sample size being j.

the nested simulation estimators. For instance, when g is a quadratic function and sampling budget is $n = 10^5$, the smallest relative error achieved by the nested simulation estimator is 8.8%, while relative error of the kNN estimator is 1%. Given that relative error of nested simulation estimators converges to zero in a rate of $n^{-1/3}$, sampling budget for the nested simulation estimator needs to be as large as almost $700 (\approx 8.8^3)$ times of the kernel smoothing estimator to achieve the same level of accuracy.

We also find that the kernel smoothing approach may outperform the regression approach significantly, especially when the sampling budget is low. An observation that is worth mentioning is that RRMSE of the regression estimator decreases in a faster rate when sampling budget increases. Despite its slower rate of convergence, the kernel estimator has smaller RRMSEs for all cases with sampling budgets as large as 10^5, compared to the regression estimator.

7. Conclusions

We have studied a simple local smoothing approach for nested estimation, the kernel smoothing approach, and analyzed its asymptotic properties. While it is not surprising that the kernel smoothing approach is preferable only for low-dimensional problems, we have shown that it may serve as a viable tool for portfolio risk measurement where the dimension of the risk factors is usually very high. The key to this successful application of the kernel smoothing approach is an observation that a high-dimensional portfolio risk measurement problem can often be decomposed into a sequence of low-dimensional ones for which the kernel smoothing approach works very well, both theoretically and practically.

We have demonstrated the efficiency of the kernel smoothing approach in conjunction with the decomposition technique. For a portfolio risk measurement example with 200 risk factors, the kernel smoothing approach outperforms existing approaches in the literature, suggesting that the kernel smoothing approach may be a promising addition to the arsenal of portfolio risk measurement for practical problems.

Acknowledgments

The authors are grateful to an anonymous associate editor and three anonymous referees for their insightful comments and valuable suggestions that significantly improve the paper. Part of the results in this paper appeared in their preliminary form in Hong and Juneja (2009), a proceedings paper of the Winter Simulation Conference.

Appendix

A.1. Proof of Lemma 3.1

Since $\tilde{m}_n(X)$ converges to $m(X)$ in probability, by the continuous mapping theorem (Durrett 2005), $g(\tilde{m}_n(X))$, and $g^2(\tilde{m}_n(X))$ converge in probability to $g(m(X))$ and $g^2(m(X))$, respectively.

Note that $g^2(\tilde{m}_n(X)) \leqslant C^2|\tilde{m}_n(X)|^{2p}$. Then, $g^2(\tilde{m}_n(X))$ is uniformly integrable, which is implied by the assumption that $\sup_n \mathrm{E}[|\tilde{m}_n(X)|^{2p+\delta}]$ for some $\delta > 0$. Thus $g(\tilde{m}_n(X))$ is also uniformly integrable. Because convergence in probability combined with uniform integrability implies convergence in mean, we have

$$\mathrm{E}[g^2(\tilde{m}_n(X))] = \mathrm{E}[g^2(m(X))] + o(1) \quad \text{and}$$
$$\mathrm{E}[g(\tilde{m}_n(X))] = \mathrm{E}[g(m(X))] + o(1).$$

Then, the conclusion of Lemma 3.1 follows from:

$$\begin{aligned}\mathrm{Var}[g(\tilde{m}_n(X))] &= \mathrm{E}[g^2(\tilde{m}_n(X)) - g^2(m(X))] + \mathrm{Var}[g(m(X))] \\ &\quad - \{\mathrm{E}^2[g(\tilde{m}_n(X))] - \mathrm{E}^2[g(m(X))]\} \\ &= \mathrm{Var}[g(m(X))] + o(1).\end{aligned}$$

A.2. Proofs of Equations

A.2.1. Asymptotic Order of the Third Term on the RHS of Equation (10).

We first note that

$$\left|(\sqrt{nh^d})^3 \frac{g'''(\Xi)}{6}(\tilde{m}_n(X) - m(X))^3\right|$$

$$\leqslant \frac{C}{6}|[\sqrt{nh^d}(\tilde{m}_n(X)-m(X))]^3| = \frac{C}{6}|Z_n^3(X)|$$

if $g'''(t)$ is bounded by a constant C for all t. Note that $\mathrm{E}[|Z_n(X)|^3] < \infty$ by Assumption 1. Then,

$$\mathrm{E}\left[\frac{g'''(\Xi)}{6}(\tilde{m}_n(X)-m(X))^3\right] = O\left(\frac{1}{(\sqrt{nh^d})^3}\right).$$

In other words, the third term on the RHS of Equation (10) is of order $(nh^d)^{-3/2}$.

A.2.2. Proof of Equation (15).

Let $\mathcal{F} = \sigma\{(X_i, Y_i), 1 \leqslant i \leqslant n\}$ be the σ-algebra generated by $\{(X_i, Y_i), 1 \leqslant i \leqslant n\}$. Note that the dependence between $g(\tilde{m}_n(X_{0,1}))$ and $g(\tilde{m}_n(X_{0,2}))$ comes from $\tilde{m}_n$, and $\{X_{0,i}, 1 \leqslant i \leqslant n\}$ are independent. Then, $g(\tilde{m}_n(X_{0,1}))$ and $g(\tilde{m}_n(X_{0,2}))$ are independent conditional on $\mathcal{F}$. Therefore

$$\begin{aligned}
&\mathrm{Cov}[g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))]\\
&= \mathrm{E}(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))g(\tilde{m}_n(X_{0,2}))|\mathcal{F}]) - \mathrm{E}^2(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}])\\
&= \mathrm{E}(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}]\mathrm{E}[g(\tilde{m}_n(X_{0,2}))|\mathcal{F}]) - \mathrm{E}^2(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}])\\
&= \mathrm{E}[(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}])^2] - \mathrm{E}^2(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}])\\
&= \mathrm{Var}(\mathrm{E}[g(\tilde{m}_n(X_{0,1}))|\mathcal{F}]) \geqslant 0.
\end{aligned}$$

A.2.3. Proof of Equation (16).

Let $H(x_1)$ and $H(x_2)$ denote hypercubes with side length h and centered at x_1 and x_2, respectively.

$$\begin{aligned}
&n\mathrm{Cov}[g(\tilde{m}_n(X_{0,1})), g(\tilde{m}_n(X_{0,2}))]\\
&= n\left\{\int_{\mathscr{D}}\int_{\mathscr{D}} \mathrm{Cov}(g(\tilde{m}_n(x_1)), g(\tilde{m}_n(x_2)))f(x_1)f(x_2)\right.\\
&\qquad \left.\cdot 1_{\{x_2\in H(x_1)\}}\,dx_1\,dx_2\right\}\\
&\leqslant \frac{n}{2}\left\{\int_{\mathscr{D}}\int_{\mathscr{D}} \left(\mathrm{Var}[g(\tilde{m}_n(x_1))] + \mathrm{Var}[g(\tilde{m}_n(x_2))]\right) f(x_1)f(x_2)\right.\\
&\qquad \left.\cdot 1_{\{x_2\in H(x_1)\}}\,dx_1\,dx_2\right\}\\
&= n\int_{\mathscr{D}}\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x_1))]f(x_1)f(x_2)\cdot 1_{\{x_2\in H(x_1)\}}\,dx_1\,dx_2, \qquad (A.1)\\
&= n\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x_1))]f(x_1)\int_{\mathscr{D}} f(x_2)1_{\{x_2\in H(x_1)\}}\,dx_2\,dx_1, \qquad (A.2)
\end{aligned}$$

where Equation (A.1) follows from symmetry, i.e., $1_{\{x_2\in H(x_1)\}} = 1_{\{x_1\in H(x_2)\}}$, and

$$\begin{aligned}
&\int_{\mathscr{D}}\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x_1))]f(x_1)f(x_2)1_{\{x_2\in H(x_1)\}}\,dx_1\,dx_2\\
&\quad = \int_{\mathscr{D}}\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x_2))]f(x_1)f(x_2)1_{\{x_1\in H(x_2)\}}\,dx_1\,dx_2.
\end{aligned}$$

Because $f(x) \leqslant C_1$ for all $x \in \mathscr{D}$, we have $\int_{\mathscr{D}} f(x_2)1_{\{x_2\in H(x_1)\}}\,dx_2 \leqslant C_1\int_{\mathscr{D}} 1_{\{x_2\in H(x_1)\}}\,dx_2 = C_1h^d$. Then, Equation (16) follows immediately from (A.2).

A.2.4. Proof of Equation (17).

By Taylor expansion,

$$\begin{aligned}
&nh^d\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x))]f(x)\,dx\\
&\leqslant nh^d\int_{\mathscr{D}} \mathrm{E}[(g(\tilde{m}_n(x)) - g(m(x)))^2]f(x)\,dx\\
&= nh^d\int_{\mathscr{D}} \mathrm{E}\left[\left(g'(m(x))(\tilde{m}_n(x)-m(x))\right.\right.\\
&\qquad \left.\left.+\frac{1}{2}g''(\xi)(\tilde{m}_n(x)-m(x))^2\right)^2\right]f(x)\,dx,
\end{aligned}$$

where ξ lies between $m(x)$ and $\tilde{m}_n(x)$.

Because $|g''(t)| \leqslant C_2$ for all t,

$$\begin{aligned}
&nh^d\int_{\mathscr{D}} \mathrm{Var}[g(\tilde{m}_n(x))]f(x)\,dx\\
&\leqslant 2nh^d\int_{\mathscr{D}}\left(\mathrm{E}[(g'(m(x)))^2(\tilde{m}_n(x)-m(x))^2]\right.\\
&\qquad \left.+\frac{C_2^2}{4}\mathrm{E}[(\tilde{m}_n(x)-m(x))^4]\right)f(x)\,dx \qquad (A.3)\\
&= 2\mathrm{E}[(g'(m(X)))^2Z_n^2(X)] + \frac{C_2^2\mathrm{E}[Z_n^4(X)]}{2nh^d}\\
&\leqslant 2\sqrt{\mathrm{E}[(g'(m(X)))^4]\mathrm{E}[Z_n^4(X)]} + \frac{C_2^2\mathrm{E}[Z_n^4(X)]}{2nh^d}, \qquad (A.4)
\end{aligned}$$

where inequality in (A.3) follows from $(a+b)^2 \leqslant 2a^2 + 2b^2$, and inequality in (A.4) follows from Hölder's inequality (Rudin 1987).

Endnotes

[1] The inner and outer expectations may be taken under a pricing martingale measure and real-world probability measure, respectively, and hence one may simulate X and Y under different measures. However, this issue is of little relevance from simulation perspective.

[2] Results presented in this paper apply to more general kernel functions with bounded supports. However, extension to unbounded kernels may involve additional technicality in asymptotic analysis, mainly in asymptotic variance analysis, which shall be presented in Section 3.1.2.

[3] To be rigorous, the last equality of Equation (11) requires $o_x(1)$ to be uniformly integrable. Throughout the paper, we implicitly assume such a condition when required to avoid too much technicality.

[4] These regularity conditions are summarized by Assumptions EC.1.1 and EC.1.2 in Section EC.1.1 of the e-companion, where discussions of these assumptions are also provided.

[5] Price dynamics are specified under real probability and risk-neutral measures for $t \leqslant \tau$ and $t > \tau$, respectively.

References

Ankenman B, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. *Oper. Res.* 58(2):371–382.

Bank for International Settlements (2012) Fundamental review of the trading book. Consultative document.

Baysal RE, Staum J (2008) Empirical likelihood for value-at-risk and expected shortfall. *J. Risk* 11(1):3–32.

Belomestny D, Kolodko A, Schoenmakers J (2010) Regression methods for stochastic control problems and their convergence analysis. *SIAM J. Control Optim.* 48(5):3562–3588.

Bierens HJ (1985) Kernel estimators of regression functions. Bewley TF, ed. *Advances in Econometrics Fifth World Congress: Volumn I* (Cambridge University Press, Cambridge, UK), 99–144.

Bosq D (1998) *Nonparametric Statistics for Stochastic Processes*, 2nd ed. (Springer, New York).

Britten-Jones M, Schaefer SM (1999) Non-linear value-at-risk. *Eur. Finance Rev.* 2:161–187.
Broadie M, Glasserman P (2004) A stochastic mesh method for pricing high-dimensional American options. *J. Computational Finance* 7 (4):35–72.
Broadie M, Du Y, Moallemi CC (2011) Efficient risk estimation via nested sequential simulation. *Management Sci.* 57(6):1172–1194.
Broadie M, Du Y, Moallemi CC (2015) Risk estimation via regression. *Oper. Res.* 63(5)(6):1077–1097.
Carriere JF (1996) Valuation of early-exercise price for options using simulations and nonparametric regression. *Insurance: Math. Econom.* 19(1):19–30.
Dimitroff G, Lorenz S, Szimayer A (2011) A parsimonious multi-asset Heston model: Calibration and derivation pricing. *Internat. J. Theoret. Appl. Finance* 14(8):1299–1333.
Duffie D (1996) *Dynamic Asset Pricing Theory* (Princeton University Press, Princeton, NJ).
Duffie D, Pan J (2001) Analytical value-at-risk with jumps and credit risk. *Finance and Stochastics* 5(2):155–180.
Durrett R (2005) *Probability: Theory and Examples*, 3rd ed. (Duxbury Press, Belmont, CA).
Fouque J-P, Papanicolaou G, Sircar KR (2000) *Derivatives in Financial Markets with Stochastic Volatility* (Cambridge University Press, New York).
Gasser T, Engel J (1990) The choice of weights in kernel regression estimation. *Biometrika* 77(2):377–381.
Glasserman P (2004) *Monte Carlo Methods in Financial Engineering* (Springer, New York).
Gordy M, Juneja S (2006) Efficient simulation for risk measurement in portfolio of CDOs. Perrone LF, Lawson BL, Liu J, Wieland FP, eds. *Proc. 2006 Winter Simulation Conf., WSC '06* (IEEE, Piscataway, NJ), 749–756.
Gordy M, Juneja S (2010) Nested simulation in portfolio risk measurement. *Management Sci.* 56(10):1833–1848.
Härdle W (1990) *Applied Nonparametric Regression* (Cambridge University Press, New York).
Hong LJ, Juneja S (2009) Estimating mean of non-linear function of conditional expectation. Mason SJ, Hill RR, Mönch L, Rose O, Jefferson T, Fowler JW, eds. *Proc. 2009 Winter Simulation Conf., WSC '09* (IEEE, Piscataway, NJ), 1223–1236.
Jennen-Steinmetz C, Gasser T (1988) A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* 83(404):1084–1089.
Lan H, Nelson BL, Staum J (2010) A confidence interval procedure for expected shortfall risk measurement via two-level simulation. *Oper. Res.* 58(5):1481–1490.
Lee SH (1998) Monte Carlo expectation of conditional quantiles. Ph.D. Thesis, Department of Operations Research, Stanford University.
Li Q, Racine JS (2007) *Nonparametric Econometrics: Theory and Practice* (Princeton University Press, Princeton, NJ).
Liu M, Staum J (2010) Stochastic kriging for efficient nested simulation of expected shortfall. *J. Risk* 12(3):3–27.
Liu M, Nelson BL, Staum J (2010) An efficient simulation procedure for point estimation of expected shortfall. Johansson B, Jain S, Montoya-Torres J, eds. *Proc. 2010 Winter Simulation Conf., WSC '10* (IEEE, Piscataway, NJ), 2782–2789.
Longstaff FA, Schwartz ES (2001) Valuing American options by simulation: A simple least-squares approach. *Rev. Financial Stud.* 14(1):113–147.
Nadaraya EA (1964) On estimating regression. *Theory Probab. ITS Appl.* 9(1):141–142.
Rouvinez C (1997) Going Greek with VaR. *Risk* 10(2):57–63.
Rudin W (1987) *Real and Complex Analysis*, 3rd ed. (McGraw-Hill, Singapore).
Seifert B, Brockmann M, Engel J, Gasser T (1994) Fast algorithms for nonparametric curve estimation. *J. Computational and Graphical Statist.* 3(2):192–213.
Shreve SE (2004) *Stochastic Calculus for Finance II: Continuous-Time Models* (Springer, New York).
Tsitsiklis JN, Van Roy B (2001) Regression method for pricing complex American-style options. *IEEE Trans. Neural Networks* 12(4):694–703.
Watson GS (1964) Smooth regression analysis. *Sankhyā, Ser. A*, 26:359–372.
World Federation of Exchanges (2014) WFE/IOMA derivative market survey 2013. Accessed October 9, 2014, www.world-exchanges.org/files/file/IOMA/2013_IOMA_Survey_final.pdf.

L. Jeff Hong is the Chair Professor of Management Sciences in the College of Business at City University of Hong Kong. His research interests include stochastic simulation, stochastic optimization, business analytics, and financial risk management.

Sandeep Juneja is a professor at the School of Technology and Computer Science at the Tata Institute of Fundamental Research in Mumbai. His research interests are primarily in applied probability with a focus on Monte Carlo methods, computational and mathematical finance, as well as game theoretic modelling in queues.

Guangwu Liu is an associate professor in the Department of Management Sciences at City University of Hong Kong. His research interests include stochastic simulation, financial engineering, and risk management.