Distributional Replication

Brendan K. Beare*†

University of California, San Diego

August 7, 2009

Abstract: Suppose that X and Y are random variables. We define a replicating function to be a function θ such that $\theta(X)$ and Y have the same distribution. In general, the set of replicating functions for a given pair of random variables may be infinite. Suppose we have some objective function, or cost function, defined over the set of replicating functions, and we seek to estimate the replicating function with the lowest cost. We develop an approach to estimating the cheapest replicating function that involves minimizing the cost function over an estimate of the set of replicating functions. Our estimated set of replicating functions is obtained by considering the functions θ in some sieve space Θ_n for which the empirical distributions of $\theta(X)$ and Y are close. Under suitable conditions, we show that our estimated function comes close to achieving distributional replication, and close to achieving the minimum cost among replicating functions. We discuss the relevance of our results to the financial literature on hedge fund replication; in this case, X is the market return, Y is the return from a hedge fund or other asset, and our estimation procedure amounts to choosing the cheapest portfolio of options on X such that the returns from our portfolio have the same distribution as the hedge fund returns.

Keywords and phrases: distributional replication; sieve estimation; hedge fund replication.

JEL classifications: primary C14; secondary G11.

*Address: Department of Economics, UCSD, 9500 Gilman Drive, La Jolla CA 92093-0508, USA. Email: bbeare@ucsd.edu.

†I thank Mark Machina, Andres Santos, Xiaoxia Shi, Max Stinchcombe, Hal White, and seminar participants at Oxford, Stanford, UC Berkeley, UC Davis and UC San Diego for helpful comments. I also thank Patrick Huggins for introducing me to the literature on hedge fund replication.

1 Introduction

Suppose that X and Y are random variables. In this paper we consider estimating a function θ such that $\theta(X)$ and Y have the same distribution. Such a function is said to be a *replicating function*. Typically, there are many different replicating functions for a given pair of random variables X and Y. We suppose that to each function θ there corresponds a "price", denoted $p(\theta)$, and we seek to estimate the replicating function θ for which $p(\theta)$ is as small as possible. That is, we seek to estimate the cheapest replicating function for a given X and Y. To estimate this function from a sample of realizations of X and Y, we first obtain an estimate of the set of all replicating functions. The estimated set is formed by choosing a rich but manageable class of functions (i.e., a sieve space) and taking all those functions θ in that class for which the distance between the empirical distributions of $\theta(X)$ and Y is small. Our estimate of the cheapest replicating function is then obtained by minimizing p over the estimated set of replicating functions.

Our research in this area is motivated by a recent literature in applied finance on "hedge fund replication". The hedge fund replication literature is concerned with the possibility of achieving financial returns that resemble those of a particular hedge fund, fund of hedge funds, or index of hedge funds, by engaging in an investment strategy that does not involve a direct investment in the fund or funds in question. Ideally, the replicating strategy should involve trading assets that are highly liquid, thereby avoiding the barriers to entry, lock-in periods and high fees that are characteristic of hedge fund investments. Several major investment banks have launched hedge fund replication products, including Goldman Sachs and Merrill Lynch in 2006 and J.P. Morgan in 2007; see Kat (2007) for further details. Hedge fund replication strategies have also attracted the attention of the popular press, with articles appearing in *The Wall Street Journal* (Laise, 2007) and *The New Yorker* (Cassidy, 2007), among other outlets.

There are two broad streams of the hedge fund replication literature. In one stream, researchers have considered the direct approximation of hedge fund returns by investing in a portfolio of other assets. By direct approximation, we mean that the returns from the selected portfolio should be close to the hedge fund returns with high probability. Typically, the replicating strategy amounts to estimating a factor model for hedge fund returns, and then investing directly in the factors rather than in the hedge fund. Hasanhodzic and Lo (2007) provide a recent study of this kind, and survey the related literature. The second stream of the hedge fund replication literature is concerned with the distributional approximation of hedge fund returns, rather than their direct approximation. The aim here is to create a trading strategy that generates returns with the same statistical distribution as the hedge fund returns. This is a more modest goal than direct approximation, because in any given period the return generated by the replicating strategy need not resemble the return from the hedge fund. Key papers in this stream of the hedge fund replication literature

include Amin and Kat (2003), Kat and Palaro (2005ab, 2006), and Kat (2007). The results in this paper extend the approach taken by these authors.

Suppose that X represents the payoff after one month from a \$1 investment in a market index, while Y represents the payoff after one month from a \$1 investment in some hedge fund. Amin and Kat (2003) propose to estimate a function θ such that $\theta(X)$ and Y have the same distribution function. Given a sample of n realizations of X and Y, their estimated replicating function is $\hat{\theta}_n = \hat{Q}_n^Y \circ \hat{F}_n^X$, where $\hat{Q}_n^Y$ is an estimate of Q^Y, the quantile function of Y, and $\hat{F}_n^X$ is an estimate of F^X, the distribution function of X. Assuming continuity of F^X, the random variable $Q^Y(F^X(X))$ has the same distribution as Y, implying that $Q^Y \circ F^X$ is a replicating function. Loosely speaking, we might therefore expect $\hat{\theta}_n(X)$ and Y to have similar distributions in large samples. The estimated function $\hat{\theta}_n$ can be thought of as describing the payoff after one month of a derivative security written on the market index. Under suitable conditions, this payoff can be achieved using a continuously rebalanced self-financed portfolio of market shares and cash, as in the hedging strategy used to justify the celebrated Black-Scholes-Merton option pricing formula (Black and Scholes, 1973; Merton, 1973). We let $p(\theta)$ denote the start-up cost of a hedging strategy with payoff $\theta(X)$, and refer to this quantity as the price of θ.

It need not be the case that $p(\theta) = 1$ when θ is a replicating function. This is because the distributional equivalence of $\theta(X)$ and Y does not imply the existence of an arbitrage opportunity when their initial investment costs differ. Indeed, two replicating functions need not have the same price. Amin and Kat (2003) aim to estimate the particular replicating function $Q^Y \circ F^X$ because it is an increasing function of the market payoff X. In Dybvig (1988ab), it is shown under very general conditions that, given a collection of payoff functions that all achieve the same payoff distribution, the cheapest such function must allocate payoffs to states as a nonincreasing function of the state prices. Amin and Kat (2003) observe that in a Black-Scholes world, the state price density (with respect to the true probability measure over states) is inversely related to X. Thus, the cheapest replicating function must be a nondecreasing function of X.

A key difference between the approach to distributional replication proposed in this paper, and the approach taken by Amin and Kat (2003), is that we do not assume that the cheapest replicating function is nondecreasing. Instead, we search for the cheapest replicating function over a large space of functions, many of which are not monotone. Empirically, there is good reason to believe that the cheapest replicating function will not be monotone. Jackwerth (2000) and Brown and Jackwerth (2004) argue that the state price density (in their terminology, pricing kernel) implied by S&P500 options with one month to expiry changed dramatically after the stock market crash of 1987, becoming nonmonotone with respect to the return on the S&P500 index. See, in particular, Figure 2 in Brown and Jackwerth (2004), in which the state price density is an increasing function of the market return for monthly return levels between approximately -3% and

3%, and decreasing elsewhere. If the relationship between the state price density and the market return is not monotone, then the results of Dybvig (1988ab) imply that the cheapest replicating function θ will not be monotone. In this case, the approach to distributional replication proposed here has a clear advantage over existing techniques.

There is a second major conceptual difference between the approach to distributional replication proposed here, and the approach taken by Amin and Kat (2003). Amin and Kat propose to implement the desired payoff function θ by engaging in a continuous time hedging strategy, trading market shares and cash. In this paper, we propose to approximate θ by investing in a portfolio of European put and call options written on the market index at various strike prices. The portfolio may also include the market index itself, and risk-free zero-coupon bonds. A key advantage of our approach is that the price of the payoff function θ corresponding to such a portfolio may be calculated directly from observed option and bond prices. By comparison, Amin and Kat price θ by taking the risk neutral expected payoff of $\theta(X)$ under Black-Scholes conditions, and they require Black-Scholes conditions to hold in order for their hedging strategy to achieve the desired payoff. The empirical limitations of the Black-Scholes pricing model have been extensively documented. We avoid these difficulties by confining ourselves to functions θ for which the market price is directly observable, and which may be implemented in practice by investing directly in a portfolio of actively traded securities.

We embed our approach in the statistical framework of sieve estimation by assuming that the set of strike prices at which options may be traded becomes more dense as the sample size n increases, at a controlled rate. The payoff functions achievable using portfolios of this kind are continuous piecewise linear functions, with kinks at the allowable strike prices. We control the complexity of this class of functions using the notion of VC-dimension (Vapnik and Červonenkis, 1971), and are thereby able to bring the machinery of empirical process theory to bear in analyzing the asymptotic properties of our technique. The use of option payoff functions to form the basis for a sieve space is not entirely without precedent. Option payoffs appear as activation functions in the regularized neural network model studied by Corradi and White (1995): take $m = 2$ in their equation (4.1). Those authors do not, however, explicitly discuss the connection to option payoffs and portfolio choice.

The approach taken by Amin and Kat (2003), and in this paper, aims to replicate the univariate distribution of Y. Typically, the joint distribution of $\theta(X)$ with any other asset payoff will differ from the joint distribution of Y and that asset payoff. In particular, the joint distribution of $\theta(X)$ and the market payoff X will differ from the joint distribution of Y and X, and for this reason we cannot expect investors to find $\theta(X)$ to be a perfect substitute for Y in general. Intuitively, if the correlation between X and Y is lower than the correlation between X and $\theta(X)$, risk-averse investors may prefer a balanced portfolio formed from X and Y to a similar portfolio formed from X and $\theta(X)$. In response to this issue, Kat and Palaro (2005ab) extend

the approach of Amin and Kat (2003) to the replication of bivariate distributions. They introduce a "reserve asset" with payoff Z, and seek to find a bivariate function θ such that the joint distribution of $\theta(X, Z)$ and X is the same as the joint distribution of Y and X. This replicating payoff function is implemented in practice using a continuously rebalanced portfolio formed by trading market shares, cash, and the reserve asset. We do not follow that approach in this paper, in part because it is generally not feasible to approximate a wide class of bivariate functions using a portfolio formed from options written on individual assets. Confining ourselves to the replication of univariate distributions may not seem unreasonable if we modify our interpretation of the random variable Y. Rather than representing the payoff from a \$1 investment in a hedge fund, Y could represent the payoff from a \$1 investment in a portfolio partly invested in the hedge fund and partly in the market index. Amin and Kat (2003) take this approach in their empirical study of hedge fund efficiency. More generally, Y could be the payoff from a \$1 investment in a portfolio formed from any number of arbitrary assets. If $\theta(X)$ has the same distribution as Y, and the price of θ is less than \$1, an investor should prefer to invest in the replicating portfolio.

The remainder of this paper is structured as follows. In sections 2 and 3 we develop a general approach to the estimation of replicating functions, without explicit reference to the financial application that serves as our motivation. In section 2 we provide some basic mathematical tools for dealing with the notion of replicating functions, while in section 3 we discuss the statistical estimation of replicating functions using the method of sieves. In section 4 we explain how the mathematical material in sections 2 and 3 can be applied to the problem of hedge fund replication. Section 5 outlines some areas for future research, and concludes. Throughout the paper, there are several numbered assumptions and theorems. In the statement of each theorem, it is implicit that all assumptions introduced prior to the theorem hold.

2 Replicating Functions

In this section we formally introduce the notion of a replicating function. We construct a pseudometric on the set of Borel measurable functions mapping the support of one random variable to the support of another, and we define a criterion function that identifies the set of replicating functions. Some useful results relating to these objects are given.

Let X and Y be real valued random variables. The probability space on which X and Y are defined is unimportant; in fact, we are concerned only with the probability distributions that X and Y induce on the real line. Let $P^X : \mathcal{B}(\mathbb{R}) \to [0,1]$ and $P^Y : \mathcal{B}(\mathbb{R}) \to [0,1]$ denote the probability measures corresponding to X and Y. $\mathcal{B}(\mathbb{R})$ denotes the usual Borel σ-field on $\mathbb{R}$. Let $F^X : \mathbb{R} \to [0,1]$ and $F^Y : \mathbb{R} \to [0,1]$ denote the distribution functions of X and Y. Let $R^X = \text{cl}(\{x \in \mathbb{R} : 0 < F^X(x) < 1\})$, and let $R^Y = \text{cl}(\{y \in \mathbb{R} : 0 < F^Y(y) < 1\})$;

here, $\text{cl}(A)$ denotes the Euclidean closure of a set $A \subseteq \mathbb{R}$. R^X and R^Y are intervals of the form $[a, b]$, $[a, \infty)$ or $(\infty, b]$, with $a, b \in \mathbb{R}$.

We place the following condition on F^X and F^Y.

Assumption 2.1. F^X and F^Y are continuous and strictly increasing on R^X and R^Y respectively.

Assumption 2.1 is stronger than is required to establish all of the results in this paper, but it will be convenient for us to maintain Assumption 2.1 throughout. Under Assumption 2.1, the restriction of F^X to R^X is a continuous and strictly increasing function, and therefore uniquely defines a continuous and strictly increasing inverse function $Q^X : F^X(R^X) \to R^X$. We refer to this function as the quantile function of X. The quantile function of Y, denoted $Q^Y : F^Y(R^Y) \to R^Y$, is defined in the same way. Note that $F^X(R^X)$ and $F^Y(R^Y)$ are equal to $(0, 1]$, $[0, 1)$, $[0, 1]$ or $(0, 1)$, depending on whether X and Y are almost surely bounded above, below, both, or neither.

Let Θ denote the set of all Borel measurable functions $\theta : R^X \to R^Y$. Though Θ depends on X and Y, we do not make this dependence explicit in our notation. We are interested in those functions $\theta \in \Theta$ for which $\theta(X)$ and Y have the same distribution.

Definition 2.1. A function $\theta \in \Theta$ is called a *replicating function for X and Y*, or simply a *replicating function* or *replicator*, if $P^X \theta^{-1} B = P^Y B$ for all $B \in \mathcal{B}(\mathbb{R})$.

Note that a replicating function does not describe a relationship between X and Y in the usual sense. $\theta(X)$ and Y may be perfectly correlated, or independent. All that matters is that they have the same marginal distribution. We will let Θ^* denote the set of all replicating functions for X and Y. Again, the dependence of Θ^* on X and Y is not made explicit in our notation.

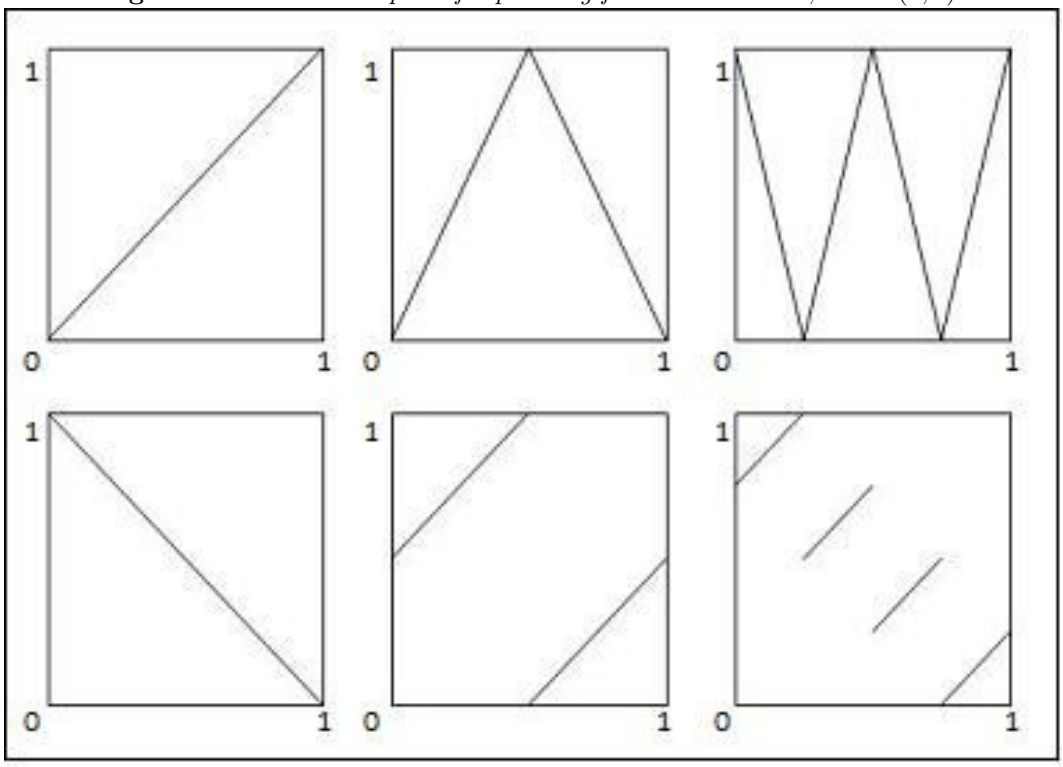
Our first result concerns the cardinality of Θ^*.

Theorem 2.1. *Θ^* is uncountably infinite. Moreoever, there exists an uncountable subset of Θ^* in which no two functions are equal on a set of positive P^X-measure.*

Remark 2.1. One example of a replicating function is the composition $Q^Y \circ F^X$, restricted to R^X. Clearly, if F^X is not continuous and F^Y is continuous, so that Assumption 2.1 is violated, Θ^* is empty.

We will sometimes find it helpful to consider the special case where X and Y are both distributed uniformly on the unit interval. In this case, the composition $\theta = Q^Y \circ F^X$ restricted to $R^X = [0, 1]$ is the identity function, $\theta(x) = x$. Another simple example of a replicating function is $\theta(x) = 1 - x$, restricted to $[0, 1]$. Graphs of these functions, and of four other replicating functions, are provided in Figure 2.1. We will let $\tilde{\Theta}$ denote the set of all Borel measurable functions $\theta : [0, 1] \to [0, 1]$, and let $\tilde{\Theta}^*$ denote the set of functions in $\tilde{\Theta}$ that are replicators when $X, Y \sim U(0, 1)$.

Figure 2.1: *Some examples of replicating functions when* $X, Y \sim U(0,1)$.



As an aid to visualizing the functions in $\tilde{\Theta}^*$, a reader familiar with the concept of local time may find it helpful to think of each function $\theta \in \tilde{\Theta}$ as a (deterministic) stochastic process on the unit interval. The functions $\theta \in \tilde{\Theta}^*$ are precisely those for which the local time at y is equal to one for each $y \in (0,1)$. That is, for $\theta \in \tilde{\Theta}$, we have $\theta \in \tilde{\Theta}^*$ if and only if

$$\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_0^1 1(|\theta(x) - y| \leq \varepsilon) \mathrm{d}x = 1$$

for each $y \in (0,1)$. This can be shown by observing that the above limit is equal to the derivative of the distribution function of $\theta(X)$ at y when $X \sim U(0,1)$.

We now introduce a pseudometric d on Θ. For $\theta_0, \theta_1 \in \Theta$, let $d : \Theta \times \Theta \to \mathbb{R}$ be given by

$$d(\theta_0, \theta_1) = \int_{R^X} |F^Y(\theta_1(x)) - F^Y(\theta_0(x))| \mathrm{d}F^X(x).$$

It is obvious that d satisfies the four axioms for a pseudometric: nonegativity, symmetry, the triangle inequality, and the requirement that $d(\theta, \theta) = 0$ for all $\theta \in \Theta$. d is not a metric because we will have $d(\theta_0, \theta_1) = 0$ when θ_0 and θ_1 are equal on a set of P^X-measure one, even if the two functions are distinct.

Note that when $X, Y \sim U(0,1)$, d corresponds to the usual L_1-seminorm for functions on $[0,1]$. When X and Y are not uniform, $d(\theta_0, \theta_1)$ is equal to the L_1 distance between the deformed functions $F^Y \circ \theta_0 \circ Q^X$ and $F^Y \circ \theta_1 \circ Q^X$.

We now introduce a nonnegative function $M : \Theta \to \mathbb{R}$ that is intended to quantify the extent to which a function $\theta \in \Theta$ achieves distributional replication. For $\theta \in \Theta$, let $F^X(\cdot;\theta)$ denote the distribution function of $\theta(X)$; that is, for $y \in \mathbb{R}$ and $\theta \in \Theta$, let $F^X(y;\theta) = P^X\theta^{-1}(-\infty, y]$. Define

$$M(\theta) = \int_{R^Y} |F^X(y;\theta) - F^Y(y)| \mathrm{d}F^Y(y).$$

Our pseudometric d endows M with a very convenient smoothness condition. Specifically, M is Lipschitz continuous with respect to d, with Lipschitz constant no greater than one.

Theorem 2.2. *For all $\theta_0, \theta_1 \in \Theta$, we have $|M(\theta_1) - M(\theta_0)| \leq d(\theta_0, \theta_1)$.*

Our next result concerns the identification of the set of replicators Θ^* using the criterion function M. It states that the set of replicators Θ^* is precisely those functions $\theta \in \Theta$ for which $M(\theta) = 0$.

Theorem 2.3. $\Theta^* = \{\theta \in \Theta : M(\theta) = 0\}$.

Theorem 2.2 and Theorem 2.3 jointly imply that, if $\theta_1, \theta_2, \ldots$ is a sequence of elements of Θ converging to some $\theta^* \in \Theta^*$ in the pseudometric d, then $M(\theta_n) \to 0$ as $n \to \infty$. We would like to interpret this to mean that θ_n gets arbitrarily close to achieving distributional replication as n becomes larger. The next result makes this notion precise.

Theorem 2.4. *Let $\theta_1, \theta_2, \ldots$ be a sequence of elements of Θ. Then, as $n \to \infty$, $M(\theta_n) \to 0$ if and only if $F^X(y;\theta_n) \to F^Y(y)$ for each $y \in \mathbb{R}$.*

Remark 2.2. Note that, since F^Y is continuous, pointwise convergence of $F^X(\cdot;\theta_n)$ to $F^Y(\cdot)$ is equivalent to the statement $P^X\theta_n^{-1} \Rightarrow P^Y$, where "$\Rightarrow$" denotes weak convergence of probability measures (see e.g. Billingsley, 1968), and $P^X\theta_n^{-1}$ is the measure on $\mathcal{B}(\mathbb{R})$ given by $P^X\theta_n^{-1}B = P^X\{x \in R^X : \theta_n(x) \in B\}$ for each $B \in \mathcal{B}(\mathbb{R})$. We could also write this statement as $\theta_n(X) \to_d Y$, where "$\to_d$" denotes convergence in distribution in the usual sense.

Our final result of this section is a modification of Theorem 2.4 that allows $\theta_1, \theta_2, \ldots$ to be random elements. An obvious first step towards defining such random elements would be to introduce a σ-field on Θ; however, such an approach leads to complications relating to the measurability of $\theta(X)$ when θ and X are both random. We will need to require each of the random elements θ_n, $n \in \mathbb{N}$, to be a random element

of some subspace $\Theta_n \subset \Theta$. Each subspace Θ_n will be equipped with a σ-field $\mathcal{T}_n$ that is well behaved in the following sense.

Definition 2.2. Given a collection of functions $\Theta' \subseteq \Theta$, an *admissible structure for* Θ' is a σ-field $\mathcal{T}'$ of subsets of Θ' such that the evaluation mapping $(\theta, x) \mapsto \theta(x)$ is a measurable map from $(\Theta' \times R^X, \mathcal{T}' \otimes \mathcal{B}(R^X))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 2.2 is a version of a definition of admissibility given in section 5.2 of Dudley (1999). $\mathcal{B}(R^X)$ denotes the Borel σ-field on R^X, while the notation $\mathcal{T}' \otimes \mathcal{B}(R^X)$ refers to the product σ-field on $\Theta' \times R^X$; that is, the σ-field on $\Theta' \times R^X$ generated by sets of the form $A \times B$, with $A \in \mathcal{T}'$ and $B \in \mathcal{B}(R^X)$. With Definition 2.2 in hand, we are now in a position to state the final result of this section.

Theorem 2.5. *Let $\Theta_1, \Theta_2, \ldots$ be a sequence of subsets of Θ, and for each $n \in \mathbb{N}$ let $\mathcal{T}_n$ be an admissible structure for Θ_n and let P^{θ_n} be a probability measure on $\mathcal{T}_n$. Let $P^{\theta_n(X)}$ be the probability measure on $\mathcal{B}(\mathbb{R})$ given by $P^{\theta_n(X)}B = P^{\theta_n} \otimes P^X\{(\theta, x) \in \Theta_n \times R^X : \theta(x) \in B\}$ for each $B \in \mathcal{B}(\mathbb{R})$, where $P^{\theta_n} \otimes P^X$ is the product measure on $\mathcal{T}_n \otimes \mathcal{B}(R^X)$. Then, as $n \to \infty$, if $\int_{\Theta_n} M dP^{\theta_n} \to 0$ then also $P^{\theta_n(X)} \Rightarrow P^Y$.*

Remark 2.3. It is possible to rephrase Theorem 2.5 in a somewhat less precise fashion that may be easier to interpret. For each $n \in \mathbb{N}$, we can think of the measure P^{θ_n} as corresponding to a random function θ_n taking values in Θ_n. The measure $P^{\theta_n(X)}$ describes the distribution of $\theta_n(X)$ when θ_n and X are both random, and θ_n is independent of X. The statement $\int_{\Theta_n} M \mathrm{d}P^{\theta_n} \to 0$ can be written as $EM(\theta_n) \to 0$. Thus, the final statement of Theorem 2.5 could be written as follows: as $n \to \infty$, if $EM(\theta_n) \to 0$ then also $\theta_n(X) \to_d Y$.

3 Sieve Estimation of Replicating Functions

In this section we turn our attention to the statistical estimation of a replicating function using a sample of observations $\{(X_i, Y_i) : 1 \leq i \leq n\}$.

Assumption 3.1. $\{X_i : i \in \mathbb{N}\}$ and $\{Y_i : i \in \mathbb{N}\}$ are iid collections of real valued random variables defined on a complete probability space $(\Omega, \mathcal{F}, P)$. Each X_i has distribution function F^X, and each Y_i has distribution function F^Y.

Remark 3.1. The iid condition in Assumption 3.1 refers to the independence of X_i and X_j, and of Y_i and Y_j, when $i \neq j$. X_i and Y_j may be dependent for any i, j.

Remark 3.2. The assumption that $(\Omega, \mathcal{F}, P)$ is complete will be useful later when we employ a result due to Stinchcombe and White (1992) that provides conditions under which certain real valued functions on Ω

are analytic. The interested reader may refer to that paper for the definition of an analytic function. When $(\Omega, \mathcal{F}, P)$ is complete, real valued functions on Ω are analytic if and only if they are measurable.

We wish to use our observed sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$ to construct an estimate of a replicating function that has good properties when n is large. As was made clear in Theorem 2.1, the set of replicating functions is uncountably infinite in a nontrivial sense. We are thus confronted with the problem of partial identification: the distributional replication property does not uniquely identify the function we are seeking to estimate. The first step in our estimation procedure is to empirically discriminate between those functions that come close to achieving distributional replication, and those that do not. In the previous section, the function $M : \Theta \to \mathbb{R}$ was used to quantify the extent to which a function $\theta \in \Theta$ achieves distributional replication. We will construct an empirical analogue to M. Given a sample of size n, let $F_n^Y : \mathbb{R} \to [0, 1]$ denote the empirical distribution function of Y, and for $\theta \in \Theta$ let $F_n^X(\cdot; \theta) : \mathbb{R} \to [0, 1]$ denote the empirical distribution function of $\theta(X)$. That is,

$$\begin{aligned} F_n^Y(y) &= \frac{1}{n} \sum_{i=1}^{n} 1(Y_i \leq y), \\ F_n^X(y; \theta) &= \frac{1}{n} \sum_{i=1}^{n} 1(\theta(X_i) \leq y). \end{aligned}$$

Let the function $M_n : \Theta \to \mathbb{R}$ be defined by

$$M_n(\theta) = \int |F_n^X(y; \theta) - F_n^Y(y)| \mathrm{d}F_n^Y(y) = \frac{1}{n} \sum_{i=1}^{n} |F_n^X(Y_i; \theta) - F_n^Y(Y_i)|.$$

M_n will serve as our empirical analogue to M. Note that we have suppressed the dependence of F_n^Y, F_n^X and M_n on $\omega \in \Omega$ in our notation.

We would like M_n to serve as a good approximation to M when n is large. Unfortunately, the space Θ is too rich for us to expect M_n to be close to M uniformly over Θ. We shall instead consider the approximation of M by M_n over a more manageable subset of the functions in Θ. We will consider a sequence of such subsets $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$, with Θ_n becoming more complex as n grows, but at a slow enough rate to allow the uniform approximation error $\sup_{\theta \in \Theta_n} |M_n(\theta) - M(\theta)|$ to decay to zero in a suitable sense. Our approach may be regarded as a version of the method of sieve estimation. For a general discussion of sieve estimation in econometrics, the reader may refer to Chen (2007).

To formalize the complexity of Θ_n, we shall employ the notion of VC-major dimension. VC-major dimension is a characterization of complexity for classes of functions that is related to the notion of VC-dimension for classes of sets.

Definition 3.1. Let $\mathcal{C}$ be a collection of subsets of $\mathbb{R}$. $\mathcal{C}$ is said to *shatter* a set of points $D = \{x_1, \ldots, x_d\} \subset \mathbb{R}$, $d \in \mathbb{N}$, if all 2^d subsets of D can be written as the intersection of D with some set in $\mathcal{C}$. $\mathcal{C}$ is said to be a *VC-class* if, for some $d \in \mathbb{N}$, $\mathcal{C}$ cannot shatter any set of size d. If $\mathcal{C}$ is a VC-class then the *VC-dimension of* $\mathcal{C}$, written $\mathcal{V}(\mathcal{C})$, is defined to be the smallest $d \in \mathbb{N}$ for which no set of size d is shattered by $\mathcal{C}$. If $\mathcal{C}$ is not a VC-class, we set $\mathcal{V}(\mathcal{C}) = \infty$.

Definition 3.1 is standard in the literature on empirical processes; see e.g. section 2.6.1 in van der Vaart and Wellner (1996). Building on Definition 3.1, we define the VC-major dimension of a subset of Θ as follows.

Definition 3.2. Consider a collection of functions $\Theta' \subseteq \Theta$. A subset of $\mathbb{R}$ is said to be *majorized* by Θ' if it can be written as $\{x \in R^X : \theta(x) > c\}$ for some $\theta \in \Theta'$ and some $c \in \mathbb{R}$. Let $\mathcal{C}$ denote the collection of all sets majorized by Θ'. We say that Θ' is a *VC-major class* if $\mathcal{C}$ is a VC-class. The *VC-major dimension of* Θ', written $\mathcal{V}(\Theta')$, is defined to be the VC-dimension of $\mathcal{C}$.

Remark 3.3. The definition of VC-major dimension should not be confused with that of VC-subgraph dimension, which also appears frequently in the empirical process literature; in general, the two are different. When Θ' is the set of indicator functions of a collection of sets $\mathcal{C}$, the VC-major dimension and VC-subgraph dimension of Θ' are both equal to the VC-dimension of $\mathcal{C}$. Sections 2.6.2 and 2.6.4 in van der Vaart and Wellner (1996) provide discussions of VC-subgraph and VC-major classes respectively.

We will control the complexity of the spaces Θ_n by bounding the growth rate of their VC-major dimension. In addition, we will need to introduce some additional technical conditions to ensure the measurability of certain real valued functions on Ω. For $\Theta' \subseteq \Theta$, let $\mathcal{B}(\Theta')$ denote the Borel σ-field on Θ' induced by the pseudometric d.

Assumption 3.2. For each $n \in \mathbb{N}$, $\Theta_n \subseteq \Theta$ is a nonempty VC-major class. Further, $\mathcal{B}(\Theta_n)$ is an admissible structure on Θ_n, and $(\Theta_n, \mathcal{B}(\Theta_n))$ is a Souslin measurable space.

Remark 3.4. The reader is referred to Stinchcombe and White (1992) for the definition of a Souslin measurable space, and further discussion. Here, we note only that for $(\Theta_n, \mathcal{B}(\Theta_n))$ to be a Souslin measurable space, it suffices that (Θ_n, d) is a Polish metric space; that is, (Θ_n, d) is a metric space that is topologically isomorphic to a complete separable metric space.

The following result shows how the magnitude of the uniform approximation error $\sup_{\theta \in \Theta_n} |M_n(\theta) - M(\theta)|$ relates to $\mathcal{V}(\Theta_n)$.

Theorem 3.1. *As $n \to \infty$, we have $E \sup_{\theta \in \Theta_n} |M_n(\theta) - M(\theta)| = O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right)$.*

Remark 3.5. In the proof of Theorem 3.1 it is established that $\sup_{\theta\in\Theta_n} |M_n(\theta) - M(\theta)|$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Thus, our statement of Theorem 3.1 uses the ordinary expectation operator. It is common in the empirical process literature to see results of this kind expressed in terms of outer expectation; see e.g. section 1.2 in van der Vaart and Wellner (1996).

Theorem 3.1 indicates that, if $\mathcal{V}(\Theta_n) = o(n)$, then when n is large we can use the empirical criterion function M_n to distinguish between those functions in Θ_n that are close to achieving distributional replication, and those that are not. We have yet to address the issue of partial identification: there may be many functions in Θ_n that are close to achieving distributional replication. We wish to entertain the possibility that not all replicating functions are created equal. Let $p : \Theta \to \mathbb{R}$ be a function describing the "price" of each function $\theta \in \Theta$. Rather than seeking to estimate an arbitrary replicating function, we will seek to estimate a replicating function θ for which $p(\theta)$ is as small as possible.

Assumption 3.3. The function $p : \Theta \to \mathbb{R}$ is nonnegative, and continuous with respect to d.

Loosely speaking, we seek to estimate the cheapest, or optimal, replicating function. The following result concerns the selection of our estimated function $\hat{\theta}_n$. In it, we make the random nature of M_n explicit by writing M_n as a function of both $\omega \in \Omega$ and $\theta \in \Theta$.

Theorem 3.2. *Let $\epsilon_1, \epsilon_2, \ldots$ and $\lambda_1, \lambda_2, \ldots$ be sequences of positive real numbers. For each $n \in \mathbb{N}$, there exists a measurable function $\hat{\theta}_n$ from $(\Omega, \mathcal{F})$ to $(\Theta_n, \mathcal{B}(\Theta_n))$ that satisfies $\hat{\theta}_n(\omega) \in \hat{\Theta}_n^*(\omega)$ and $p(\hat{\theta}_n(\omega)) \leq \inf_{\theta\in\hat{\Theta}_n^*(\omega)} p(\theta) + \epsilon_n$ for all $\omega \in \Omega$, where*

$$\hat{\Theta}_n^*(\omega) = \left\{ \theta \in \Theta_n : M_n(\omega, \theta) \leq \inf_{\vartheta\in\Theta_n} M_n(\omega, \vartheta) + \lambda_n \right\}.$$

Remark 3.6. The mathematical content of Theorem 3.2 is the existence of a random function $\hat{\theta}_n$ satisfying the stated conditions. The proof applies the Sainte-Beauve measurable selection theorem (see Corollary 5.3.2 in Dudley, 1999) and Theorem 2.17 of Stinchcombe and White (1992), which concerns the measurability of the suprema of random functions over random sets. Theorem 3.2 also serves to define our estimated replicating function $\hat{\theta}_n$. That is, we take $\hat{\theta}_n$ to be any random function satisfying the conditions given in Theorem 3.2.

Remark 3.7. The random set $\hat{\Theta}_n^*$ can be viewed as our estimate of the set of replicators Θ^*. It consists of all those functions $\theta \in \Theta_n$ such that $M_n(\theta)$ comes close to achieving its infimum over Θ_n. Note that this infimum is not necessarily achieved by any $\theta \in \Theta_n$. The tuning parameter λ_n governs how close $M_n(\theta)$ must be to $\inf_{\vartheta\in\Theta_n} M_n(\vartheta)$ before θ is admitted into the set $\hat{\Theta}_n^*$. We will require that $\lambda_n \to 0$ as $n \to \infty$, but at a

rate that is not too fast. $\hat{\theta}_n$ is chosen such that $\hat{\theta}_n(\omega) \in \hat{\Theta}_n^*(\omega)$ for each $\omega \in \Omega$. Thus, if $\hat{\Theta}_n^*$ is an effective estimator of Θ^*, we can expect $\hat{\theta}_n$ to come close to achieving distributional replication.

Remark 3.8. The sequence $\epsilon_1, \epsilon_2, \ldots$ should be thought of as converging to zero very quickly. We would like to choose $\hat{\theta}_n$ such that $p(\hat{\theta}_n(\omega))$ is equal to the infimum of p over $\hat{\Theta}_n^*(\omega)$ for each $\omega \in \Omega$, but in general this is not possible because the set $\hat{\Theta}_n^*(\omega)$ need not be compact. So instead, we choose $\hat{\theta}_n$ such that $p(\hat{\theta}_n(\omega))$ is very close to the infimum of p over $\hat{\Theta}_n^*(\omega)$, with arbitrarily small approximation error ϵ_n. This technical argument relates closely to what Chen (2007, p. 5561) refers to as an approximate sieve extremum estimate. Though λ_n and ϵ_n appear to play similar roles in Theorem 3.2, from a more substantive perspective we wish ϵ_n to be as small as possible, while λ_n plays a more involved role in the asymptotic results to follow, and must be chosen to converge to zero at a suitable rate.

Remark 3.9. If there is no relevant notion of "price" over the space of functions Θ, we may simply take p to be constant over Θ. In this case, the sequence $\epsilon_1, \epsilon_2, \ldots$ and the function p play no role in Theorem 3.2. Instead, Theorem 3.2 merely asserts the existence of a measurable function $\hat{\theta}_n$ from $(\Omega, \mathcal{F})$ to $(\Theta_n, \mathcal{B}(\Theta_n))$ that satisfies $\hat{\theta}_n(\omega) \in \hat{\Theta}_n^*(\omega)$ for each $\omega \in \Omega$.

It remains to show that our estimator $\hat{\theta}_n$ has desirable asymptotic properties. To ensure that $\hat{\theta}_n$ is well-behaved, the rate at which the sieve space Θ_n expands, and at which the tuning parameter λ_n decays, must be suitably controlled. The following assumption provides a sufficient condition of this kind.

Assumption 3.4. As $n \to \infty$, we have $\lambda_n \to 0$, $n^{-1}\lambda_n^{-2}\mathcal{V}(\Theta_n) \to 0$ and $\lambda_n^{-1} \inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger) \to 0$ for each $\theta^\dagger \in \Theta^\dagger$, where $\Theta^\dagger$ is some dense subset of Θ^* under d.

Remark 3.10. The requirement that $n^{-1}\lambda_n^{-2}\mathcal{V}(\Theta_n) \to 0$ and $\lambda_n^{-1} \inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger) \to 0$ for each $\theta^\dagger \in \Theta^\dagger$ places opposing constraints on the rate of expansion of Θ_n as $n \to \infty$. The complexity of Θ_n must increase sufficiently fast for the sieve approximation error $\inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger)$ to tend to zero faster than λ_n for each $\theta^\dagger \in \Theta^\dagger$, but not so fast that $\mathcal{V}(\Theta_n)$ increases faster than $n\lambda_n^2$. On the other hand, the rate of decay of λ_n may be arbitrarily slow, provided that $\lambda_n \to 0$.

Our final result of this section indicates that, when the above assumptions are satisfied, in large samples we can expect our estimated function to be close to achieving distributional replication, and close to achieving the minimum cost among replicators. We first require some additional notation. Let $P^{\hat{\theta}_n}$ be the probability measure on $\mathcal{B}(\Theta_n)$ given by $P^{\hat{\theta}_n}B = P\hat{\theta}_n^{-1}B$ for each $B \in \mathcal{B}(\Theta_n)$, and let $P^{\hat{\theta}_n(X)}$ be the probability measure on $\mathcal{B}(\mathbb{R})$ given by $P^{\hat{\theta}_n(X)}B = P^{\hat{\theta}_n} \otimes P^X\{(\theta, x) \in \Theta_n \times R^X : \theta(x) \in B\}$ for each $B \in \mathcal{B}(\mathbb{R})$. Note that for $P^{\hat{\theta}_n(X)}$ to be well defined we need $\mathcal{B}(\Theta_n)$ to be an admissible structure for Θ_n; this condition was given

in Assumption 3.2. We can think of $P^{\hat{\theta}_n(X)}$ as the probability distribution of $\hat{\theta}_n(X)$ when $\hat{\theta}_n$ and X are distributed independently of one another.

Theorem 3.3. *As $n \to \infty$, $P^{\hat{\theta}_n(X)} \Rightarrow P^Y$ and $P\{\omega : p(\hat{\theta}_n(\omega)) > \inf_{\theta \in \Theta^*} p(\theta) + \varepsilon\} \to 0$ for any $\varepsilon > 0$.*

Remark 3.11. Theorem 3.3 indicates that $\hat{\theta}_n$ can be expected to perform well with respect to the dual goals of distributional replication and cost minimization in large samples. This duality complicates any discussion of the optimal selection of the tuning parameter λ_n. When λ_n is large, we include functions in our estimated set $\hat{\Theta}_n^*$ for which the empirical evidence for distributional replication is weaker, but we also minimize the function p over a larger set. In applications, the best choice of λ_n would depend on an individual's relative preference for distributional replication, quantified by $M(\theta)$, and cost minimization, quantified by $p(\theta)$.

4 Distributional Replication Using Options

In this section we consider the problem of choosing a portfolio of options on some financial asset such that the payoff from our portfolio after a specified period of time has approximately the same statistical distribution as the payoff from a \$1 investment in some other asset over the same time period. We would like to find the cheapest portfolio of options such that distributional replication is achieved; in particular, we would like the cost of the portfolio to be \$1 or less. We will show how this problem of portfolio selection can be interpreted and solved using the mathematical and statistical machinery developed in the previous two sections.

We suppose that the random variables X and Y represent the dollar denominated payoffs after one period from a \$1 investment in each of two assets. The asset with payoff X will be referred to as the base asset, and the asset with payoff Y will be referred to as the target asset. The price of a one share investment in either asset is taken to be \$1. We assume that X and Y are nonnegative and may be arbitrarily large with nonzero probability, so that $R^X = R^Y = [0, \infty)$ under Assumption 2.1. We may thus replace Assumption 2.1 with the following, more restrictive condition.

Assumption 4.1. F^X and F^Y are continuous and strictly increasing on $[0, \infty)$, and zero on $(-\infty, 0]$.

We find the payoff distribution of the target asset to be desirable, but we seek to achieve this distribution by investing in a portfolio composed of the base asset itself and a basket of European put and call options written on the base asset, with the options expiring after one period. The payoff of such a portfolio after one period is a deterministic function of X; for instance, the payoff from a European call option with strike price s after one period is given by $\max\{0, X - s\}$, while the payoff from a European put option with strike price s after one period is given by $\max\{0, s - X\}$. We also allow our portfolio to include an investment in

risk-free zero-coupon bonds with \$1 par value, expiring after one period. The payoff from such a bond after one period is simply \$1. We allow our portfolio to include long or short positions in each of the component assets, but the payoff from the complete portfolio must be nonnegative.

The payoff from a portfolio of options and bonds after one period is a deterministic function of X. Thus, we can think of a portfolio as a function $\theta \in \Theta$, and write the payoff from the portfolio as $\theta(X)$. Suppose our portfolio includes options at m different strike prices $s_1, \ldots, s_m$, with $0 < s_1 < \cdots < s_m < \infty$. Without loss of generality, we may consider all options to be call options, since the payoff function for a put option with strike price s_i can be replicated by selling one share of the base asset, purchasing a call option with strike price s_i, and purchasing s_i zero-coupon bonds. Suppose we form a portfolio by purchasing β_1 bonds, β_2 shares in the base asset, and β_{i+2} call options at strike price s_i, with $i = 1, \ldots, m$. The payoff function corresponding to our portfolio is then given by

$$\theta(x; \beta, s) = \beta_1 + \beta_2 x + \sum_{i=1}^{m} \beta_{i+2} \max\{0, x - s_i\},$$

where $x \in [0, \infty)$. For fixed $s = (s_1, \ldots, s_m)$, the collection of functions $\{\theta(\cdot; \beta, s) : \beta \in \mathbb{R}^{m+2}\}$ consists of all the continuous functions from $[0, \infty)$ to $\mathbb{R}$ that are linear on each of the $m + 1$ subintervals $(0, s_1), (s_1, s_2), \ldots, (s_m, \infty)$. To ensure that the payoff from our portfolio is nonnegative, we require that β lies in a suitable subset of $\mathbb{R}^{m+2}$. We will let $\Psi_m(s)$ denote the collection of all continuous functions from $[0, \infty)$ to $[0, \infty)$ that are linear on each of the $m + 1$ subintervals $(0, s_1), (s_1, s_2), \ldots, (s_m, \infty)$, and let $\mathcal{B}(\Psi_m(s))$ denote the Borel σ-field on $\Psi_m(s)$ generated by d.

Theorem 4.1. *For fixed $s \in \mathbb{R}^m$ with $0 < s_1 < \cdots < s_m$, we have (i) $\mathcal{V}(\Psi_m(s)) = m + 3$; (ii) $(\Psi_m(s), d)$ is a Polish metric space; and (iii) $\mathcal{B}(\Psi_m(s))$ is an admissible structure for $\Psi_m(s)$.*

We can see from Theorem 4.1 and Remark 3.4 that $\Psi_m(s)$ satisfies the conditions placed on Θ_n in Assumption 3.2. The main idea behind the application discussed in this section is that $\Psi_m(s)$, the space of nonnegative payoff functions achievable using strike prices s, can be used to play the role of the sieve space Θ_n described in the previous section. We obtain an expanding sequence of sieve spaces by assuming that the collection of strike prices s varies with the sample size n, becoming more dense (in a sense soon to be made precise) as n increases. Suppose that $m_1, m_2, \ldots$ is a nondecreasing sequence of natural numbers with $m_n \to \infty$ and $m_n/n \to 0$ as $n \to \infty$. Let $\{s_{i,n} : i = 1, \ldots, m_n; n \in \mathbb{N}\}$ be a triangular array of positive real numbers satisfying (i) $0 < s_{1,n} < \cdots < s_{m_n,n}$ for each $n \in \mathbb{N}$, and (ii) $\{s_{1,n}, \ldots, s_{m_n,n}\} \subseteq \{s_{1,n+1}, \ldots, s_{m_{n+1},n+1}\}$ for each $n \in \mathbb{N}$. We define our expanding sequence of sieve spaces by setting $\Theta_n = \Psi_{m_n}(s_{1,n}, \ldots, s_{m_n,n})$. Theorem 4.1 implies that this choice of Θ_n satisfies Assumption 3.2, with $\mathcal{V}(\Theta_n) = m_n + 3$.

In the context of the present application, the function p introduced in the previous section describes, quite literally, the price of each payoff function $\theta \in \Theta$. For a payoff function $\theta \in \Theta_n$, we can calculate the price $p(\theta)$ directly from the prices of bonds and options. Consider the function $\theta(x; \beta, s) = \beta_1 + \beta_2 x + \sum_{i=1}^{m} \beta_{i+2} \max\{0, x - s_i\}$ defined earlier. Let p_1 denote the price of a bond, p_2 denote the price of a share in the base asset, and p_{i+2} denote the price of a call option with strike price s_i, where $i = 1, \ldots, m$. Note that $p_2 = 1$ by assumption. The price of $\theta(\cdot; \beta, s)$ is simply $\sum_{i=1}^{m+2} p_i \beta_i$. In this way we can calculate $p(\theta)$ for any $\theta \in \Theta_n$, provided we observe the bond price p_1 and the prices of call options at strike prices $s_{1,n}, \ldots, s_{m_n,n}$.

Assumption 3.4 imposes a condition on the rate of decay of the sieve approximation error: we require that $\lambda_n^{-1} \inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger) \to 0$ for each $\theta^\dagger \in \Theta^\dagger$, where $\Theta^\dagger$ is some dense subset of Θ^* under d. The following result shows how $\Theta^\dagger$ may be chosen such that this condition is satisfied when our sieve space corresponds to portfolios of options.

Theorem 4.2. *Let $\Theta^\dagger$ denote the set of all functions $\theta^\dagger \in \Theta^*$ such that $F^Y \circ \theta^\dagger \circ Q^X$ is Lipschitz continuous. Then $\Theta^\dagger$ is dense in Θ^* under d. Also, when $\Theta_n = \Psi_{m_n}(s_{1,n}, \ldots, s_{m_n,n})$, as $n \to \infty$ we have*

$$\inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger) = O\left(m_n \sup_{0 \leq i \leq m_n} P^X(s_{i,n}, s_{i+1,n})^2 \right)$$

for each $\theta^\dagger \in \Theta^\dagger$, where $s_{0,n} = 0$ and $s_{m_n+1,n} = \infty$.

Theorem 4.2 reveals that our sequence of sieve spaces constructed using option payoffs can approximate replicating functions satisfying a deformed Lipschitz condition, provided that $\sup_{0 \leq i \leq m_n} P^X(s_{i,n}, s_{i+1,n})$ decays to zero at a suitable rate. Further, that set of deformed Lipschitz continuous replicating functions is dense in the set of all replicating functions. If we could choose our strike prices such that $P^X(s_{i,n}, s_{i+1,n})$ was constant across $i = 0, \ldots, m_n$, we would have $\inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger) = O(m_n^{-1})$ for each $\theta^\dagger \in \Theta^\dagger$.

Theorem 4.2 and part (i) of Theorem 4.1 show how the choice of strike prices is constrained by Assumption 3.4. Specifically, the conditions on Θ_n imposed by Assumption 3.4 may be rewritten as follows: $n^{-1}\lambda_n^{-2} m_n \to 0$ and $\lambda_n^{-1} m_n \sup_{0 \leq i \leq m_n} P^X(s_{i,n}, s_{i+1,n})^2 \to 0$ as $n \to \infty$. If our strike prices are chosen such that $P^X(s_{i,n}, s_{i+1,n})$ is constant across $i = 0, \ldots, m_n$, Assumption 3.4 will be satisfied provided that $\lambda_n = o(1)$, $m_n = o(n\lambda_n^2)$ and $m_n^{-1} = o(\lambda_n)$. For instance, we could choose $m_n \sim n^a$ and $\lambda_n \sim n^{-b}$, with $0 < b < a < 1 - 2b$. As noted in Remark 3.11, it is difficult to see how an optimal choice of m_n and λ_n could be made in practice, because the two parameters may have different effects on the twin criterion functions $M(\theta)$ and $p(\theta)$, and one's relative preference for optimizing with respect to those two functions may be subjective. It is perhaps best to experiment with a range of different values for m_n and λ_n. Further, the choice of strike prices may be constrained by the strike prices being actively traded on the market.

5 Conclusion

In this paper we have developed a mathematical framework for considering the estimation of a function θ such that $\theta(X)$ has the same distribution as Y. We have discussed the relevance of our results to financial applications in which one seeks to find the cheapest way to achieve a desired payoff distribution by trading liquid assets. We now briefly discuss two possible extensions of our results that may prove fruitful.

In terms of the relevance of our technical conditions in financial applications, the elephant in the room is clearly Assumption 3.1, which imposes an iid condition on the random variables (X_i, Y_i), $i = 1, \ldots, n$. It is certainly the case that time series of financial returns typically do not behave as though they were distributed independently over time, as is clear from the voluminous literature on stochastic volatility. The iid condition comes into play in the proof of Theorem 3.1, in which results in empirical process theory are used to establish a uniform bound on the error in the approximation of M by M_n over our sieve space Θ_n. The results we apply are based on iid conditions, but generalizations suitable for dependent data have been developed by Doukhan, Massart and Rio (1995) and Rio (1998, 2000). It seems likely that, with some strengthening of the rate conditions in Assumption 3.4, the results in this paper could be adapted to allow for dependent data. However, by allowing for the possibility of serial dependence, a further question is raised. The methods we have proposed are designed such that the unconditional distribution of $\theta(X)$ is approximately equal to the unconditional distribution of Y. If data are serially dependent, the more relevant objects may be distributions that are conditional on past information. Though we acknowledge the importance of this issue, it goes beyond the scope of this paper; and, indeed, beyond the scope of the financial literature on hedge fund replication to date.

A second potential extension of our results would be to consider the replication of multivariate distributions. As discussed in the introduction, Kat and Palaro (2005ab) consider estimating a transformation θ of a pair of random variables X and Z such that X and $\theta(X, Z)$ have the same joint distribution as X and Y. The difficulty with adapting our own method to this approach is that the class of bivariate functions that can be approximated by portfolios of options written on individual assets is rather small. One possible solution would be to consider portfolios formed from derivative securities that are written on multiple underlying assets; another would be to forgo exact distributional replication, and seek the closest distributional match from a smaller class of multivariate payoff functions that is approximable using portfolios of simple options. We leave these possibilities for future research.

A Mathematical Appendix

Proof of Theorem 2.1. Choose a point $c \in (0,1)$, and let $\tilde{\theta}^c : [0,1] \to (0,1)$ be defined by $\tilde{\theta}^c(u) = u/c$ for $u \in (0,c)$, $\tilde{\theta}^c(u) = 1-(u-c)/(1-c)$ for $u \in (c,1)$, and $\tilde{\theta}^c(u) = \frac{1}{2}$ for $u \in \{0,c,1\}$. Let $\theta^c : R^X \to R^Y$ be defined by $\theta^c(x) = Q^Y \circ \tilde{\theta}^c \circ F^X(x)$. F^X is continuous under Assumption 2.1, so $F^X(X) \sim U(0,1)$. Hence, for any $a \in (0,1)$, we have

$$P^X\{x : \tilde{\theta}^c \circ F^X(x) \leq a\} = \int_0^1 1\{\tilde{\theta}^c(u) \leq a\}\mathrm{d}u = \int_0^1 1\{u \leq ac \text{ or } u \geq 1-a+ac\}\mathrm{d}u = a,$$

implying that $\tilde{\theta}^c \circ F^X(X) \sim U(0,1)$. Therefore, $\theta^c(X) \sim Y$, and so θ^c is a replicating function for X and Y. If we choose $c_0, c_1 \in (0,1)$ with $c_0 \neq c_1$, then $\tilde{\theta}^{c_0}(u) \neq \tilde{\theta}^{c_1}(u)$ for $u \notin \{0,c_0,c_1,1\}$. Since Q^Y is strictly increasing under Assumption 2.1, it follows that $\theta^{c_0}(x) \neq \theta^{c_1}(x)$ for $x \notin \{Q^X(0), Q^X(c_0), Q^X(c_1), Q^X(1)\}$, and so continuity of F^X implies that $\theta^{c_0}(x) \neq \theta^{c_1}(x)$ for all x in a set of P^X-measure one. Thus, by allowing c to vary over $(0,1)$, we obtain an uncountable collection of replicating functions, no two of which are equal on a set of positive P^X-measure. □

Proof of Theorem 2.2. For $\theta_0, \theta_1 \in \Theta$ we can use the triangle inequality to show that

$$|M(\theta_1) - M(\theta_0)| \leq \int_{R^Y} \left|F^X(y;\theta_1) - F^X(y;\theta_0)\right| \mathrm{d}F^Y(y).$$

For each $y \in R^Y$ we have

$$\begin{aligned} \left|F^X(y;\theta_1) - F^X(y;\theta_0)\right| &= \left|P^X\{x : \theta_1(x) \leq y\} - P^X\{x : \theta_0(x) \leq y\}\right| \\ &= \left|P^X\{x : \theta_1(x) \leq y < \theta_0(x)\} - P^X\{x : \theta_0(x) \leq y < \theta_1(x)\}\right|, \end{aligned}$$

and so applying the triangle inequality again we obtain

$$|M(\theta_1) - M(\theta_0)| \leq \int_{R^Y} P^X\{x : \theta_1(x) \leq y < \theta_0(x)\}\mathrm{d}F^Y(y) + \int_{R^Y} P^X\{x : \theta_0(x) \leq y < \theta_1(x)\}\mathrm{d}F^Y(y).$$

Tonelli's theorem implies that

$$\int_{R^Y} P^X\{x : \theta_1(x) \leq y < \theta_0(x)\}\mathrm{d}F^Y(y) = \int_{R^X}\int_{R^Y} 1\{\theta_1(x) \leq y < \theta_0(x)\}\mathrm{d}F^Y(y)\mathrm{d}F^X(x).$$

F^Y is continuous under Assumption 2.1, and so we have

$$\int_{R^Y} 1\{\theta_1(x) \leq y < \theta_0(x)\} \mathrm{d}F^Y(y) = \max\{F^Y(\theta_0(x)) - F^Y(\theta_1(x)), 0\}$$

for each $x \in R^X$. Therefore,

$$\int_{R^Y} P^X\{x : \theta_1(x) \leq y < \theta_0(x)\} \mathrm{d}F^Y(y) = \int_{R^X} \max\{F^Y(\theta_0(x)) - F^Y(\theta_1(x)), 0\} \mathrm{d}F^X(x).$$

Similarly, we have

$$\int_{R^Y} P^X\{x : \theta_0(x) \leq y < \theta_1(x)\} \mathrm{d}F^Y(y) = \int_{R^X} \max\{F^Y(\theta_1(x)) - F^Y(\theta_0(x)), 0\} \mathrm{d}F^X(x),$$

and so

$$|M(\theta_1) - M(\theta_0)| \leq \int_{R^X} \left|F^Y(\theta_1(x)) - F^Y(\theta_0(x))\right| \mathrm{d}F^X(x) = d(\theta_0, \theta_1).$$

□

Proof of Theorem 2.3. It is obvious that $M(\theta) = 0$ if $\theta \in \Theta^*$. We will prove the reverse implication. Suppose $M(\theta) = 0$. Then $F^X(y;\theta) = F^Y(y)$ for all y in a set of P^Y-measure one. Suppose $F^X(c_0;\theta) \neq F^Y(c_0)$ for some $c_0 \in \mathbb{R}$. Right continuity of $F^X(\cdot;\theta)$ and F^Y ensures that $F^X(y;\theta) \neq F^Y(y)$ for all y in some open interval (c_0, c_1). Since F^Y is strictly increasing under Assumption 2.1, (c_0, c_1) must have strictly positive P^Y-measure, leading to a contradiction. Thus it must be the case that $F^X(y;\theta) = F^Y(y)$ for all $y \in \mathbb{R}$, implying that $\theta \in \Theta^*$. □

Proof of Theorem 2.4. If $F^X(y;\theta_n) \to F^Y(y)$ for each $y \in \mathbb{R}$, then $M(\theta_n) \to 0$ by dominated convergence. Suppose $F^X(c_0;\theta_n) \nrightarrow F^Y(c_0)$ for some $c_0 \in \mathbb{R}$. Then there must be an increasing sequence of natural numbers $n_1, n_2, \ldots$ and a real number $\varepsilon > 0$ (or $\varepsilon < 0$) such that $F^X(c_0;\theta_{n_k}) \geq F^Y(c_0) + \varepsilon$ (resp. $F^X(c_0;\theta_{n_k}) \leq F^Y(c_0) + \varepsilon$) for all k sufficiently large. Suppose $\varepsilon > 0$. F^Y is continuous under Assumption 2.1, so we may choose $c_1 > c_0$ such that $F^Y(c_1) = F^Y(c_0) + \varepsilon/2$. Monotonicity of $F^X(\cdot;\theta_{n_k})$ and F^Y then ensures that $F^X(y;\theta_{n_k}) \geq F^Y(y) + \varepsilon/2$ for all $y \in [c_0, c_1]$, for all k sufficiently large. Consequently, we have

$$\int |F^X(y;\theta_{n_k}) - F^Y(y)| \mathrm{d}F^Y(y) \geq \frac{\varepsilon}{2} \int_{c_0}^{c_1} \mathrm{d}F^Y(y) = \frac{\varepsilon}{2}(F^Y(c_1) - F^Y(c_0)) = \frac{\varepsilon^2}{4} > 0$$

for all k sufficiently large, implying that $M(\theta_n) \nrightarrow 0$. □

Proof of Theorem 2.5. Since $(\theta, x) \mapsto \theta(x)$ is $\mathcal{T}_n \otimes \mathcal{B}(R^X)$-measurable, $(\theta, x, y) \mapsto 1\{\theta(x) \leq y\}$ is $\mathcal{T}_n \otimes$

$\mathcal{B}(R^X)\otimes\mathcal{B}(\mathbb{R})$-measurable. Tonelli's theorem thus implies that $(\theta, y) \mapsto \int_{R^X} 1\{\theta(x) \leq y\} \mathrm{d}P^X(x) = F^X(y;\theta)$ is $\mathcal{T}_n \otimes \mathcal{B}(\mathbb{R})$-measurable, justifying the following interchange of integrals:

$$\int_{\Theta_n} M \mathrm{d}P^{\theta_n} = \int_{\Theta_n} \int_{R^Y} |F^X(y;\theta) - F^Y(y)| \mathrm{d}F^Y(y) \mathrm{d}P^{\theta_n}(\theta) = \int_{R^Y} \int_{\Theta_n} |F^X(y;\theta) - F^Y(y)| \mathrm{d}P^{\theta_n}(\theta) \mathrm{d}F^Y(y).$$

We thus have

$$\int_{\Theta_n} M \mathrm{d}P^{\theta_n} \geq \int_{R^Y} \left| \int_{\Theta_n} F^X(y;\theta) \mathrm{d}P^{\theta_n}(\theta) - F^Y(y) \right| \mathrm{d}F^Y(y).$$

Again using the $\mathcal{T}_n \otimes \mathcal{B}(\mathbb{R})$-measurability of $(\theta, y) \mapsto F^X(y;\theta)$, Tonelli's theorem implies that

$$\int_{\Theta_n} F^X(y;\theta) \mathrm{d}P^{\theta_n}(\theta) = \int_{\Theta_n} \int_{R^X} 1\{\theta(x) \leq y\} \mathrm{d}P^X(x) \mathrm{d}P^{\theta_n}(\theta) = P^{\theta_n(X)}(-\infty, y]$$

for each $y \in \mathbb{R}$. Letting $F^{\theta_n(X)}$ denote the cdf corresponding to $P^{\theta_n(X)}$, we now have

$$\int_{\Theta_n} M \mathrm{d}P^{\theta_n} \geq \int_{R^Y} \left| F^{\theta_n(X)}(y) - F^Y(y) \right| \mathrm{d}F^Y(y).$$

Arguing as we did in the proof of Theorem 2.4, we can show that $\int_{R^Y} \left| F^{\theta_n(X)}(y) - F^Y(y) \right| \mathrm{d}F^Y(y) \to 0$ if and only if $F^{\theta_n(X)}(y) \to F^Y(y)$ for each $y \in \mathbb{R}$. Thus, $\int_{\Theta_n} M \mathrm{d}P^{\theta_n} \to 0$ implies $F^{\theta_n(X)} \to F^Y$ pointwise, which is equivalent to $P^{\theta_n(X)} \Rightarrow P^Y$. □

Proof of Theorem 3.1. Elementary arguments can be used to show that, for all $\theta \in \Theta$,

$$\begin{aligned} |M_n(\theta) - M(\theta)| \quad \leq \quad & \int \left| F_n^Y(y) - F^Y(y) \right| \mathrm{d}F_n^Y(y) + \int \left| F_n^X(y;\theta) - F^X(y;\theta) \right| \mathrm{d}F_n^Y(y) \\ & + \left| \int \left| F^X(y;\theta) - F^Y(y) \right| \mathrm{d}F_n^Y(y) - \int \left| F^X(y;\theta) - F^Y(y) \right| \mathrm{d}F^Y(y) \right|. \end{aligned} \tag{A.1}$$

We will establish a uniform bound on the order of each of the three terms on the right-hand side of (A.1). These bounds will be expressed in terms of the outer expectation operator E^*, denoting outer integration of nonnegative functions defined on the underlying probability space $(\Omega, \mathcal{F}, P)$; see e.g. section 1.2 in van der Vaart and Wellner (1996). Obtaining a bound for the first term is simple as it does not depend on θ: Donsker's theorem yields

$$E^* \int \left| F_n^Y(y) - F^Y(y) \right| \mathrm{d}F_n^Y(y) \leq E^* \sup_{y \in \mathbb{R}} \left| F_n^Y(y) - F^Y(y) \right| = O(n^{-1/2}).$$

For the second term on the right-hand side of (A.1), we have

$$E^* \sup_{\theta\in\Theta_n} \int \left|F_n^X(y;\theta) - F^X(y;\theta)\right| \mathrm{d}F_n^Y(y) \leq E^* \sup_{\theta\in\Theta_n} \sup_{y\in\mathbb{R}} \left|F_n^X(y;\theta) - F^X(y;\theta)\right| = E^* \sup_{g\in\mathcal{G}_n} |P_n^X g - P^X g|,$$

where $\mathcal{G}_n$ is the class of indicator functions of sets of the form $\{x \in R^X : \theta(x) \leq y\}$ with $\theta \in \Theta_n$ and $y \in \mathbb{R}$. Note that $\mathcal{G}_n$ is the collection of indicators of all complements of sets majorized by Θ_n. Since Θ_n is VC-major with dimension $\mathcal{V}(\Theta_n)$, $\mathcal{G}_n$ must be VC-subgraph with dimension $\mathcal{V}(\Theta_n)$. Hence Theorem 2.6.7 in van der Vaart and Wellner (1996) implies that, for any $\varepsilon \in (0,1)$ and any probability measure Q on $\mathcal{B}(\mathbb{R})$, there exists $K < \infty$ such that $N(\varepsilon, \mathcal{G}_n, L_2(Q)) \leq K\mathcal{V}(\Theta_n)(16\mathrm{e})^{\mathcal{V}(\Theta_n)}\varepsilon^{-2(\mathcal{V}(\Theta_n)-1)}$. Theorem 2.14.1 in the same book thus gives $E^* \sup_{g\in\mathcal{G}_n} |P_n^X g - P^X g| = O(\sqrt{\mathcal{V}(\Theta_n)/n})$, implying that

$$E^* \sup_{\theta\in\Theta_n} \int \left|F_n^X(y;\theta) - F^X(y;\theta)\right| \mathrm{d}F_n^Y(y) = O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right).$$

For the third term on the right-hand side of (A.1), we have

$$E^* \sup_{\theta\in\Theta_n} \left| \int \left|F^X(y;\theta) - F^Y(y)\right| \mathrm{d}F_n^Y(y) - \int \left|F^X(y;\theta) - F^Y(y)\right| \mathrm{d}F^Y(y) \right| = E^* \sup_{h\in\mathcal{H}_n} |P_n^Y h - P^Y h|,$$

where $\mathcal{H}_n$ is the class of functions $\{|F^X(\cdot;\theta) - F^Y(\cdot)| : \theta \in \Theta_n\}$. Consider the simpler class of functions $\mathcal{H}_n^0 = \{F^X(\cdot;\theta) : \theta \in \Theta_n\}$. Since $\mathcal{H}_n^0$ is a subset of the collection of monotone increasing functions from $\mathbb{R}$ to $[0,1]$, Theorem 2.7.5 in van der Vaart and Wellner (1996) implies the existence of $K < \infty$ such that $N_{[]}(\varepsilon, \mathcal{H}_n^0, L_2(P^Y)) \leq K\varepsilon^{-1}$ for all $\varepsilon \in (0,1)$. It is straightforward to show that $N_{[]}(\varepsilon, \mathcal{H}_n, L_2(P^Y)) \leq N_{[]}(\varepsilon, \mathcal{H}_n^0, L_2(P^Y))$, and so Theorem 2.14.2 in van der Vaart and Wellner (1996) gives $E^* \sup_{h\in\mathcal{H}_n} |P_n^Y h - P^Y h| = O(n^{-1/2})$, implying that

$$E^* \sup_{\theta\in\Theta_n} \left| \int \left|F^X(y;\theta) - F^Y(y)\right| \mathrm{d}F_n^Y(y) - \int \left|F^X(y;\theta) - F^Y(y)\right| \mathrm{d}F^Y(y) \right| = O(n^{-1/2}).$$

Collecting together these bounds on the order of the terms on the right-hand side of (A.1), we obtain

$$E^* \sup_{\theta\in\Theta_n} |M_n(\theta) - M(\theta)| = O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right).$$

It remains only to show that $\omega \mapsto \sup_{\theta\in\Theta_n} |M_n(\omega,\theta) - M(\theta)|$ is $\mathcal{F}$-measurable. Corollary 5.3.5 of Dudley (1999) implies that $\omega \mapsto \sup_{\theta\in\Theta_n} |M_n(\omega,\theta) - M(\theta)|$ is universally $\mathcal{F}$-measurable if $(\Theta_n, \mathcal{B}(\Theta_n))$ is a Souslin measurable space and $(\omega,\theta) \mapsto |M_n(\omega,\theta) - M(\theta)|$ is $\mathcal{F}\otimes\mathcal{B}(\Theta_n)$-measurable. Since $(\Omega, \mathcal{F}, P)$ is complete under Assumption 3.1, universal $\mathcal{F}$-measurability implies $\mathcal{F}$-measurability. Assumption 3.2 states that $(\Theta_n, \mathcal{B}(\Theta_n))$

is a Souslin measurable space, and $\theta \mapsto M(\theta)$ is continuous and hence $\mathcal{B}(\Theta_n)$-measurable by Theorem 2.3, so it suffices for us to show that $(\omega, \theta) \mapsto M_n(\omega, \theta)$ is $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable. Observe that

$$M_n(\omega, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{n} \sum_{j=1}^{n} 1\{\theta(X_j(\omega)) \leq Y_i(\omega)\} - \frac{1}{n} \sum_{j=1}^{n} 1\{Y_j(\omega) \leq Y_i(\omega)\} \right|. \tag{A.2}$$

It should be clear from (A.2) that $(\omega, \theta) \mapsto M_n(\omega, \theta)$ will be $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable if $(\omega, \theta) \mapsto \theta(X_j(\omega))$ is $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable for any j. This condition is satisfied for each j since $\mathcal{B}(\Theta_n)$ is an admissible structure for Θ_n under Assumption 3.2, and each X_j is $\mathcal{F}$-measurable. □

Proof of Theorem 3.2. We showed in the proof of Theorem 3.1 that $(\omega, \theta) \mapsto M_n(\omega, \theta)$ is $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable. Since $(\Theta_n, \mathcal{B}(\Theta_n))$ is a Souslin measurable space, it follows from Corollary 5.3.5 of Dudley (1999) that $\omega \mapsto \inf_{\theta \in \Theta_n} M_n(\omega, \theta)$ is $\mathcal{F}$-measurable. Consequently, we have

$$\text{gr}(\hat{\Theta}_m^*) = \left\{ (\omega, \theta) \in \Omega \times \Theta_n : M_n(\omega, \theta) \leq \inf_{\vartheta \in \Theta_n} M_n(\omega, \vartheta) + \lambda_n \right\} \in \mathcal{F} \otimes \mathcal{B}(\Theta_n),$$

where $\text{gr}(\hat{\Theta}_m^*)$ denotes the graph of $\hat{\Theta}_m^*$. Therefore, since p is continuous and hence $\mathcal{B}(\Theta_n)$-measurable under Assumption 3.3, Theorem 2.17(a) of Stinchcombe and White (1992) implies that $\omega \mapsto \inf_{\theta \in \hat{\Theta}_m^*(\omega)} p(\theta)$ is $\mathcal{F}$-analytic, and hence $\mathcal{F}$-measurable given that $(\Omega, \mathcal{F}, P)$ is complete under Assumption 3.1. Let $H_n : \Omega \times \Theta_n \to \mathbb{R}$ be the function defined by

$$H_n(\omega, \theta) = \max \left\{ M_n(\omega, \theta) - \inf_{\vartheta \in \Theta_n} M_n(\omega, \vartheta) - \lambda_n,\ p(\theta) - \inf_{\vartheta \in \hat{\Theta}_n^*(\omega)} p(\vartheta) - \epsilon_n \right\}.$$

We have shown that $(\omega, \theta) \mapsto M_n(\omega, \theta)$ is $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable, $\omega \mapsto \inf_{\theta \in \Theta_n} M_n(\omega, \theta)$ is $\mathcal{F}$-measurable, $\theta \mapsto p(\theta)$ is $\mathcal{B}(\Theta_n)$-measurable, and $\omega \mapsto \inf_{\theta \in \hat{\Theta}_m^*(\omega)} p(\theta)$ is $\mathcal{F}$-measurable, so it must be the case that $(\omega, \theta) \mapsto H_n(\omega, \theta)$ is $\mathcal{F} \otimes \mathcal{B}(\Theta_n)$-measurable. Observe that, for all $\omega \in \Omega$, there exists $\theta \in \Theta_n$ such that $H_n(\omega, \theta) \leq 0$; here we use the fact that M_n and p are bounded from below. The Sainte-Beuve selection theorem (see Theorem 5.3.2 in Dudley, 1999) thus implies the existence of a measurable function $\hat{\theta}_n$ from $(\Omega, \mathcal{F})$ to $(\Theta_n, \mathcal{B}(\Theta_n))$ such that $H_n(\omega, \hat{\theta}_n(\omega)) \leq 0$ for all $\omega \in \Omega$. Clearly, $\hat{\theta}_n$ must therefore satisfy $\hat{\theta}_n(\omega) \in \hat{\Theta}_n^*(\omega)$ and $p(\hat{\theta}_n(\omega)) \leq \inf_{\theta \in \hat{\Theta}_n^*(\omega)} p(\theta) + \epsilon_n$ for all $\omega \in \Omega$, as required. □

Proof of Theorem 3.3. We first show that $P^{\hat{\theta}_n(X)} \Rightarrow P^Y$ as $n \to \infty$. For each $\omega \in \Omega$, we have $\hat{\theta}_n(\omega) \in \Theta_n$ and $\hat{\theta}_n(\omega) \in \hat{\Theta}_n^*(\omega)$. These two inclusions justify the following two inequalities respectively:

$$M(\hat{\theta}_n(\omega)) \leq M_n(\omega, \hat{\theta}_n(\omega)) + \sup_{\theta \in \Theta_n} |M_n(\omega, \theta) - M(\theta)| \leq \inf_{\theta \in \Theta_n} M_n(\omega, \theta) + \lambda_n + \sup_{\theta \in \Theta_n} |M_n(\omega, \theta) - M(\theta)|.$$

Since $M_n(\omega,\theta) \leq M(\theta) + \sup_{\vartheta\in\Theta_n} |M_n(\omega,\vartheta) - M(\vartheta)|$ for all $\theta \in \Theta_n$, we therefore have

$$M(\hat{\theta}_n(\omega)) \leq \inf_{\theta\in\Theta_n} M(\theta) + \lambda_n + 2 \sup_{\theta\in\Theta_n} |M_n(\omega,\theta) - M(\theta)|.$$

Theorem 2.2 implies that $M(\theta) \leq M(\vartheta) + d(\theta,\vartheta)$ for any $\theta,\vartheta \in \Theta$. Theorem 2.3 implies that $M(\theta^\dagger) = 0$ for any arbitrary $\theta^\dagger \in \Theta^\dagger$, so we have $M(\theta) \leq d(\theta,\theta^\dagger)$. Consequently, $\inf_{\theta\in\Theta_n} M(\theta) \leq \inf_{\theta\in\Theta_n} d(\theta,\theta^\dagger)$, and so

$$M(\hat{\theta}_n(\omega)) \leq \inf_{\theta\in\Theta_n} d(\theta,\theta^\dagger) + \lambda_n + 2 \sup_{\theta\in\Theta_n} |M_n(\omega,\theta) - M(\theta)|.$$

Integrating both sides over Ω, we obtain

$$\int_\Omega M(\hat{\theta}_n(\omega))\mathrm{d}P(\omega) \leq \inf_{\theta\in\Theta_n} d(\theta,\theta^\dagger) + \lambda_n + 2\int_\Omega \sup_{\theta\in\Theta_n} |M_n(\omega,\theta) - M(\theta)|\mathrm{d}P(\omega). \tag{A.3}$$

The first two terms on the right-hand side of (A.3) converge to zero as $n \to \infty$ under Assumption 3.4. The third term is $O(\sqrt{\mathcal{V}(\Theta_n)/n})$ by Theorem 3.1, and therefore must also converge to zero under Assumption 3.4. Thus we have $\int_\Omega M(\hat{\theta}_n(\omega))\mathrm{d}P(\omega) \to 0$ as $n \to \infty$. But $\int_\Omega M(\hat{\theta}_n(\omega))\mathrm{d}P(\omega) = \int_{\Theta_n} M(\theta)\mathrm{d}P^{\hat{\theta}_n}(\theta)$, and so Theorem 2.5 implies that $P^{\hat{\theta}_n(X)} \Rightarrow P^Y$ as $n \to \infty$.

We next show that $P\{\omega : p(\hat{\theta}_n(\omega)) > \inf_{\theta\in\Theta^*} p(\theta) + \varepsilon\} \to 0$ for any $\varepsilon > 0$ as $n \to \infty$. Fix $\varepsilon > 0$, and choose $\theta^\dagger \in \Theta^\dagger$ such that $p(\theta^\dagger) \leq \inf_{\theta\in\Theta^*} p(\theta) + \varepsilon/2$. Choose a sequence $\theta_n \in \Theta_n$, $n \in \mathbb{N}$, such that $d(\theta_n,\theta^\dagger) = O(\inf_{\theta\in\Theta_n} d(\theta,\theta^\dagger))$. Observe that

$$\begin{aligned} P\left\{\omega : p(\hat{\theta}_n(\omega)) > \inf_{\theta\in\Theta^*} p(\theta) + \varepsilon\right\} \quad &\leq \quad P\left\{\omega : p(\hat{\theta}_n(\omega)) > p(\theta^\dagger) + \varepsilon/2 \text{ and } \theta_n \in \hat{\Theta}_n^*(\omega)\right\} \\ &\quad + P\left\{\omega : \theta_n \notin \hat{\Theta}_n^*(\omega)\right\}. \end{aligned} \tag{A.4}$$

We will show that the two terms on the right-hand side of (A.4) tend to zero as $n \to \infty$. First we consider the first term. If $\theta_n \in \hat{\Theta}_n^*(\omega)$, we must have $p(\hat{\theta}_n(\omega)) \leq p(\theta_n) + \epsilon_n$. Therefore,

$$\begin{aligned} P\left\{\omega : p(\hat{\theta}_n(\omega)) > p(\theta^\dagger) + \varepsilon/2 \text{ and } \theta_n \in \hat{\Theta}_n^*(\omega)\right\} \quad &\leq \quad P\left\{\omega : p(\theta_n) + \epsilon_n > p(\theta^\dagger) + \varepsilon/2 \text{ and } \theta_n \in \hat{\Theta}_n^*(\omega)\right\} \\ &\leq \quad P\left\{\omega : p(\theta_n) + \epsilon_n > p(\theta^\dagger) + \varepsilon/2\right\} \\ &= \quad 1\left\{p(\theta_n) + \epsilon_n > p(\theta^\dagger) + \varepsilon/2\right\}. \end{aligned}$$

Assumption 3.3 states that p is continuous. Therefore, since $d(\theta_n,\theta^\dagger) \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$, we must have $1\{p(\theta_n) + \epsilon_n > p(\theta^\dagger) + \varepsilon/2\} = 0$ for all sufficiently large n. Thus the first term on the right-hand side of (A.4) is zero for all sufficiently large n. Next we consider the second term on the right-hand side of (A.4).

We have

$$P\left\{\omega : \theta_n \notin \hat{\Theta}_n^*(\omega)\right\} \leq P\left\{\omega : M_n(\omega, \theta_n) > \lambda_n\right\} \leq \lambda_n^{-1} \int_\Omega M_n(\omega, \theta_n) \mathrm{d}P(\omega),$$

using Markov's inequality to obtain the second inequality. Theorem 3.1 gives

$$\int_\Omega M_n(\omega, \theta_n) \mathrm{d}P(\omega) \leq M(\theta_n) + \int_\Omega \sup_{\theta \in \Theta_n} |M_n(\omega, \theta_n) - M(\theta_n)| \mathrm{d}P(\omega) = M(\theta_n) + O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right).$$

Theorem 2.2 implies that $M(\theta_n) \leq M(\theta^\dagger) + d(\theta_n, \theta^\dagger)$, while Theorem 2.3 implies that $M(\theta^\dagger) = 0$, so we have

$$\int_\Omega M_n(\omega, \theta_n) \mathrm{d}P(\omega) \leq d(\theta_n, \theta^\dagger) + O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right) = O\left(\inf_{\theta \in \Theta_n} d(\theta, \theta^\dagger)\right) + O\left(\sqrt{\mathcal{V}(\Theta_n)/n}\right).$$

It now follows from Assumption 3.4 that the second term on the right-hand side of (A.4) tends to zero as $n \to \infty$. □

Proof of Theorem 4.1. We first prove (i). Let $\Lambda_m(s)$ denote the collection of all continuous functions from $[0, \infty)$ to $\mathbb{R}$ that are linear on each of the $m+1$ subintervals $(0, s_1), (s_1, s_2), \ldots, (s_m, \infty)$. $\Lambda_m(s)$ is an $(m+2)$-dimensional real vector space of functions, and so Theorem 7.2 of Dudley (1978) implies that $\Lambda_m(s)$ is a VC-major class with $\mathcal{V}(\Lambda_m(s)) = m+3$. Since $\Psi_m(s) \subset \Lambda_m(s)$, it must be the case that $\mathcal{V}(\Psi_m(s)) \leq \mathcal{V}(\Lambda_m(s))$. Moreover, given any $\theta_0 \in \Lambda_m(s)$ and any interval $[a, b] \subset \mathbb{R}$, we can find $\theta_1 \in \Psi_m(s)$ and $c \in \mathbb{R}$ such that $\theta_1(x) = \theta_0(x) + c$ for all $x \in [a, b]$. It is easy to see that this implies $\mathcal{V}(\Psi_m(s)) \geq \mathcal{V}(\Lambda_m(s))$. This proves (i).

We next prove (ii). First, observe that d is a metric (rather than merely a pseudometric) on $\Psi_m(s)$, because any two distinct functions $\theta_0, \theta_1 \in \Psi_m(s)$ must differ everywhere on some open interval, which must be of positive P^X-measure under Assumption 4.1. Since F^Y is strictly increasing on $[0, \infty)$ under Assumption 4.1, we will thus have $F^Y \circ \theta_0$ and $F^Y \circ \theta_1$ differing everywhere on the interval in question, forcing $d(\theta_0, \theta_1)$ to be nonzero. It remains to show that the metric space $(\Psi_m(s), d)$ is topologically isomorphic to a complete separable metric space. Each function $\theta \in \Psi_m(s)$ can be written in the form $\theta(x) = \sum_{i=1}^{m+2} \beta_i f_i(x)$ for some unique $\beta \in \tilde{\Psi}_m(s)$, where $f_1(x) = 1$, $f_2(x) = x$, $f_{i+2}(x) = \max\{0, x - s_i\}$ for $i = 1, \ldots, m$, and

$$\tilde{\Psi}_m(s) = \left\{\beta \in \mathbb{R}^{m+2} : \sum_{i=1}^{m+2} \beta_i f_i(x) \geq 0 \text{ for all } x \in R^X\right\}.$$

Similarly, each $\beta \in \tilde{\Psi}_m(s)$ uniquely identifies a function $\theta \in \Psi_m(s)$. We will denote this bijection between $\Psi_m(s)$ and $\tilde{\Psi}_m(s)$ by $\mathcal{S} : \Psi_m(s) \to \tilde{\Psi}_m(s)$. It is easy to see that $\tilde{\Psi}_m(s)$ is a complete separable metric space. We will show that $\mathcal{S}$ defines a topological isomorphism between $(\Psi_m(s), d)$ and $(\tilde{\Psi}_m(s), \tilde{d})$, where $\tilde{d}$ is the usual Euclidean metric on $\tilde{\Psi}_m(s)$. That is, we will show that $\mathcal{S}$ and $\mathcal{S}^{-1}$ are continuous. Suppose $\beta_1, \beta_2, \ldots$

is a sequence in $\tilde{\Psi}_m(s)$ converging to some $\beta^* \in \tilde{\Psi}_m(s)$, and let $\theta^* = \mathcal{S}^{-1}\beta^*$ and $\theta_n = \mathcal{S}^{-1}\beta_n$ for each $n \in \mathbb{N}$. For $x \in [0,\infty)$, Cauchy's inequality gives

$$|\theta_n(x) - \theta^*(x)| = \left|\sum_{i=1}^{m+2}(\beta_{n,i} - \beta_i^*)f_i(x)\right| \leq \tilde{d}(\beta_n,\beta^*)\left(\sum_{i=1}^{m+2} f_i(x)^2\right)^{1/2},$$

and hence θ_n converges to θ^* pointwise. It then follows from dominated convergence that $d(\theta_n,\theta^*) \to 0$, which proves that $\mathcal{S}^{-1}$ is continuous. Suppose now that $\beta_1,\beta_2,\ldots$ does not converge to $\beta^* \in \tilde{\Psi}_m(s)$. Then we can choose a subsequence $\beta_{n_1},\beta_{n_2},\ldots$ and a constant $\varepsilon > 0$ such that $\tilde{d}(\beta_{n_k},\beta^*) > \varepsilon$ for all k. For $x \in R^X$ and all k, we have

$$|\theta_{n_k}(x) - \theta^*(x)| = \tilde{d}(\beta_{n_k},\beta^*)\left|\sum_{i=1}^{m+2}\gamma_{n_k,i}f_i(x)\right| \geq \varepsilon\left|\sum_{i=1}^{m+2}\gamma_{n_k,i}f_i(x)\right|,$$

where $\gamma_{n_k} = \tilde{d}(\beta_{n_k},\beta^*)^{-1}(\beta_{n_k} - \beta^*)$. The subsequence $\gamma_{n_1},\gamma_{n_2},\ldots$ takes values in the unit sphere in $\mathbb{R}^{m+2}$, which is compact, and so we have a further subsequence $\gamma_{n_{k_1}},\gamma_{n_{k_2}},\ldots$ that converges to some γ^* in the unit sphere. Therefore, arguing as we did above with Cauchy's inequality, we have $\sum_{i=1}^{m+2}\gamma_{n_{k_j},i}f_i(x) \to \sum_{i=1}^{m+2}\gamma_i^* f_i(x)$ pointwise in x as $j \to \infty$. Noting that $\sum_{i=1}^{m+2}\gamma_i^* f_i(x) \neq 0$ on a set of positive P^X-measure, we conclude that the subsequence $\theta_{n_{k_j}}$ cannot contain a further subsequence that converges to θ^* pointwise on a set of P^X-measure one. Recall (see e.g. Theorem 9.2.1 in Dudley, 2002) that a sequence of random variables converges in probability if and only if every subsequence of random variables contains a further subsequence that is almost surely convergent. It must therefore be the case that $\theta_n(X) \nrightarrow_p \theta^*(X)$. Since F^Y has a continuous inverse, this implies that $F^Y(\theta_n(X)) \nrightarrow_p F^Y(\theta^*(X))$, which implies that $E|F^Y(\theta_n(X)) - F^Y(\theta^*(X))| \nrightarrow 0$. That is, $d(\theta_n,\theta^*) \nrightarrow 0$. This establishes that $\mathcal{S}$ is continuous, which proves (ii).

To prove (iii), we must show that $(\theta,x) \mapsto \theta(x)$ is $\mathcal{B}(\Psi_m(s)) \otimes \mathcal{B}(R^X)$-measurable. By Theorem 12.2.1 in Dudley (2002), it suffices to show that $(\Psi_m(s),d)$ is a separable metric space, and that $\theta \mapsto \theta(x)$ is continuous in θ for each $x \in R^X$. The former assertion was established in (ii). To demonstrate the latter assertion, we let $\theta_1,\theta_2,\ldots$ be a sequence in $\Psi_m(s)$ converging to some $\theta^* \in \Psi_m(s)$. Using the topological isomorphism $\mathcal{S}$ introduced above, we identify our sequence in $\Psi_m(s)$ with another sequence $\beta_1,\beta_1,\ldots$ in $\tilde{\Psi}_m(s)$ converging to $\beta^* \in \tilde{\Psi}_m(s)$. As above, we have $|\theta_n(x) - \theta^*(x)| \leq \tilde{d}(\beta_n,\beta^*)(\sum_{i=1}^{m+2} f_i(x)^2)^{1/2}$ for each $x \in R^X$, so that θ_n converges to θ^* pointwise. This proves that $\theta \mapsto \theta(x)$ is continuous in θ for each $x \in R^X$. □

Proof of Theorem 4.2. We first show that $\Theta^\dagger$ is dense in Θ^* under d. Fix a function $\theta^* \in \Theta^*$. The function $\tilde{\theta}^* = F^Y \circ \theta^* \circ Q^X$ is a Borel measurable mapping from $[0,1)$ to $[0,1)$; we extend the domain and range of $\tilde{\theta}^*$ to $[0,1]$ by setting $\tilde{\theta}^*(1) = 1$. A well known consequence of Urysohn's lemma (see e.g. Lemma 2.6.3

in Dudley, 2002) is that the continuous functions on $[0,1]$ form a dense subset of the Lebesgue integrable functions on $[0,1]$ under the L_1-seminorm. It is also well known (see e.g. Theorem 11.2.4 and the following example in Dudley, 2002) that any continuous function on $[0,1]$ can be approximated arbitrarily well by a continuous piecewise linear function on $[0,1]$, with finitely many kinks. Such a function can in turn be approximated arbitrarily well by a continuous piecewise linear function on $[0,1]$, with finitely many kinks, for which the slope of the function is nonzero wherever it is defined. We will let $\tilde{\theta}_1, \tilde{\theta}_2, \ldots$ be a sequence of such functions, chosen such that $\lim_{k\to\infty} \int |\tilde{\theta}^*(u) - \tilde{\theta}_k(u)| \mathrm{d}u = 0$ and $0 \leq \tilde{\theta}_k \leq 1$ for each $k \in \mathbb{N}$.

Let U be a random variable distributed uniformly on $[0,1]$, and let $F^U(\cdot; \tilde{\theta}_k)$ denote the distribution function of $\tilde{\theta}_k(U)$. The piecewise linearity and nonzero slope of each $\tilde{\theta}_k$ ensures that $F^U(\cdot; \tilde{\theta}_k)$ is Lipschitz continuous for each k. Specifically, for $a, b \in [0,1]$ we have $|F^U(b; \tilde{\theta}_k) - F^U(a; \tilde{\theta}_k)| \leq \nu_k^{-1}(N_k+1)|b-a|$, where ν_k is a lower bound on the derivative of $\tilde{\theta}_k$, and N_k is the number of kinks. Let $\tilde{\theta}_k^\dagger : [0,1] \to [0,1]$ be defined by $\tilde{\theta}_k^\dagger(u) = F^U(\tilde{\theta}_k(u); \tilde{\theta}_k)$. Since $\tilde{\theta}_k$ and $F^U(\cdot; \tilde{\theta}_k)$ are Lipschitz continuous for each k, each $\tilde{\theta}_k^\dagger$ must also be Lipschitz continuous. Let $\theta_k^\dagger = Q^Y \circ \tilde{\theta}_k^\dagger \circ F^X$, and observe that $\theta_k^\dagger(X) = Q^Y(\tilde{\theta}_k^\dagger(F^X(X))) \sim Q^Y(\tilde{\theta}_k^\dagger(U)) = Q^Y(F^U(\tilde{\theta}_k(U); \tilde{\theta}_k)) \sim Q^Y(U) \sim Y$. This establishes that $\theta_k^\dagger \in \Theta^\dagger$ for each k. It remains to show that $\lim_{k\to\infty} d(\theta_k^\dagger, \theta^*) = 0$. Since $d(\theta_k^\dagger, \theta^*) = \int |\tilde{\theta}_k^\dagger(u) - \tilde{\theta}^*(u)| \mathrm{d}u$, it suffices to show $\lim_{k\to\infty} \int |\tilde{\theta}_k^\dagger(u) - \tilde{\theta}_k(u)| \mathrm{d}u = 0$. Observe that $\int |\tilde{\theta}_k^\dagger(u) - \tilde{\theta}_k(u)| \mathrm{d}u = \int |F^U(\tilde{\theta}_k(u); \tilde{\theta}_k) - \tilde{\theta}_k(u)| \mathrm{d}u \leq \sup_{0\leq v\leq 1} |F^U(v; \tilde{\theta}_k) - v|$. Since $\tilde{\theta}^*(U) \sim U$ and $\lim_{k\to\infty} \int |\tilde{\theta}^*(u) - \tilde{\theta}_k(u)| \mathrm{d}u = 0$, Theorems 2.2, 2.3 and 2.4 jointly imply that $\tilde{\theta}_k(U) \to_d U$ as $k \to \infty$. But this implies that $\lim_{k\to\infty} \sup_{0\leq v\leq 1} |F^U(v; \tilde{\theta}_k) - v| = 0$, which proves $\lim_{k\to\infty} \int |\tilde{\theta}_k^\dagger(u) - \tilde{\theta}_k(u)| \mathrm{d}u = 0$. We conclude that $\Theta^\dagger$ is dense in Θ^* under d.

Next we show that $\inf_{\theta\in\Theta_n} d(\theta, \theta^\dagger) = O(m_n \sup_{0\leq i\leq m_n} P^X(s_{i,n}, s_{i+1,n})^2)$ for each $\theta^\dagger \in \Theta^\dagger$. Choose $\theta_n \in \Theta_n$ such that $\theta_n(s_{i,n}) = \theta^\dagger(s_{i,n})$ for $i = 0, \ldots, m_n$, and such that $\theta_n(x)$ is constant for $x \geq s_{m_n,n}$. Then we have

$$\begin{aligned} d(\theta_n, \theta^\dagger) &= \sum_{i=0}^{m_n} \int_{s_{i,n}}^{s_{i+1,n}} |F^Y(\theta_n(x)) - F^Y(\theta^\dagger(x))| \mathrm{d}F^X(x) \\ &= \sum_{i=0}^{m_n} \int_{F^X(s_{i,n})}^{F^X(s_{i+1,n})} \left|F^Y \circ \theta_n \circ Q^X(u) - F^Y \circ \theta^\dagger \circ Q^X(u)\right| \mathrm{d}u, \end{aligned}$$

where $F^X(\infty) = 1$. Our choice of θ_n and the Lipschitz property of $F^Y \circ \theta^\dagger \circ Q^X$ ensure that

$$\sup_{F^X(s_{i,n})<u<F^X(s_{i+1,n})} \left|F^Y \circ \theta_n \circ Q^X(u) - F^Y \circ \theta^\dagger \circ Q^X(u)\right| \leq K \left|F^X(s_{i+1,n}) - F^X(s_{i,n})\right|$$

for $i = 0, \ldots, m_n$, where K is the Lipschitz coefficient of $F^Y \circ \theta^\dagger \circ Q^X$. We thus have

$$d(\theta_n, \theta^\dagger) \leq K \sum_{i=0}^{m_n} \left| F^X(s_{i+1,n}) - F^X(s_{i,n}) \right|^2 = K \sum_{i=0}^{m_n} P^X(s_{i,n}, s_{i+1,n})^2 \leq K m_n \sup_{0 \leq i \leq m_n} P^X(s_{i,n}, s_{i+1,n})^2,$$

giving the desired result. □

References

Amin, G. S. and Kat, H. M. (2003). Hedge fund performance 1990-2000: do the "money machines" really add value? *Journal of Financial and Quantitative Analysis* **38** 251–274.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Black, F. and Scholes, M. (1973). Pricing of options and corporate liabilities. *Journal of Political Economy* **81** 637–654.

Brown, D. P. and Jackwerth, J. C. (2004). The pricing kernel puzzle: reconciling index option data and economic theory. Working paper, University of Wisconsin at Madison, School of Business.

Cassidy, J. (2007). Hedge Clipping. *The New Yorker* 7/2/2007.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Chapter 76 in Heckman, J. J. and Leamer, E. E. (eds) *Handbook of Econometrics,* vol. 6B. Elsevier, Amsterdam.

Corradi, V. and White, H. (1995). Regularized neural networks: some convergence rate results. *Neural Computation* **7** 1225–1244.

Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques* **31** 393–427.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability* **6** 899–929.

Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge.

Dudley, R. M. (2002). *Real Analysis and Probability,* 2nd ed. Cambridge University Press, Cambridge.

Dybvig, P. H. (1988a). Distributional analysis of portfolio choice. *Journal of Business* **63** 369–393.

Dybvig, P. H. (1988b). Inefficient dynamic portfolio strategies or how to throw away a million dollars in the stock market. *Review of Financial Studies* **1** 67–88.

Hasanhodzic, J. and Lo, A. W. (2007). Can hedge-fund returns be replicated? The linear case. *Journal of Investment Management* **5** 5–45.

Jackwerth, J. C. (2000). Recovering risk aversion from option prices and realized returns. *Review of Financial Studies* **13** 433–451.

Kat, H. M. (2007). Alternative routes to hedge fund return replication. *Journal of Wealth Management* **10** 29–39.

Kat, H. M. and Palaro, H. P. (2005a) Hedge fund returns: you can make them yourself! *Journal of Wealth Management* **8** 62–68.

Kat, H. M. and Palaro, H. P. (2005b) Who needs hedge funds? A copula based approach to hedge fund return replication. Alternative Investment Research Centre Working Paper No. 27, Cass Business School Research Papers.

Kat, H. M. and Palaro, H. P. (2006) Replicating hedge fund returns using futures: a European perspective. Alternative Investment Research Centre Working Paper No. 32, Cass Business School Research Papers.

Laise, E. (2007). The hedge fund 'clones'. *The Wall Street Journal* 7/1/2007.

Merton, R. (1973). An intertemporal capital asset pricing model. *Econometrica* **41** 867–887.

Rio, E. (1998). Processus empiriques absolument réguliers et entropie universelle. *Probability Theory and Related Fields* **111** 585–608.

Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants.* Springer-Verlag, Berlin.

Stinchcombe, M. B. and White, H. (1992). Some measurability results for extrema of random functions over random sets. *Review of Economic Studies* **59** 495–512.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Vapnik, V. N. and Červonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* **16** 264–280.