Semiparametric Dynamic Portfolio Choice with Multiple Conditioning Variables[*]

Jia Chen[†]   Degui Li[‡]   Oliver Linton[§]   Zudi Lu[¶]

September 15, 2015

Abstract

Dynamic portfolio choice has been a central and essential objective for investors in active asset management. In this paper, we study the dynamic portfolio choice with multiple conditioning variables, where the dimension of the conditioning variables can be either fixed or diverging to infinity at certain polynomial rate of the sample size. We propose a novel data-driven method to estimate the optimal portfolio choice, motivated by the model averaging marginal regression approach suggested by Li, Linton and Lu (2015). More specifically, in order to avoid the curse of dimensionality associated with the multivariate nonparametric regression problem and to make it practically implementable, we first estimate the marginal optimal portfolio choice by maximising the conditional utility function for each univariate conditioning variable, and then construct the joint dynamic optimal portfolio through the weighted average of the marginal optimal portfolio across all the conditioning variables. Under some regularity conditions, we establish the large sample properties for the developed portfolio choice procedure. Both the simulation study and empirical application well demonstrate the finite-sample performance of the proposed methodology.

[†]Department of Economics and Related Studies, University of York, Heslington, YO10 5DD, UK. E-mail: jia.chen@york.ac.uk.

[‡]Department of Mathematics, University of York, Heslington, YO10 5DD, UK. E-mail: degui.li@york.ac.uk.

[§]Faculty of Economics, Cambridge University, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. E-mail: obl20@cam.ac.uk.

[¶]School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. E-mail: Z.Lu@soton.ac.uk.

1   Introduction

Portfolio choice is a central issue for investors and asset managers. Financial research has clarified how this might be carried out to meet various objectives. Fundamental contributions to this literature have been made, inter alia, by: Markowitz (1952), Sharpe (1963), Merton (1969), Samuelson (1969), and Fama (1970). See Back (2010) and Brandt (2010) for some recent surveys. In practice, it is not uncommon that dynamic portfolio choice depends on many conditioning (or forecasting) variables, which reflect the varying investment opportunities over time. Generally speaking, there are two ways to characterize the dependence of portfolio choice on the conditioning variables. One is to assume a parametric statistical model that relates the returns of risky assets to the conditioning variables and then solve for an investor's portfolio choice by using some traditional econometric approaches to estimate the conditional distribution of the returns. However, the pre-assumed parametric model might be misspecified, which would lead to inconsistent or biased estimation of the optimal portfolio. The other way, which avoids the possible issue of model misspecification, is to use some nonparametric methods such as the kernel estimation method to characterize the dependence of the portfolio choice on conditioning variables. This latter method is first introduced by Brandt (1999), who also establishes the asymptotic properties for the estimated portfolio choice and provides an empirical application.

Although the nonparametric method allows the financial data to "speak for themselves" and is robust to model misspecification, its performance is often poor when the dimension of the conditioning variables is large (say, larger than three), owing to the so-called "curse of dimensionality" (c.f., Fan and Yao, 2003). This indicates that a direct use of Brandt (1999)'s nonparametric method may be inappropriate when there are multiple conditioning variables. Our main objective in this paper is to address this issue in dynamic portfolio choice problem with multiple conditioning variables and propose a novel data-driven method to estimate the optimal portfolio choice, where the dimension of the conditioning variables and the number of the risky assets can be either fixed or diverging to infinity at certain polynomial rate of the sample size.

In order to avoid the curse of dimensionality issue, we first consider the optimal portfolio choice for a given univariate conditioning variable, and then construct the joint dynamic optimal portfolio

choice through a weighted average of the marginal optimal portfolio across all the conditioning variables. This method is partly motivated by the Model Averaging MArginal Regression (MAMAR) approach suggested in a recent paper by Li, Linton and Lu (2015), which shows that such a method performs well in estimating the conditional multivariate mean regression function and out-of-sample prediction. Furthermore, we introduce a semiparametric data-driven method to choose the optimal weights in model averaging. Under some mild conditions, we establish the large sample properties for the developed portfolio choice procedure to show its advantages over the conventional nonparametric kernel smoothing method in terms of convergence. Both simulation studies and an empirical application are carried out to examine the finite sample performance of the proposed methodology.

The structure of the paper is as follows. The methodology for estimating the dynamic portfolio choice is introduced in Section 2, and the relevant large sample theory is presented in Section 3. The data-driven choice of the optimal weights in model averaging of the marginal optimal portfolios across all conditioning variables is developed in Section 4. Numerical studies including both simulation and an empirical application are reported in Section 5. Section 6 concludes the paper. The assumptions and the technical proofs of the main results are relegated to Appendices A and B, respectively.

2 Methodology for estimating dynamic portfolio choice

Suppose that there are n risky assets with $\mathbf{R}_t = (R_{1t}, \ldots, R_{nt})^\intercal$ as a vector of gross returns at time t, $t = 1, \ldots, T$, where n can be either fixed or diverging to infinity with the sample size T. Let $\mathbf{X}_t = (X_{1t}, \ldots, X_{Jt})^\intercal$, where J is the number of conditioning or forecasting variables X_{jt}. The dynamic portfolio choice aims to choose the portfolio weights at each time period t by maximising the conditional utility function defined by

$$\mathsf{E}\left[u(w^\intercal \mathbf{R}_t)|\mathbf{X}_{t-1}\right] = \mathsf{E}\left[u(w^\intercal \mathbf{R}_t)|(X_{1,t-1}, \ldots, X_{J,t-1})\right], \tag{2.1}$$

subject to $\mathbf{1}_n^\intercal w = \sum_{i=1}^n w_i = 1$, where $w = (w_1, \ldots, w_n)^\intercal$, $\mathbf{1}_n$ is the n-dimensional column vector of ones, $u(\cdot)$ is a concave utility function which measures the investor's utility with wealth $w^\intercal \mathbf{R}_t$ at time t. For simplicity, we only focus on the problem of single-period portfolio choice. Furthermore, we assume that the investors can borrow assets and sell them (short selling), which indicates that some of the portfolio weights may take negative values.

The classic mean-variance paradigm considers the quadratic utility function $u(v) = v - (\gamma/2)v^2$ or the CARA (Constant Absolute Risk Aversion) utility function $u(v) = -\exp(-\gamma v)$ plus normality, in which case the solution (with covariates) is explicitly defined in terms of the conditional

3

mean vector $\boldsymbol{\mu}(\mathbf{x}) = \mathsf{E}[\mathbf{R}_t|\mathbf{X}_{t-1} = \mathbf{x}]$, $\mathbf{x} = (x_1, \ldots, x_J)^\mathsf{T}$, and the conditional covariance matrix $\boldsymbol{\Sigma}(\mathbf{x}) = \mathsf{E}[(\mathbf{R_t}-\boldsymbol{\mu}(\mathbf{x}))(\mathbf{R_t}-\boldsymbol{\mu}(\mathbf{x}))^\mathsf{T}|\mathbf{X}_{t-1} = \mathbf{x}]$ of returns, i.e.,

$$w(\mathbf{x}) = \frac{1}{\gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x})\left[\boldsymbol{\mu}(\mathbf{x}) - \theta(\mathbf{x})\mathbf{1}_n\right], \quad \theta(\mathbf{x}) = \frac{\boldsymbol{\mu}(\mathbf{x})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x})\mathbf{1}_n - \gamma}{\mathbf{1}_n^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x})\mathbf{1}_n}.$$

In this case, it suffices to know $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$. One may also work with the more general CRRA (Constant Relative Risk Aversion) utility function with risk aversion parameter γ

$$u(v) = \left\{ \begin{array}{ll} \frac{v^{1-\gamma}}{1-\gamma}, & \gamma \neq 1 \\ \log v, & \gamma = 1, \end{array} \right.$$

in which case the solution for the optimal weights is not typically explicit, and generally depends on more features of the conditional distribution. More discussion on different classes of utility functions $u(\cdot)$ can be found in Chapter 1 of the book by Back (2010).

In order to solve the general maximisation problem in (2.1), Brandt (1999) proposes a nonparametric conditional method of moments approach, which can be seen as an extension of the method of moments approach in Hansen and Singleton (1982). Taking the first-order derivative of $u(\cdot)$ in (2.1) with respect to w_i and considering the constraint of $\mathbf{1}_n^\mathsf{T}w = \sum_{i=1}^n w_i = 1$, we may obtain the dynamic portfolio choice by solving the following equations for $w_1, \ldots, w_{n-1}$:

$$\mathsf{E}\left[(R_{it} - R_{nt})\dot{u}(w^\mathsf{T}\mathbf{R}_t)|X_{1,t-1}, \ldots, X_{J,t-1}\right] = 0 \quad a.s., \quad i = 1, \ldots, n-1, \tag{2.2}$$

where $\dot{u}(\cdot)$ is the derivative of the utility function $u(\cdot)$. The last element w_n in w can be determined by using the constraint $\sum_{i=1}^n w_i = 1$. Brandt (1999) suggests a kernel-based smoothing method to estimate the solution to (2.2). However, when J is large, the kernel-based nonparametric conditional method of moments approach would perform quite poorly due to the curse of dimensionality discussed in Section 1. Therefore, we propose a novel dimension-reduction technique to address this problem.

We start with the portfolio choice for each univariate conditioning variable in $\mathbf{X}_{t-1}$. For $j = 1, \ldots, J$, we define the marginal conditional utility function as

$$\mathsf{E}\left[u(w^\mathsf{T}\mathbf{R}_t)|X_{j,t-1} = x_j\right] \tag{2.3}$$

with the constraint $\mathbf{1}_n^\mathsf{T}w = \sum_{i=1}^n w_i = 1$. The associated first-order conditions for the marginal optimal portfolio weights $w_j^*(x_j)$ evaluated at x_j for the conditioning variables are:

$$\mathsf{E}\left[(R_{it} - R_{nt})\dot{u}(w_j^{*\mathsf{T}}(x_j)\mathbf{R}_t)|X_{j,t-1} = x_j\right] = 0, \quad i = 1, \cdots, n-1, \tag{2.4}$$

4

where

$$w_j^*(x_j) = \left[w_{1j}^*(x_j), \ldots, w_{nj}^*(x_j) \right]^\top \quad \text{with} \quad w_{nj}^*(x_j) = 1 - \sum_{i=1}^{n-1} w_{ij}(x_j).$$

For a given j, this is essentially the problem posed and solved by Brandt (1999). For given $\mathbf{x} = (x_1, \ldots, x_J)^\top$, (2.3) and (2.4) may be understood as the utility function and the corresponding first order conditions for portfolio choice in a "fictitious economy", where the realization of each univariate conditioning variable determines the state of the economy.

We next consider how to combine the marginal portfolios selected above to form a joint portfolio. We shall consider a weighted average of the marginal portfolio choices $w_j(x_j)$ over $j = 1, \ldots, J$, and obtain the joint portfolio choice as

$$w_{\mathbf{a}}(\mathbf{x}) = \sum_{j=1}^{J} a_j w_j(x_j) \quad \text{with} \quad \sum_{j=1}^{J} a_j = 1, \tag{2.5}$$

where negative values for a_j can be allowed. In Section 4 below, we will discuss how to choose the weights $\mathbf{a} = (a_1, \ldots, a_J)^\top$ in the combination (2.5) by using a data-driven method.

The joint portfolio choice $w_{\mathbf{a}}(\mathbf{x})$ defined in (2.5) can, in some sense, be seen as an *approximation* of the true optimal portfolio choice, as we next discuss. Consider the following class of weights (that are measurable functions of the covariates):

$$\mathcal{W} = \left\{ w(\cdot) : \sum_{i=1}^{n} w_i(\mathbf{x}) = 1 \right\}$$

$$\mathcal{W}^* = \left\{ w_a(\cdot) : w_a(\mathbf{x}) = \sum_{j=1}^{J} a_j w_j^*(x_j), \ \sum_{i=1}^{J} a_j = 1 \right\}$$

$$\mathcal{W}_{add} = \left\{ w_{add}(\cdot) : w_{add}(\mathbf{x}) = \sum_{j=1}^{J} w_j(x_j), \ \sum_{i=1}^{n} w_{ji}(x_j) = 1 \right\},$$

where $w_j^*(\cdot)$ with $\sum_{i=1}^{n} w_{ji}^*(x_j) = 1$ are the optimal weights to (2.3). Note that $\mathcal{W}^* \subset \mathcal{W}_{add} \subset \mathcal{W}$. The solution to (2.2) is a member of $\mathcal{W}$, and our proposed method is a member of $\mathcal{W}^*$. One could solve the intermediate additive portfolio problem, i.e., $\max \mathsf{E}\left[u(w^\top \mathbf{R}_t) | \mathbf{X}_{t-1} \right]$ with respect to $w \in \mathcal{W}_{add}$, which also only requires finding one dimensional functions, but this will require us to solve a nonlinear integral equation system in the $n \times J$ scalar functions $w_{ji}(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$. This is likely to be computationally difficult and theoretically challenging to analyze. We consider our approach to be a practically feasible alternative to this, in the same way as the MAMAR approximation introduced

5

in Li, Linton and Lu (2015) can be considered a practically feasible alternative to high dimensional additive regression.[1]

We next introduce a kernel-based nonparametric estimation methodology to estimate the approximate joint portfolio choice defined in (2.5). Let $K(\cdot)$ be a kernel function and h be a bandwidth that converges to zero as T tends to infinity. Using the sample information, we may express the first-order conditions (2.4) for the marginal optimal portfolio as

$$\frac{1}{Th}\sum_{t=1}^{T}(R_{it} - R_{nt})\dot{u}(w^{\intercal}\mathbf{R}_t)K\left(\frac{X_{j,t-1} - x_j}{h}\right) = 0, \quad i = 1, \ldots, n-1, \tag{2.6}$$

for each $j = 1, \cdots, J$. Denote $\widehat{w}_j^*(x_j) = [\widehat{w}_{1j}^*(x_j), \ldots, \widehat{w}_{nj}^*(x_j)]^{\intercal}$ the solution to (2.6), where

$$\widehat{w}_{nj}^*(x_j) = 1 - \sum_{i=1}^{n-1}\widehat{w}_{ij}^*(x_j). \tag{2.7}$$

Then we estimate the joint optimal portfolio choice by the weighted average

$$\widehat{w}_{\mathbf{a}}(\mathbf{x}) = \sum_{j=1}^{J}a_j\widehat{w}_j^*(x_j), \quad \sum_{j=1}^{J}a_j = 1. \tag{2.8}$$

The asymptotic properties for $\widehat{w}_{\mathbf{a}}(\mathbf{x})$ will be given in Section 3 below for various scenarios.

3 Large sample theory

In this section, we give the asymptotic theorems for the estimation developed in the previous section for the following four cases: (i) both $J = J_0$ and $n = n_0$ are fixed positive integers, (ii) $J = J_T$ is a diverging positive integer whereas $n = n_0$ is fixed, (iii) $n = n_T$ is a diverging positive integer whereas $J = J_0$ is fixed, and (iv) both $J = J_T$ and $n = n_T$ are diverging with the sample size T.

We start with case (i): $J = J_0$ and $n = n_0$. Given (2.7), we only need study the asymptotic theory for $\widehat{w}_j^*(x_j) = [\widehat{w}_{1j}^*(x_j), \ldots, \widehat{w}_{n_0-1,j}^*(x_j)]^{\intercal}$, the estimate of $w_j^*(x_j) = \left[w_{1j}^*(x_j), \ldots, w_{n_0-1,j}^*(x_j)\right]^{\intercal}$. Before

---

[1]We could also instead seek to approximate the objective function $Q(w; \mathbf{x}) = \mathsf{E}\left[u(w^{\intercal}\mathbf{R}_t)|X_{t-1} = \mathbf{x}\right]$ by using the MAMAR method directly, i.e., to approximate this conditional expectation by a weighted sum of one dimensional nonparametric regressions $\widetilde{Q}(w; \mathbf{x}) = \sum_{j=1}^{J}\alpha_j\mathsf{E}\left[u(w^{\intercal}\mathbf{R}_t)|X_{jt-1} = x_j\right]$ for some weights α_j, and then optimizing $\widetilde{Q}(w; \mathbf{x})$ with respect to w. We conjecture that such a method may give similar results except that it provides less diagnostic information. However, it is perhaps harder to define a data-driven method for selecting α_j as more constraints might be involved.

stating the asymptotic theorems, we introduce some notation. For $j = 1, \ldots, J_0$ and $t = 1, \ldots, T$, we define

$$\boldsymbol{\Lambda}_j(x_j) = f_j(x_j)\mathsf{E}\left[\mathbf{R}_t^*(\mathbf{R}_t^*)^{\mathsf{T}}\ddot{u}\left(w_j^{*\mathsf{T}}(x_j)\mathbf{R}_t\right)|X_{j,t-1} = x_j\right],$$

$$\mathbf{Z}_{jt}(x_j) = \mathbf{R}_t^*\dot{u}\left(w_j^{*\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\right)K\left(\frac{X_{j,t-1} - x_j}{h}\right),$$

where $\mathbf{R}_t^* = (R_{1t} - R_{n_0 t}, \ldots, R_{n_0-1,t} - R_{n_0 t})^{\mathsf{T}}$, $\ddot{u}(\cdot)$ is the second-order derivative of $u(\cdot)$ and $f_j(\cdot)$ is the marginal density function of X_{jt}. Define

$$\mathbf{W}_{jt}(x_j) = \boldsymbol{\Lambda}_j^{-1}(x_j)\mathbf{Z}_{jt}(x_j) \quad \text{and} \quad \mathbf{W}_t(\mathbf{x}|\mathbf{a}) = \sum_{j=1}^{J_0} a_j \mathbf{W}_{jt}(x_j)$$

for $t = 1, \ldots, T$. Following the argument in the proof of Theorem 3.1 in Appendix B and letting

$$\widehat{w}_{\mathbf{a}}^*(\mathbf{x}) = \sum_{j=1}^{J} a_j \widehat{w}_j^*(x_j) \quad \text{and} \quad w_{\mathbf{a}}^*(\mathbf{x}) = \sum_{j=1}^{J} a_j w_j^*(x_j),$$

we may show that

$$\sqrt{Th}\left[\widehat{w}_{\mathbf{a}}^*(\mathbf{x}) - w_{\mathbf{a}}^*(\mathbf{x})\right] = \frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\mathbf{W}_t(\mathbf{x}|\mathbf{a}) + o_P(1) \tag{3.1}$$

for given $\mathbf{a} = (a_1, \ldots, a_{J_0})^{\mathsf{T}}$. The asymptotic distribution theory for $\widehat{w}_j^*(x_j)$ and $\widehat{w}_{\mathbf{a}}^*(\mathbf{x})$ is given in Theorem 3.1 below.

THEOREM 3.1. *Suppose that Assumptions 1–5 in Appendix A are satisfied, both $J = J_0$ and $n = n_0$ are fixed positive integers.*

(i) For $j = 1, \ldots, J_0$, we have

$$\sqrt{Th}\left[\widehat{w}_j^*(x_j) - w_j^*(x_j)\right] \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \; \boldsymbol{\Omega}_j(x_j)\right), \tag{3.2}$$

where $\boldsymbol{\Omega}_j(x_j) = \mathsf{E}\left[\mathbf{W}_{jt}(x_j)\mathbf{W}_{jt}^{\mathsf{T}}(x_j)\right] = \boldsymbol{\Lambda}_j^{-1}(x_j)\mathsf{E}\left[\mathbf{Z}_{jt}(x_j)\mathbf{Z}_{jt}^{\mathsf{T}}(x_j)\right]\boldsymbol{\Lambda}_j^{-1}(x_j)$.

(ii) For a set of given weights $(a_1, \cdots, a_{J_0})$ with $\sum_{j=1}^{J_0} a_j = 1$, we have

$$\sqrt{Th}\left[\widehat{w}_{\mathbf{a}}^*(\mathbf{x}) - w_{\mathbf{a}}^*(\mathbf{x})\right] \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \; \boldsymbol{\Omega}(\mathbf{x}|\mathbf{a})\right), \tag{3.3}$$

where $\boldsymbol{\Omega}(\mathbf{x}|\mathbf{a}) = \mathsf{E}\left[\mathbf{W}_t(\mathbf{x}|\mathbf{a})\mathbf{W}_t^{\mathsf{T}}(\mathbf{x}|\mathbf{a})\right]$.

Although there are multiple conditioning variables in the nonparametric dynamic portfolio choice, we can still achieve the root-(Th) convergence rates as shown in the above theorem. This means

7

that we can successfully overcome the curse of dimensionality problem. The main reason is that, in the estimation of the joint optimal portfolio, we apply the univariate kernel smoothing to estimate the marginal optimal portfolio choice for each univariate conditioning variable and then obtain the joint portfolio choice through a weighted average defined in (2.8). In contrast, if we directly use the multivariate kernel smoothing as is done in Brandt (1999), the convergence rate for the resulting estimation of the joint portfolio choice would be root-(Th^{J_0}), slower than the rates in (3.2) and (3.3) when $J_0 > 1$.

We next consider case (ii) that $n = n_0$ is fixed but $J = J_T$ is diverging, and give the relevant asymptotic distribution theory in the following theorem.

THEOREM 3.2. *Suppose that Assumptions 1–4 and 5′ in Appendix A are satisfied, the number of the risky assets is fixed, and the number of conditioning variables is a positive integer J_T which is diverging with the sample size T. Then, the asymptotic distributions in (3.2) and (3.3) of Theorem 3.1 still hold.*

Theorem 3.2 above indicates that the root-(Th) convergence rate remains even when the number of the potential conditioning variables is diverging at a rate that satisfies the restriction in Assumption 5′, i.e.,

$$\frac{T^{1-1/(2+\delta)}h}{J_T^{1/(2+\delta)}\log T} \to \infty, \quad \delta > 0,$$

see also the second condition in (3.6) below. Such a restriction means that J_T can be larger than T, if we are only interested in $\widehat{w}_j^*(\cdot)$ or $\widehat{w}_{\mathbf{a}}^*(\cdot)$ for a given $\mathbf{a}$. However, some additional restrictions on J_T would be needed when we also consider the optimal choice of $\mathbf{a} = (a_1, \ldots, a_{J_T})^\intercal$, see Section 4 below for details.

For cases (i) and (ii), we can further establish the uniform consistency results for $\widehat{w}_j(x_j)$ over $x_j \in \mathcal{X}_j$ with $\mathcal{X}_j$ being the support of X_{jt}.

THEOREM 3.3. *Suppose that Assumptions 1–4 in Appendix A are satisfied.*

(i) If both $J = J_0$ and $n = n_0$ are fixed positive integers, and

$$h \to 0, \quad \frac{T^{1-2/(2+\delta)}h}{\log T} \to \infty, \tag{3.4}$$

where $\delta > 0$ is specified in Assumption 3 in Appendix A, then

$$\max_{1 \le j \le J_0} \sup_{x_j \in \mathcal{X}_j} \|\widehat{w}_j(x_j) - w_j(x_j)\| = O_P\left(h^2 + \sqrt{\log T/(Th)}\right), \tag{3.5}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector or the Frobenius norm of a matrix.

8

(ii) If $n = n_0$ is fixed and $J = J_T$ is diverging with the sample size T, and

$$h \to 0, \quad \frac{T^{1-2/(2+\delta)}h}{J_T^{2/(2+\delta)} \log T} \to \infty, \tag{3.6}$$

then (3.5) still holds with J_0 replaced by J_T.

In fact, Theorem 3.3 (i) can be considered as a special case of Theorem 3.3 (ii), and the above uniform consistency results can be seen as an extension of the uniform consistency results for the nonparametric kernel-based estimation for stationary time series (Hansen, 2008; Kristensen, 2009; Li, Lu and Linton, 2012) to the case of nonparametric portfolio choice. Note that the order h^2 contributed by the bias term would be asymptotically dominated by the order $\sqrt{\log T/(Th)}$ if $Th^4 = o(1)$ in Assumption 5 (or 5′) is satisfied. In the latter case, the uniform convergence rate would become $O_P\left(\sqrt{\log T/(Th)}\right)$. By modifying the proof in Appendix B, we may further generalise (3.5) to the case where $\mathcal{X}_j$ is an expanding set.

We finally give a brief discussion on how to generalize Theorems 3.1–3.3 to cases (iii) and (iv) in which the number of the risky assets is diverging with the sample size T. Note that the dimension of $\widehat{w}_j(x_j)$ is n_T, which may slow down the convergence rate when $n_T \to \infty$. However, to derive some sensible asymptotic results, n_T cannot diverge to infinity too fast. Following the arguments in the proof of Theorem 3.3 in Appendix B and those in the high-dimensional variable selection literature (c.f., Fan and Peng, 2004), we may show that when

$$n_T^4 = o(Th), \quad h \to 0, \quad \frac{T^{1-2/(2+\delta)}h}{(n_T J)^{2/(2+\delta)} \log T} \to \infty, \tag{3.7}$$

the uniform consistency result in (3.5) can be generalized to

$$\max_{1 \leq j \leq J} \sup_{x_j \in \mathcal{X}_j} \|\widehat{w}_j(x_j) - w_j(x_j)\| = O_P\left(\sqrt{n_T}\left(h^2 + \sqrt{\log T/(Th)}\right)\right), \tag{3.8}$$

where the uniform convergence rate is slower than that in (3.5). The above result holds no matter when J is fixed or diverging to infinity. We may also generalize the asymptotic distribution theory in Theorems 3.1 and 3.2 to the case of diverging n_T. In the latter case, as the dimension of the estimated portfolio weights is diverging, we need to apply a transformation matrix with full row rank to reduce the dimension from n_T to n_* (a fixed positive integer) when deriving the asymptotic normal distribution. As the relevant argument is similar to that in proving Theorem 4.2 below, we omit the details here to save space.

9

4 Data-driven weight choice in model averaging

The performance of the dynamic portfolio choice defined in (2.8) relies on the choice of the model averaging weights $a_1, \ldots, a_J$. Let $w_{\mathbf{a}t} \equiv w_a(\mathbf{X}_t) = \sum_{j=1}^{J} a_j w_j(X_{jt})$ and define the objective function:

$$U(\mathbf{a}) = \mathsf{E}\left[u(w_{\mathbf{a},t-1}^{\mathsf{T}} \mathbf{R}_t)\right] = \mathsf{E}\left\{ u\Big[\sum_{j=1}^{J} a_j w_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] \right\}, \tag{4.1}$$

which is the expected utility associated with a particular choice of $\mathbf{a}$. We choose the optimal weights $\mathbf{a}_0$ by maximising $U(\mathbf{a})$, i.e.,

$$\mathbf{a}_0 = \arg\max_{\mathbf{a}} U(\mathbf{a}) = \arg\max_{\mathbf{a}} \mathsf{E}\left\{ u\Big[\sum_{j=1}^{J} a_j w_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] \right\}. \tag{4.2}$$

subject to the constraint $\sum_{j=1}^{J} a_j = 1$. This leads to the following first-order conditions

$$\mathsf{E}\left[(R_{jt}^w - R_{Jt}^w)\dot{u}\Big(\sum_{j=1}^{J} a_{j0} R_{jt}^w \Big) \right] = 0 \ \ \text{for} \ \ j = 1, \ldots, J-1, \tag{4.3}$$

and $a_{J0} = 1 - \sum_{j=1}^{J-1} a_{j0}$, where $R_{jt}^w = w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t$, $R_{Jt}^w = w_J^{\mathsf{T}}(X_{J,t-1})\mathbf{R}_t$ and a_{j0} is the j-th element of the J-dimensional vector $\mathbf{a}_0$.

We next propose a data-driven procedure for the choice of the model averaging weights $\mathbf{a}$. By replacing the unobservable quantity $w_j(X_{j,t-1})$ by its estimated value $\widehat{w}_j(X_{j,t-1})$ which is constructed in (2.6), we may estimate $\mathbf{a}_0 = (a_{10}, \ldots, a_{J0})^{\mathsf{T}}$ by $\widehat{\mathbf{a}} = (\widehat{a}_1, \ldots, \widehat{a}_J)^{\mathsf{T}}$, which is the solution to the following equations

$$\frac{1}{T} \sum_{t=1}^{T} (\widehat{R}_{jt}^w - \widehat{R}_{Jt}^w)\dot{u}\Big(\sum_{j=1}^{J} \widehat{a}_j \widehat{R}_{jt}^w \Big) = 0 \ \ \text{for} \ \ j = 1, \ldots, J-1, \tag{4.4}$$

and $\widehat{a}_J = 1 - \sum_{j=1}^{J-1} \widehat{a}_j$, where $\widehat{R}_{jt}^w = \widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t$.

We next study the asymptotic property for the estimator $\widehat{\mathbf{a}}$. As $\widehat{a}_J = 1 - \sum_{j=1}^{J-1} \widehat{a}_j$, it suffices to consider $\widehat{\mathbf{a}}^* \equiv (\widehat{a}_1, \ldots, \widehat{a}_{J-1})^{\mathsf{T}}$, the estimate of $\mathbf{a}_0^* \equiv (a_{10}, \ldots, a_{J-1,0})^{\mathsf{T}}$. Define $\eta_t = \dot{u}\big[\sum_{j=1}^{J} a_{j0} w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t \big]$, $\eta_t^* = \ddot{u}\big[\sum_{j=1}^{J} a_{j0} w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t \big]$, $\mathbf{R}_t^*(w) = (R_{1t}^w, \ldots, R_{J-1,t}^w)^{\mathsf{T}}$, $\mathbf{V}_t^* = \mathbf{R}_t^*(w) - R_{Jt}^w \mathbf{1}_{J-1}$ and $\mathbf{\Delta}_1 = \mathsf{E}\big[\eta_t^* \mathbf{V}_t^* (\mathbf{V}_t^*)^{\mathsf{T}} \big]$. For $j = 1, \ldots, J$, define $\varepsilon_{jt} = \dot{u}\big(w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t \big) = \dot{u}\big(R_{jt}^w \big)$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t} a_{10}, \ldots, \varepsilon_{Jt} a_{J0})^{\mathsf{T}}$

and $\mathbf{Q}_t = (Q_{1t}, \ldots, Q_{Jt})^\intercal$ with $Q_{jt} = \left\{ \mathsf{E}\left[\eta_s^* \mathbf{V}_s^* \mathbf{R}_s^\intercal \mathbf{W} \boldsymbol{\Lambda}_j^{-1}(X_{j,s-1}) | X_{j,s-1} = X_{j,t-1} \right] \right\} f_j(X_{j,t-1}) \mathbf{R}_t^*$ and

$$\mathbf{W} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \\ -1 & \cdots & -1 \end{pmatrix}$$

being an $n \times (n-1)$ matrix. Define

$$\boldsymbol{\Delta}_2 = \sum_{t=-\infty}^{\infty} \mathsf{Cov}(\mathbf{V}_0^* \eta_0 + \mathbf{Q}_0^\intercal \boldsymbol{\varepsilon}_0, \mathbf{V}_t^* \eta_t + \mathbf{Q}_t^\intercal \boldsymbol{\varepsilon}_t).$$

Throughout the paper, we assume that the mean of $\mathbf{V}_t^* \eta_t + \mathbf{Q}_t^\intercal \boldsymbol{\varepsilon}_t$ is zero. In the following theorem, we give the asymptotic distribution theory for $\widehat{\mathbf{a}}^*$ when both $J = J_0$ and $n = n_0$ are fixed positive integers, which is the case (i) discussed in Section 3.

THEOREM 4.1. *Suppose that the assumptions in Theorem 3.3(i) are satisfied and the matrix $\boldsymbol{\Delta}_1$ is non-singular. Then we have*

$$\sqrt{T} \left(\widehat{\mathbf{a}}^* - \mathbf{a}_0^* \right) \xrightarrow{d} \mathsf{N} \left(\mathbf{0}, \boldsymbol{\Delta}_1^{-1} \boldsymbol{\Delta}_2 \boldsymbol{\Delta}_1^{-1} \right). \tag{4.5}$$

We next deal with the case (ii) where $J = J_T$ is diverging with the sample size T whereas $n = n_0$ is fixed. Let $\boldsymbol{\Delta}_T = \boldsymbol{\Delta}_1^{-1} \boldsymbol{\Delta}_2 \boldsymbol{\Delta}_1^{-1}$, where we have used the subscript T to denote the dependence of the size of the matrix on T. As the number of the potential conditioning variables J_T tends to infinity, we cannot state the asymptotic normal distribution theory in the same way as in Theorem 4.1 above. As in Fan and Peng (2004), we let $\boldsymbol{\Psi}_T$ be a $J_* \times (J_T - 1)$ matrix with full row rank such that as $T \to \infty$, $\boldsymbol{\Psi}_T \boldsymbol{\Psi}_T^\intercal \to \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a $J_* \times J_*$ non-negative definite matrix with J_* being a fixed positive integer. The role of the matrix $\boldsymbol{\Psi}_T$ is to reduce the dimension from $(J_T - 1)$ to J_* in the derivation of the asymptotic normality, so it is only involved in the asymptotic analysis. If we are only interested in the asymptotic behavior for the first J_* components of $\widehat{\mathbf{a}}$, we may choose $\boldsymbol{\Psi}_T = \left[I_{J_*}, \ O_{J_* \times (J_T - J_*)} \right]$, where I_p is a $p \times p$ identity matrix and $O_{k \times j}$ is a $k \times j$ null matrix. We next state the asymptotic distribution theory for $\widehat{\mathbf{a}}^*$ when $J = J_T$ is diverging.

THEOREM 4.2. *Suppose that the assumptions in Theorem 3.3(ii) are satisfied, the matrix $\boldsymbol{\Delta}_1$ is non-singular and*

$$J_T^2 \left(h^2 + \sqrt{\frac{\log T}{Th}} \right) \to 0. \tag{4.6}$$

Then we have

$$\sqrt{T}\mathbf{\Psi}_T\mathbf{\Delta}_T^{-1/2}\left(\widehat{\mathbf{a}}^* - \mathbf{a}_0^*\right) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \mathbf{\Psi}\right). \tag{4.7}$$

The above theorem is similar to some results in the existing literature such as Theorem 2(ii) in Fan and Peng (2004) and Theorem 4.3 in Li, Linton and Lu (2015). The condition (4.6) implies that the dimension J_T should not diverge too fast to infinity, and (4.7) indicates that the convergence rate is $\sqrt{T/J_T}$ due to the diverging number of the conditioning variables.

Although $n = n_0$ is assumed to be fixed in Theorems 4.1 and 4.2 above, the developed asymptotic results can be generalized to the more general case of $n = n_T$ which is diverging with the sample size T (see the cases (iii) and (iv) in Section 3). By imposing some additional restrictions on n_T such as (3.7), the corresponding proofs in Appendix B need to be slightly modified with the role of Theorem 3.3 replaced by the uniform consistency result in (3.8). As the technical arguments are quite similar, we omit details here to save space.

5   Numerical studies

In the simulated examples, we set the number of assets under consideration for investment to be $n = 5$. This value of n is chosen primarily for convenience of computation. Computation procedures for larger values of n are exactly the same.

EXAMPLE 5.1. The time series of gross returns vectors $\mathbf{R}_t$ on the n assets are generated via the conditioning variables by the following regression:

$$\log(\mathbf{R}_t) = 0.06 * \mathbf{1}_n + \mathbf{A} * \log(\mathbf{X}_{t-1}) + \mathbf{e}_t, \tag{5.1}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones, $\mathbf{A}$ is an $n \times J$ full-rank matrix so generated such that the elements of $1000 * \mathbf{A}$ are random integers ranging between 1 and 30; $\mathbf{e}_t$ are i.i.d. random vectors distributed as $\mathbf{e}_t \sim \mathsf{N}(\mathbf{0}, 0.001 * \mathbf{I}_n)$ in which $\mathbf{I}_n$ is the $n \times n$ identity matrix; $\{\log(\mathbf{X}_t)\}$ is a J-dimensional AR(1) process generated as

$$\log(\mathbf{X}_t) = -0.01 * \mathbf{1}_J + \psi \log(\mathbf{X}_{t-1}) + \mathbf{u}_t, \tag{5.2}$$

in which $\psi = 0.9$ or $\psi = 0.4$, and $\mathbf{u}_t$ are i.i.d. random vectors generated from $\mathsf{N}(\mathbf{0}, 0.002 * \Sigma_\mathbf{u})$, where $\Sigma_\mathbf{u}$ is a $J \times J$ matrix with diagonal elements being 1 and off-diagonal elements being 0.4. Since the thus generated components in $\mathbf{u}_t$ are correlated, so are the components in $\mathbf{X}_t$, which will be

used as the conditioning variables. Such a design is aimed to mimic the real world situation where economic, finance and social indicators, which are often correlated with each other, are chosen as the conditioning variables. The dimension of $\mathbf{X}_t$ is set to satisfy $J = [0.5 * \sqrt{T}]$, where $[\cdot]$ denotes the operator that rounds a number to the nearest integer less than or equal to that number.

We use a CRRA utility function with $\gamma = 1$, 5, and 10 (Campbell (1996) found γ in range 2.7 to 21). For each $j = 1, \ldots, J$, $t = 1, \ldots, T$, and the observed value, $X_{j,t-1}$, of the conditioning variable in the previous time period $t - 1$, we calculate the j-th set of conditional optimal portfolio weights, $\widehat{w}_j(X_{j,t-1})$, by solving the nonparametric version of the conditioning equations, i.e., (2.6). Then by solving the equations in (4.4) with respect to a_j, we can obtain the joint optimal portfolio weights, $\widehat{w}_\mathbf{a}(\mathbf{X}_{t-1}) = \sum_{j=1}^{J} \widehat{a}_j \widehat{w}_j(X_{j,t-1})$, conditional on the values of all the conditioning variables in time period $t - 1$, where $\mathbf{X}_{t-1} = (X_{1,t-1}, \ldots, X_{J,t-1})^\mathsf{T}$. Note that in calculating the $\widehat{w}_j(X_{j,t-1})$ and $\widehat{a}_j$, we have imposed $\sum_{j=1}^{J} \widehat{w}_j(X_{j,t-1}) = 1$ and $\sum_{j=1}^{J} \widehat{a}_j = 1$ so that the budget constraint is satisfied.

We compare the single-period returns of portfolios constructed with weights calculated from the proposed semiparametric model averaging method (SMAM) and the unconditional parametric method (UPM) which solves for the weights that maximise the unconditional utility, i.e., $\frac{1}{T}\sum_{t=1}^{T} u(w^\mathsf{T}\mathbf{R}_t)$, subject to $w^\mathsf{T}\mathbf{1}_n = 1$. Table 5.1 reports the averages of the mean difference in returns (MDR) and averages of the mean difference in utilities (MDU) between the SMAM- and UPM-constructed portfolios:

$$\text{MDR} = \frac{1}{T}\sum_{t=1}^{T} \left(R_t^s - R_t^u\right), \quad \text{MDU} = \frac{1}{T}\sum_{t=1}^{T} \left[u(R_t^s) - u(R_t^u)\right],$$

where $R_t^s = \widehat{w}_\mathbf{a}^\mathsf{T}(\mathbf{X}_{t-1})\mathbf{R}_t$ and $R_t^u = \widehat{w}_u^\mathsf{T}\mathbf{R}_t$ with $\widehat{w}_\mathbf{a}(\cdot)$ and $\widehat{w}_u$ chosen by SMAM and UPM, respectively, and $u(\cdot)$ is a given utility function. Also reported in Table 5.1 are the averages of positive frequency (PF) of the SMAM, i.e., the frequency at which the return on the SMAM-constructed portfolio exceeds that of the UPM-constructed portfolio. These results are based on 100 independent samples of $T = 100$, 300, or 500 observations. The numbers in parentheses are the respective standard errors.

It can be seen from Table 5.1 that in most time periods, the return on the portfolio chosen by the SMAM is larger than the return on the portfolio chosen by the UPM. This is especially so when the sample size is relatively small. For example, when $\psi = 0.9$ and $\gamma = 5$, the average gain in choosing portfolios by the SMAM than by the UPM is an additional 2.86% return when $T = 100$, and this reduces to 1.25% when $T = 500$. As γ measures the level of risk aversion of an investor with a higher value representing less willingness for risk taking, the portfolio returns generally decrease as γ increases. Hence, we see a decreasing trend in the MDR values as γ increases in Table 5.1. Furthermore, when the persistence of the data generating process for the conditioning

variables increases (as represented by the increase in the autoregressive coefficient ψ), the average MDR values generally increase. This occurs because when the conditioning process is more persistent, its current values have more predicative ability for the its next-period value and thus the next-period asset returns.

TABLE 5.1. Averages of MDR, MDU, and PF for Example 5.1

ψ	γ \ T		$T = 100$	$T = 300$	$T = 500$
$\psi = 0.4$	$\gamma = 1$	MDR	0.0505(0.1682)	0.0474(0.1092)	0.0585(0.1120)
		MDU	0.0500(0.3751)	0.0031(0.1627)	0.0441(0.1450)
		PF	0.5239(0.0606)	0.5232(0.0340)	0.5255(0.0347)
	$\gamma = 5$	MDR	0.0143(0.0101)	0.0078(0.0040)	0.0066(0.0034)
		MDU	0.0028(0.0206)	0.0066(0.0310)	0.0011(0.0129)
		PF	0.5934(0.0451)	0.5734(0.0276)	0.5704(0.0192)
	$\gamma = 10$	MDR	0.0060(0.0038)	0.0038(0.0020)	0.0032(0.0013)
		MDU	0.0021(0.0010)	0.0014(0.0005)	0.0011(0.0004)
		PF	0.5821(0.0486)	0.5747(0.0284)	0.5676(0.0213)
$\psi = 0.9$	$\gamma = 1$	MDR	0.1045(0.3192)	0.0763(0.3311)	0.0937(0.3418)
		MDU	0.0958(0.2397)	0.0954(0.2086)	0.1842(0.3147)
		PF	0.5229(0.0557)	0.5199(0.0354)	0.5169(0.0382)
	$\gamma = 5$	MDR	0.0286(0.0153)	0.0160(0.0064)	0.0125(0.0047)
		MDU	0.0113(0.0111)	0.0059(0.0220)	0.0058(0.0086)
		PF	0.6388(0.0443)	0.6211(0.0267)	0.6059(0.0225)
	$\gamma = 10$	MDR	0.0160(0.0098)	0.0087(0.0034)	0.0065(0.0023)
		MDU	0.0052(0.0024)	0.0033(0.0010)	0.0028(0.0008)
		PF	0.6469(0.0445)	0.6247(0.0305)	0.6162(0.0229)

EXAMPLE 5.2. In this example, the gross returns vectors $\mathbf{R}_t$ are generated from a stationary VAR:

$$\log(\mathbf{R}_t) = 0.01 * \mathbf{1}_n + \mathbf{B} * \log(\mathbf{R}_{t-1}) + \mathbf{e}_t, \tag{5.3}$$

where $\mathbf{e}_t$ are generated in the same way as in Example 5.1, the AR coefficient matrix $\mathbf{B}$ is set as the transpose of $0.01 * \text{magic}(n)$, in which $\text{magic}(n)$ denotes the $n \times n$ magic matrix constructed from the integers 1 through n^2 with equal row and column sums.

The conditioning variables are taken as the lag-one and lag-two returns, i.e. $\mathbf{X}_{t-1} = (\mathbf{R}_{t-1}^{\mathsf{T}}, \mathbf{R}_{t-2}^{\mathsf{T}})^{\mathsf{T}}$. Hence, the number of conditioning variables is $J = 2n$. The results based on 100 independent samples of this example are given in Table 5.2. Similar findings can be obtained as those in Example 5.1.

TABLE 5.2. Averages of MDR, MDU, and PF for Example 5.2

γ \ T		$T = 100$	$T = 300$	$T = 500$
$\gamma = 1$	MDR	0.2187(0.3298)	0.2599(0.1431)	0.2448(0.1282)
	MDU	0.1518(0.2794)	0.0314(0.1257)	0.0232(0.0561)
	PF	0.5500(0.0590)	0.5373(0.0489)	0.5903(0.0403)
$\gamma = 5$	MDR	0.0444(0.0270)	0.0189(0.0079)	0.0165(0.0077)
	MDU	0.0140(0.0458)	0.0099(0.0033)	0.0075(0.0132)
	PF	0.6633(0.0430)	0.6274(0.0305)	0.6196(0.0215)
$\gamma = 10$	MDR	0.0233(0.0131)	0.0112(0.0049)	0.0097(0.0042)
	MDU	0.0088(0.0041)	0.0049(0.0017)	0.0044(0.0016)
	PF	0.6659(0.0475)	0.6287(0.0306)	0.6220(0.0253)

EXAMPLE 5.3. We have a sample of daily returns data on Dow Jones stocks over the period 2010-2014. We use a total of 1000 observations on the stock returns of the 30 companies comprising the Dow Jones index. The lag-1 returns on the 30 stocks are used as the conditioning variables. We apply a CRRA utility with $\gamma = 5$. The proposed data-driven semiparametric model averaging method is used to construct an optimal portfolio from the 30 stocks that gives the investor the maximum next-period utility. The portfolio is rebalanced in each time period to reflect latest market conditions. The returns on the selected portfolio are calculated and compared to those on the portfolio constructed from the unconditional parametric method. Figure 5.1 plots the returns on the portfolios constructed from the above two methods. It shows that in most of the time periods under consideration, the semiparametric model averaging gives a portfolio that provides higher returns. In some periods, the additional return is as high as 40%–60%. The mean difference between these returns over the sampling period is calculated as 3.2%.
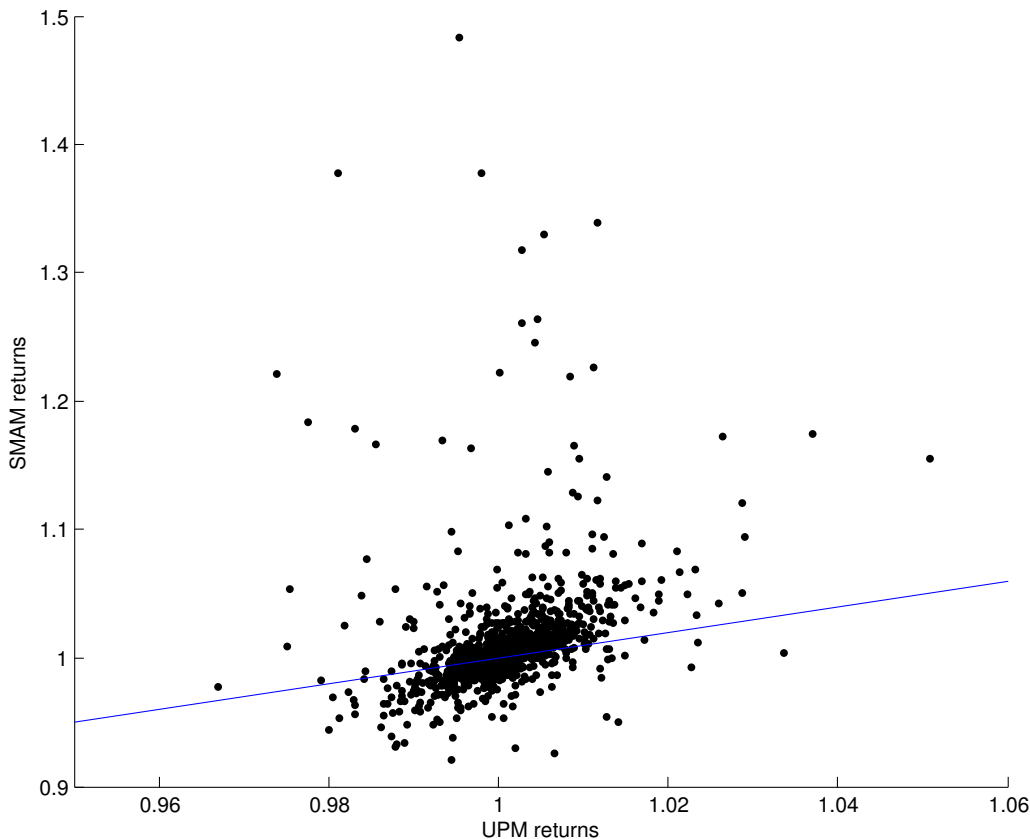
Figure 5.1. SMAM returns vs UPM returns: the returns on the portfolio constructed from the semiparametric model average method are on the vertical axis and those on the portfolio constructed from the unconditional parametric method are on the horizontal axis; the straight line is $y = x$.

6 Conclusions and extensions

In this paper, we have introduced a new data-driven method to estimate the dynamic portfolio choice with multiple conditioning variables, where the number of the conditioning variables can be either fixed or divergent with the sample size. Motivated by the MAMAR method proposed in Li, Linton and Lu (2015), we first solve a portfolio choice problem for each univariate conditioning variable, and then combine the portfolio weights from each of those "experts" through the model averaging approach. The optimal weights in the model averaging are determined by the semiparametric data-driven method introduced in Section 4. The asymptotic theorems demonstrate that the proposed dynamic portfolio choice approach can circumvent the curse of dimensionality issue, and the numerical

studies show that the proposed approach performs reasonably well in the finite sample case. In fact, this approach is quite common in the machine learning literature, see, for example, Györfi, Ottucsák and Urbán (2011). It is also possible to introduce constraints such as absence of short selling or position limits at each stage of our method at the cost of computational complexity.

A   Assumptions

We next list the regularity conditions which are used to prove the asymptotic results. Some of these conditions might not be the weakest possible and can be relaxed at the cost of more lengthy proofs.

ASSUMPTION 1. (i) The utility function $u(\cdot)$ is concave and has continuous derivatives up to the second order.

(ii) The optimal weight functions (for the scalar conditioning variable) $w_j^*(\cdot)$, $j = 1, \ldots, J$, have continuous derivatives up to the second order.

ASSUMPTION 2. The joint process of the conditioning variables $\{\mathbf{X}_t = (X_{1t}, \ldots, X_{Jt})^\intercal\}$ and asset returns $\{\mathbf{R}_t = (R_{1t}, \ldots, R_{nt})^\intercal\}$ is strictly stationary and α-mixing with the mixing coefficient decaying at a geometric rate, $\alpha_k \sim \gamma_0^k$, $0 < \gamma_0 < 1$.

ASSUMPTION 3. For $1 \le j \le J$, each component variable X_{jt} has a continuous marginal density function $f_j(\cdot)$ on a compact support denoted by $\mathcal{X}_j$. For all $t > 1$, the joint density function of $(\mathbf{X}_1, \mathbf{X}_t)$ exists and is uniformly bounded. There exists a $\delta > 0$ such that

$$\max_{1 \le j \le J} \max_{1 \le i,k \le n} \mathsf{E}\left[\left|R_{it}R_{kt}\ddot{u}\left(w_j^\intercal(X_{j,t-1})\mathbf{R}_t\right)\right|^{2+\delta} + \left|R_{it}\dot{u}\left(w_j^\intercal(X_{j,t-1})\mathbf{R}_t\right)\right|^{2+\delta}\right] < \infty.$$

The matrix

$$\mathsf{E}\left[\mathbf{R}_t^*(\mathbf{R}_t^*)^\intercal \ddot{u}\left(w_j^\intercal(x_j)\mathbf{R}_t\right) | X_{j,t-1} = x_j\right]$$

is non-singular uniformly for $x_j \in \mathcal{X}_j$, $j = 1, \ldots, J$, where $\mathbf{R}_t^*$ is defined in Section 3.

ASSUMPTION 4. The kernel function $K(\cdot)$ is positive, Lipschitz continuous and symmetric about zero with a compact support, and $\int K(z)dz = 1$.

ASSUMPTION 5. The bandwidth h satisfies $h \to 0$,

$$Th^4 = o(1) \quad \text{and} \quad \frac{T^{1-1/(2+\delta)}h}{\log T} \to \infty.$$

ASSUMPTION 5′. The bandwidth h satisfies $h \to 0$,

$$Th^4 = o(1) \quad \text{and} \quad \frac{T^{1-1/(2+\delta)}h}{J_T^{1/(2+\delta)} \log T} \to \infty.$$

The above assumptions are mild and justifiable. Some of the assumptions are similar to those in Brandt (1999). Under some conditions, Assumption 1(ii) follows from Assumption 1(i). For example when $u(x) = x - (\gamma/2)x^2$, it suffices that the conditional mean and conditional covariance matrix of returns are twice continuously differentiable. We impose in Assumption 2 the stationarity and mixing dependence condition on the joint processes of the returns of the risky assets and the conditioning variables. The methodology and theory developed in the present paper are also applicable to the more general dependence structure, say the near epoch dependent process (Li, Lu and Linton, 2012). To facilitate our proofs, we assume that the mixing coefficients decay at a geometric rate, which can be relaxed to a polynomial rate at the cost of more lengthy proofs. The bandwidth conditions in Assumptions 5 and 5′ indicate that there is a trade-off between the moment conditions and the bandwidth restriction. And the condition $Th^4 = o(1)$ shows that certain under-smoothing is needed in the asymptotic analysis, which is not uncommon in semiparametric estimation.

B   Proofs of the theoretical results

We next give the proofs of the theoretical results stated in Sections 3 and 4. In this appendix, we let C be a positive constant whose value may change from line to line.

PROOF OF THEOREM 3.1. By the definition of $\widehat{w}_j^*(x_j) = [\widehat{w}_{1j}(x_j), \ldots, \widehat{w}_{n_0-1,j}(x_j)]^{\mathsf{T}}$ or $\widehat{w}_j(x_j) = [\widehat{w}_{1j}(x_j), \ldots, \widehat{w}_{n_0 j}(x_j)]^{\mathsf{T}}$, we have

$$\frac{1}{Th} \sum_{t=1}^{T} (R_{it} - R_{nt})\dot{u}\left(\widehat{w}_j^{\mathsf{T}}(x_j)\mathbf{R}_t\right) K\left(\frac{X_{j,t-1} - x_j}{h}\right) = 0 \tag{B.1}$$

for $i = 1, \ldots, n_0 - 1$ and $j = 1, \ldots, J_0$. By Assumption 1 and using the Taylor's expansion for $\dot{u}(\cdot)$,

$$\dot{u}\left(\widehat{w}_j^{\mathsf{T}}(x_j)\mathbf{R}_t\right) = \dot{u}\left(w_j^{\mathsf{T}}(x_j)\mathbf{R}_t\right) + \ddot{u}\left(w_\diamond^{\mathsf{T}}(x_j)\mathbf{R}_t\right) \left\{(\mathbf{R}_t^*)^{\mathsf{T}} \left[\widehat{w}_j^*(x_j) - w_j^*(x_j)\right]\right\},$$

where $w_\diamond(x_j)$ lies between $\widehat{w}_j(x_j)$ and $w_j(x_j)$, and $w_j^*(x_j) = [w_{1j}(x_j), \ldots, w_{n_0-1,j}(x_j)]^{\mathsf{T}}$. Then we may prove that

$$\widehat{w}_j^*(x_j) - w_j^*(x_j) = \mathcal{A}_{nj}^{-1}(x_j)\mathcal{B}_{nj}(x_j), \tag{B.2}$$

18

for $j = 1, \cdots, J_0$, where

$$\mathcal{A}_{nj}(x_j) = \frac{1}{Th} \sum_{t=1}^{T} \mathbf{R}_t^* (\mathbf{R}_t^*)^{\intercal} \ddot{u} \left(w_{\diamond}^{\intercal}(x_j) \mathbf{R}_t \right) K \left(\frac{X_{j,t-1} - x_j}{h} \right),$$

$$\mathcal{B}_{nj}(x_j) = \frac{1}{Th} \sum_{t=1}^{T} \mathbf{R}_t^* \dot{u} \left(w_j^{\intercal}(x_j) \mathbf{R}_t \right) K \left(\frac{X_{j,t-1} - x_j}{h} \right).$$

By Assumptions 2–5 in Appendix A and following the standard argument in nonparametric kernel-based smoothing in time series (c.f., Robinson, 1983), we can show that

$$\mathcal{A}_{nj}(x_j) = \mathbf{\Lambda}_j(x_j) + o_P(1) \tag{B.3}$$

when $n = n_0$ is fixed and $\widehat{w}_j(x_j)$ is sufficiently close to $w_j(x_j)$, where $\mathbf{\Lambda}_j(x_j)$ is defined in Section 3. The convergence in (B.3) holds uniformly for $x_j \in \mathcal{X}_j$ and $j = 1, \ldots, J_0$ (c.f., the proof of Theorem 3.3 below). On the other hand, we recall that $\mathbf{Z}_{jt}(x_j) = \mathbf{R}_t^* \dot{u} \left(w_j^{\intercal}(X_{j,t-1}) \mathbf{R}_t \right) K \left(\frac{X_{j,t-1} - x_j}{h} \right)$. By Assumptions 1(i)(ii) and the Taylor's expansion for $\dot{u} \left(w_j^{\intercal}(\cdot) \mathbf{R}_t \right)$, we may show that

$$\mathcal{B}_{nj}(x_j) = \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_j) + O_P(h^2). \tag{B.4}$$

Noting that $Th^4 = o(1)$ in Assumption 5 and by (B.2)–(B.4),

$$\sqrt{Th} \left[\widehat{w}_j^*(x_j) - w_j^*(x_j) \right] = \mathbf{\Lambda}_j^{-1}(x_j) \cdot \frac{1}{\sqrt{Th}} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_j) + o_P(1). \tag{B.5}$$

Then, using the central limit theorem for the stationary α-mixing sequence (e.g., Section 2.6.4 in Fan and Yao, 2003), we can complete the proof of (3.2) in Theorem 3.1(i).

As in Section 3, let

$$\mathbf{W}_{jt}(x_j) = \mathbf{\Lambda}_j^{-1}(x_j) \mathbf{Z}_{jt}(x_j), \quad \mathbf{W}_t(\mathbf{x}|\mathbf{a}) = \sum_{j=1}^{J_0} a_j \mathbf{W}_{jt}(x_j).$$

By (B.5) and the definitions of $\widehat{w}_{\mathbf{a}}^*(\mathbf{x})$ and $w_{\mathbf{a}}^*(\mathbf{x})$, we have

$$\sqrt{Th} \left[\widehat{w}_{\mathbf{a}}^*(\mathbf{x}) - w_{\mathbf{a}}^*(\mathbf{x}) \right] = \frac{1}{\sqrt{Th}} \sum_{t=1}^{T} \mathbf{W}_t(\mathbf{x}|\mathbf{a}) + o_P(1). \tag{B.6}$$

Using (B.6), we can readily prove (3.3) in Theorem 3.1(ii). ∎

19

PROOF OF THEOREM 3.2. The proof of this theorem is similar to the proof of Theorem 3.1 above. The only difference that the stronger bandwidth condition in Assumption 5′ is needed when we prove (B.3) uniformly for $x_j \in \mathcal{X}_j$ and $j = 1, \ldots, J_T$. ∎

PROOF OF THEOREM 3.3. We only consider the proof of (3.5) for the case when $J = J_T$ is diverging, as the proof for the case of $J = J_0$ is similar and simpler. Noting that $\widehat{w}_{n_0 j}(x_j) = 1 - \sum_{i=1}^{n_0-1} \widehat{w}_{ij}(x_j)$ and using (B.2) and (B.3) in the proof of Theorem 3.1, we only need to show that

$$\max_{1 \leq j \leq J_T} \sup_{x_j \in \mathcal{X}_j} \left\| \frac{1}{Th} \sum_{t=1}^{T} \mathbf{R}_t^* \dot{u} \left(w_j^{\mathsf{T}}(x_j) \mathbf{R}_t \right) K\left(\frac{X_{j,t-1} - x_j}{h} \right) \right\| = O_P\left(h^2 + \sqrt{\log T/(Th)} \right), \qquad \text{(B.7)}$$

as $\boldsymbol{\Lambda}_j(x_j)$ is nonsingular uniformly for $x_j \in \mathcal{X}_j$, $1 \leq j \leq J_T$ (see Assumption 3). Note that the convergence result in (B.4) can be strengthened from the point-wise convergence to the uniform convergence over $x_j \in \mathcal{X}_j$, $1 \leq j \leq J_T$. Hence, in order to prove (B.7), we only need to show that

$$\max_{1 \leq j \leq J_T} \sup_{x_j \in \mathcal{X}_j} \left\| \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_j) \right\| = O_P\left(\sqrt{\log T/(Th)} \right), \qquad \text{(B.8)}$$

where $\mathbf{Z}_{jt}(x_j)$ is defined in the proof of Theorem 3.1.

For notational simplicity, denote $\xi_T = \left(\frac{\log T}{Th} \right)^{1/2}$. The main idea of proving (B.8) is to consider covering the compact support $\mathcal{X}_j$ by a finite number of disjoint subsets $\mathcal{X}_j(k)$ which are centered at x_{jk} with radius $r_T = \xi_T h^2$, $k = 1, \ldots, \mathcal{N}_j$. It is easy to show that

$$\max_{1 \leq j \leq J_T} \mathcal{N}_j = O(r_T^{-1}) = O(\xi_T^{-1} h^{-2})$$

and

$$\begin{aligned}
\max_{1 \leq j \leq J_T} \sup_{x_j \in \mathcal{X}_j} \left\| \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_j) \right\| &\leq \max_{1 \leq j \leq J_T} \max_{1 \leq k \leq \mathcal{N}_j} \left\| \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_{jk}) \right\| + \\
&\qquad \max_{1 \leq j \leq J_T} \max_{1 \leq k \leq \mathcal{N}_j} \sup_{x_j \in \mathcal{X}_j(k)} \left\| \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_j) - \frac{1}{Th} \sum_{t=1}^{T} \mathbf{Z}_{jt}(x_{jk}) \right\| \\
&\equiv \Pi_{T1} + \Pi_{T2}.
\end{aligned} \qquad \text{(B.9)}$$

By the continuity condition on $K(\cdot)$ in Assumption 4 and using the definition of r_T, we readily have

$$\Pi_{T2} = O_P\left(\frac{r_T}{h^2} \right) = O_P(\xi_T). \qquad \text{(B.10)}$$

20

For Π_{T1}, we apply the truncation technique and the Bernstein-type inequality for the α-mixing dependent random variables which can be found in Bosq (1998) and Fan and Yao (2003) to obtain the convergence rate. Let $M_T = M_1(TJ_T)^{1/(2+\delta)}$,

$$\overline{\mathbf{Z}}_{jt}(x_{jk}) = \mathbf{Z}_{jt}(x_{jk}) \cdot \mathrm{I}\left\{\left\|\mathbf{R}_t^* \dot{u}\left(R_{jt}^w\right)\right\| \leq M_T\right\}$$

and

$$\widetilde{\mathbf{Z}}_{jt}(x_{jk}) = \mathbf{Z}_{jt}(x_{jk}) \cdot \mathrm{I}\left\{\left\|\mathbf{R}_t^* \dot{u}\left(R_{jt}^w\right)\right\| > M_T\right\},$$

where $\mathrm{I}\{\cdot\}$ is an indicator function and $R_{jt}^w = w_j^\intercal(X_{j,t-1})\mathbf{R}_t$. Then we have

$$
\begin{aligned}
\Pi_{T1} &\leq \max_{1 \leq j \leq J_T} \max_{1 \leq k \leq \mathcal{N}_j} \left\|\frac{1}{Th} \sum_{t=1}^T \left\{\overline{\mathbf{Z}}_{jt}(x_{jk}) - \mathsf{E}\left[\overline{\mathbf{Z}}_{ij}(x_{jk})\right]\right\}\right\| + \\
&\quad \max_{1 \leq j \leq J_T} \max_{1 \leq k \leq \mathcal{N}_j} \left\|\frac{1}{Th} \sum_{i=1}^n \left\{\widetilde{\mathbf{Z}}_{jt}(x_{jk}) - \mathsf{E}\left[\widetilde{\mathbf{Z}}_{jt}(x_{jk})\right]\right\}\right\| \\
&\equiv \Pi_{T3} + \Pi_{T4}.
\end{aligned}
\tag{B.11}
$$

For $M_2 > 0$ and any $\epsilon > 0$, by the moment condition in Assumption 3 and the Markov inequality,

$$
\begin{aligned}
\mathsf{P}\left(\Pi_{T4} > M_2 \xi_T\right) &\leq \mathsf{P}\left(\max_{1 \leq j \leq J_T} \max_{1 \leq k \leq \mathcal{N}_j} \max_{1 \leq t \leq T} \left\|\widetilde{\mathbf{Z}}_{jt}(x_{jk})\right\| > M_2 \xi_T\right) \\
&\leq \sum_{j=1}^{J_T} \sum_{t=1}^T \mathsf{P}\left(\left\|\mathbf{R}_t^* \dot{u}\left(R_{jt}^w\right)\right\| > M_T\right) \\
&\leq M_1^{-(2+\delta)} \cdot \max_{1 \leq j \leq J_T} \mathsf{E}\left[\left\|\mathbf{R}_t^* \dot{u}\left(R_{jt}^w\right)\right\|^{2+\delta}\right] < \epsilon,
\end{aligned}
$$

if we choose

$$M_1 > \left\{\max_{1 \leq j \leq J_T} \mathsf{E}\left[\left\|\mathbf{R}_t^* \dot{u}(R_{jt}^w)\right\|^{2+\delta}\right]\right\}^{1/(2+\delta)} \epsilon^{-1/(2+\delta)}.$$

Then, by letting ϵ be arbitrarily small, we can show that

$$\Pi_{T4} = O_P(\xi_T). \tag{B.12}$$

On the other hand, note that

$$\left\|\overline{\mathbf{Z}}_{jt}(x_{jk}) - \mathsf{E}\left[\overline{\mathbf{Z}}_{jt}(x_{jk})\right]\right\| \leq C_0 M_T \tag{B.13}$$

and

$$\mathsf{Var}\left[\overline{\mathbf{Z}}_{jt}(x_{jk})\right] \leq C_0 h, \tag{B.14}$$

21

where C_0 is a positive constant. By (B.13), (B.14) and Theorem 1.3(2) in Bosq (1998) with $p = [(M_2 M_T \xi_T/4)^{-1}]$ which tends to infinity by (3.6), we have

$$
\begin{aligned}
\mathsf{P}(\Pi_{T3} > M_2 \xi_T) &= \mathsf{P}\left(\max_{1 \le j \le J_T} \max_{1 \le k \le \mathcal{N}_j} \left\| \frac{1}{Th} \sum_{t=1}^{T} \left\{ \overline{\mathbf{Z}}_{jt}(x_{jk}) - \mathsf{E}\left[\overline{\mathbf{Z}}_{ij}(x_{jk})\right] \right\} \right\| > M_2 \xi_T \right) \\
&= \sum_{j=1}^{J_T} \mathcal{N}_j \left(4 \exp\left\{ \frac{-q M_2^2 \xi_T^2}{4 C_0 M_2 M_T \xi_T/h + 16 C_0/(ph))} \right\} + 22 \left[1 + 4 C_0 M_T/(M_2 h \xi_T) \right] q \gamma_0^p \right) \\
&\le C \sum_{j=1}^{J_T} \mathcal{N}_j \left[\exp\left\{ -M_2^{1/2} \log T \right\} + T M_T^2 \gamma_0^p \right] = o(1),
\end{aligned}
$$

where M_2 is chosen sufficiently large and $q = T/(2p)$. Hence we have

$$
\Pi_{T3} = O_P(\xi_T). \tag{B.15}
$$

In view of (B.10)–(B.12) and (B.15), we have shown (B.8), completing the proof of Theorem 3.3. ∎

PROOF OF THEOREM 4.1. Recall that

$$
\begin{aligned}
\widehat{\mathbf{a}}^* &= (\widehat{a}_1, \ldots, \widehat{a}_{J-1})^{\mathsf{T}}, \quad \mathbf{a}_0^* = (a_{10}, \ldots, a_{J-1,0})^{\mathsf{T}}, \\
\mathbf{R}_t(w) &= (R_{1t}^w, \ldots, R_{Jt}^w)^{\mathsf{T}}, \quad \mathbf{R}_t^*(w) = \left(R_{1t}^w, \ldots, R_{J-1,t}^w \right)^{\mathsf{T}}, \\
\widehat{\mathbf{R}}_t(w) &= \left(\widehat{R}_{1t}^w, \ldots, \widehat{R}_{Jt}^w \right)^{\mathsf{T}}, \quad \widehat{\mathbf{R}}_t^*(w) = \left(\widehat{R}_{1t}^w, \ldots, \widehat{R}_{J-1,t}^w \right)^{\mathsf{T}}.
\end{aligned}
$$

As $\widehat{a}_{J_0} = 1 - \sum_{j=1}^{J_0-1} \widehat{a}_j$ and $a_{J_00} = 1 - \sum_{j=1}^{J_0-1} a_{j0}$, by Theorem 3.3(i), Assumption 1(i) and the Taylor's expansion for $\dot{u}(\cdot)$, we may show that

$$
\begin{aligned}
&\dot{u}\Big[\sum_{j=1}^{J_0} \widehat{a}_j \widehat{w}_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] - \dot{u}\Big[\sum_{j=1}^{J_0} a_{j0} \widehat{w}_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] \\
&= \ddot{u}\Big[\sum_{j=1}^{J_0} a_{j0} w_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] \sum_{j=1}^{J_0} (\widehat{a}_j - a_{j0}) \, \widehat{w}_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t + O_P(\|\widehat{\mathbf{a}} - \mathbf{a}_0\|^2) \\
&= \ddot{u}\Big[\sum_{j=1}^{J_0} a_{j0} w_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] (\widehat{\mathbf{a}} - \mathbf{a}_0)^{\mathsf{T}} \widehat{\mathbf{R}}_t(w) + O_P(\|\widehat{\mathbf{a}} - \mathbf{a}_0\|^2) \\
&= \ddot{u}\Big[\sum_{j=1}^{J_0} a_{j0} w_j^{\mathsf{T}}(X_{j,t-1}) \mathbf{R}_t \Big] (\widehat{\mathbf{a}}^* - \mathbf{a}_0^*)^{\mathsf{T}} \left[\widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0 t}^w \mathbf{1}_{J_0-1} \right] + O_P(\|\widehat{\mathbf{a}}^* - \mathbf{a}_0^*\|^2)
\end{aligned}
$$

and

$$\dot{u}\big[\sum_{j=1}^{J_0} a_{j0}\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] - \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$$

$$= \ddot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]\sum_{j=1}^{J_0} a_{j0}\big[\widehat{w}_j(X_{j,t-1}) - w_j(X_{j,t-1})\big]^{\mathsf{T}}\mathbf{R}_t + O_P\left(h^4 + \frac{\log T}{Th}\right)$$

$$= \ddot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]\mathbf{a}_0^{\mathsf{T}}\big[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\big] + O_P\left(h^4 + \frac{\log T}{Th}\right).$$

Hence, we have

$$\dot{u}\big[\sum_{j=1}^{J_0} \widehat{a}_j\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] = \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] + \dot{u}\big[\sum_{j=1}^{J_0} \widehat{a}_j\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] - \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$$

$$= \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] + \dot{u}\big[\sum_{j=1}^{J_0} \widehat{a}_j\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] - \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$$

$$+\dot{u}\big[\sum_{j=1}^{J_0} a_{j0}\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big] - \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$$

$$= \eta_t + \eta_t^*\big[\widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0t}^w\mathbf{1}_{J_0-1}\big]^{\mathsf{T}}(\widehat{\mathbf{a}}^* - \mathbf{a}_0^*) + \eta_t^*\big[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\big]^{\mathsf{T}}\mathbf{a}_0$$

$$+O_P\left(h^4 + \frac{\log T}{Th}\right) + O_P(\|\widehat{\mathbf{a}}^* - \mathbf{a}_0^*\|^2), \tag{B.16}$$

where $\eta_t = \dot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$ and $\eta_t^* = \ddot{u}\big[\sum_{j=1}^{J_0} a_{j0}w_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$.

By (4.4) and (B.16), we have

$$\mathbf{0} = \frac{1}{T}\sum_{t=1}^{T}\big[\widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0t}^w\mathbf{1}_{J_0-1}\big]\dot{u}\big[\sum_{j=1}^{J} \widehat{a}_j\widehat{w}_j^{\mathsf{T}}(X_{j,t-1})\mathbf{R}_t\big]$$

$$= \frac{1}{T}\sum_{t=1}^{T}\big[\widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0t}^w\mathbf{1}_{J_0-1}\big]\Big\{\eta_t + \eta_t^*\big[\widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0t}^w\mathbf{1}_{J_0-1}\big]^{\mathsf{T}}(\widehat{\mathbf{a}}^* - \mathbf{a}_0^*)$$

$$+\eta_t^*\big[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\big]^{\mathsf{T}}\mathbf{a}_0\Big\} + O_P\left(h^4 + \frac{\log T}{Th}\right) + O_P(\|\widehat{\mathbf{a}}^* - \mathbf{a}_0^*\|^2). \tag{B.17}$$

By (B.17), we readily have

$$\sqrt{T}\,(\widehat{\mathbf{a}}^* - \mathbf{a}_0^*) \overset{P}{\sim} \left[\frac{1}{T}\sum_{t=1}^{T}\eta_t^*\widehat{\mathbf{V}}_t^*(\widehat{\mathbf{V}}_t^*)^{\mathsf{T}}\right]^{-1}\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widehat{\mathbf{V}}_t^*\eta_t + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widehat{\mathbf{V}}_t^*\eta_t^*\big[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\big]^{\mathsf{T}}\mathbf{a}_0\right\},$$
$$\tag{B.18}$$

where $\widehat{\mathbf{V}}_t^* = \widehat{\mathbf{R}}_t^*(w) - \widehat{R}_{J_0 t}^w \mathbf{1}_{J_0-1}$ and $\alpha_n \overset{P}{\sim} \beta_n$ denotes that $\alpha_n/\beta_n = 1 + o_P(1)$.

By Theorem 3.3 and the law of larger numbers, we readily have

$$\frac{1}{T} \sum_{t=1}^T \eta_t^* \widehat{\mathbf{V}}_t^* (\widehat{\mathbf{V}}_t^*)^\intercal = \frac{1}{T} \sum_{t=1}^T \eta_t^* \mathbf{V}_t^* (\mathbf{V}_t^*)^\intercal + o_P(1) = \boldsymbol{\Delta}_1 + o_P(1), \tag{B.19}$$

where $\boldsymbol{\Delta}_1$ is defined in Section 4. Note that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\mathbf{V}}_t^* \eta_t = \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t^* \eta_t + \frac{1}{\sqrt{T}} \sum_{t=1}^T \left(\widehat{\mathbf{V}}_t^* - \mathbf{V}_t^* \right) \eta_t.$$

By Assumptions 2 and 3 and following the argument in the proof of Lemma B.3 in Li, Linton and Lu (2015), we may show that the second term on the right hand side of (B.19) is asymptotically negligible. Hence, we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\mathbf{V}}_t^* \eta_t \overset{P}{\sim} \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t^* \eta_t. \tag{B.20}$$

We next consider $\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\mathbf{V}}_t^* \eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right]^\intercal \mathbf{a}_0$. It is easy to see that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\mathbf{V}}_t^* \eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right]^\intercal \mathbf{a}_0 \overset{P}{\sim} \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t^* \eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right]^\intercal \mathbf{a}_0 \tag{B.21}$$

by using Theorem 3.3. Let $\mathbf{W}$ be an $n_0 \times (n_0 - 1)$ matrix which is defined by

$$\mathbf{W} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \\ -1 & \cdots & -1 \end{pmatrix}.$$

It is easy to show that for any $j = 1, \ldots, J_0$ and $x_i \in \mathcal{X}_j$,

$$\widehat{w}_j(x_j) - w_j(x_j) = \mathbf{W} \left[\widehat{w}_j^*(x_j) - w_j^*(x_j) \right]. \tag{B.22}$$

Hence, by (B.22) and using the argument in the proofs of Theorems 3.1 and 3.3, we may show that

$$
\begin{aligned}
\widehat{R}_{jt}^w - R_{jt}^w &= \left[\widehat{w}_j(X_{j,t-1}) - w_j(X_{j,t-1})\right]^\mathsf{T} \mathbf{R}_t \\
&= \left[\widehat{w}_j^*(X_{j,t-1}) - w_j^*(X_{j,t-1})\right]^\mathsf{T} \mathbf{W}^\mathsf{T}\mathbf{R}_t \\
&\stackrel{P}{\sim} \mathbf{R}_t^\mathsf{T}\mathbf{W}\boldsymbol{\Lambda}_j^{-1}(X_{j,t-1}) \cdot \left[\frac{1}{Th}\sum_{s=1}^T \mathbf{Z}_{js}(X_{j,t-1})\right] \\
&= \mathbf{R}_t^\mathsf{T}\mathbf{W}\boldsymbol{\Lambda}_j^{-1}(X_{j,t-1}) \cdot \left[\frac{1}{Th}\sum_{s=1}^T \mathbf{R}_s^* \dot{u}\left(w_j^\mathsf{T}(X_{j,s-1})\mathbf{R}_s\right) K\left(\frac{X_{j,s-1} - X_{j,t-1}}{h}\right)\right] \\
&= \mathbf{R}_t^\mathsf{T}\mathbf{W}\boldsymbol{\Lambda}_j^{-1}(X_{j,t-1}) \cdot \left[\frac{1}{Th}\sum_{s=1}^T \mathbf{R}_s^*\varepsilon_{js} K\left(\frac{X_{j,s-1} - X_{j,t-1}}{h}\right)\right], \qquad \text{(B.23)}
\end{aligned}
$$

where $\varepsilon_{js} = \dot{u}\left(w_j^\mathsf{T}(X_{j,s-1})\mathbf{R}_s\right) = \dot{u}\left(R_{js}^w\right)$ and $\mathbf{Z}_{js}(\cdot)$ is defined in the proof of Theorem 3.1. By (B.23), we readily have

$$
\left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\right]^\mathsf{T} \mathbf{a}_0 = \mathbf{R}_t^\mathsf{T}\mathbf{W} \cdot \left[\frac{1}{Th}\sum_{s=1}^T \sum_{j=1}^{J_0} \boldsymbol{\Lambda}_j^{-1}(X_{j,t-1})\mathbf{R}_s^*\varepsilon_{js}a_{j0}K\left(\frac{X_{j,s-1} - X_{j,t-1}}{h}\right)\right],
$$

which indicates that

$$
\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbf{V}_t^*\eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\right]^\mathsf{T} \mathbf{a}_0 \\
&\stackrel{P}{\sim} \frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbf{V}_t^*\eta_t^*\mathbf{R}_t^\mathsf{T}\mathbf{W} \cdot \left[\frac{1}{Th}\sum_{s=1}^T \sum_{j=1}^{J_0} \boldsymbol{\Lambda}_j^{-1}(X_{j,t-1})\mathbf{R}_s^*\varepsilon_{js}a_{j0}K\left(\frac{X_{j,s-1} - X_{j,t-1}}{h}\right)\right] \\
&= \frac{1}{\sqrt{T}}\sum_{s=1}^T \sum_{j=1}^{J_0} \varepsilon_{js}a_{j0}\left[\frac{1}{Th}\sum_{t=1}^T \eta_t^*\mathbf{V}_t^*\mathbf{R}_t^\mathsf{T}\mathbf{W}\boldsymbol{\Lambda}_j^{-1}(X_{j,t-1})K\left(\frac{X_{j,s-1} - X_{j,t-1}}{h}\right)\right]\mathbf{R}_s^* \\
&\stackrel{P}{\sim} \frac{1}{\sqrt{T}}\sum_{s=1}^T \sum_{j=1}^{J_0} \varepsilon_{js}a_{j0}Q_{js}, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(B.24)}
\end{aligned}
$$

where

$$
Q_{js} = \left\{ \mathsf{E}\left[\eta_t^*\mathbf{V}_t^*\mathbf{R}_t^\mathsf{T}\mathbf{W}\boldsymbol{\Lambda}_j^{-1}(X_{j,t-1})|X_{j,t-1} = X_{j,s-1}\right]\right\}f_j(X_{j,s-1})\mathbf{R}_s^*.
$$

Recall that $\boldsymbol{\varepsilon}_t = \left[\varepsilon_{1t}a_{10}, \ldots, \varepsilon_{J_0t}a_{J_00}\right]^\mathsf{T}$ and $\mathbf{Q}_t = (Q_{1t}, \ldots, Q_{J_0t})^\mathsf{T}$. Then we have

$$
\frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbf{V}_t^*\eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w)\right]^\mathsf{T} \mathbf{a}_0 \stackrel{P}{\sim} \frac{1}{\sqrt{T}}\sum_{t=1}^T \mathbf{Q}_t^\mathsf{T}\boldsymbol{\varepsilon}_t. \qquad\text{(B.25)}
$$

25

By (B.20), (B.21) and (B.25), we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\mathbf{V}}_t^* \eta_t + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{V}_t^* \eta_t^* \left[\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right]^{\intercal} \mathbf{a}_0 \overset{P}{\sim} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\mathbf{V}_t^* \eta_t + \mathbf{Q}_t^{\intercal} \boldsymbol{\varepsilon}_t \right). \tag{B.26}$$

By the central limit theorem for the α-mixing sequence, we can prove that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\mathbf{V}_t^* \eta_t + \mathbf{Q}_t^{\intercal} \boldsymbol{\varepsilon}_t \right) \overset{d}{\to} \mathsf{N}\left(\mathbf{0}, \boldsymbol{\Delta}_2 \right) \tag{B.27}$$

Then, we can complete the proof of Theorem 4.1 by (B.18)–(B.21), (B.26) and (B.27). ∎

PROOF OF THEOREM 4.2. The main idea in this proof is similar to the proof of Theorem 4.1 above with some modifications. Hence, we next only sketch the proof.

Following the proof of (B.18) and using the condition (4.6), we may show that

$$\begin{aligned}
\sqrt{T} \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-1/2} \left(\widehat{\mathbf{a}}^* - \mathbf{a}_0^* \right) = \ & \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-\frac{1}{2}} \left[\frac{1}{T} \sum_{t=1}^{T} \eta_t^* \widehat{\mathbf{V}}_t^* (\widehat{\mathbf{V}}_t^*)^{\intercal} \right]^{-1} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\mathbf{V}}_t^* \eta_t \right. \\
& \left. + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\mathbf{V}}_t^* \eta_t^* \left(\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right)^{\intercal} \mathbf{a}_0 \right].
\end{aligned} \tag{B.28}$$

Note that (B.19) still holds by using Theorem 3.3(ii) and (4.6). Hence, by (B.28), we have

$$\sqrt{T} \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-1/2} \left(\widehat{\mathbf{a}}^* - \mathbf{a}_0^* \right) \overset{P}{\sim} \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-1/2} \boldsymbol{\Delta}_1^{-1} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\mathbf{V}}_t^* \eta_t + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\mathbf{V}}_t^* \eta_t^* \left(\widehat{\mathbf{R}}_t(w) - \mathbf{R}_t(w) \right)^{\intercal} \mathbf{a}_0 \right]. \tag{B.29}$$

Furthermore, using the argument in proving (B.26), we obtain

$$\sqrt{T} \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-1/2} \left(\widehat{\mathbf{a}} - \mathbf{a}_0 \right) \overset{P}{\sim} \boldsymbol{\Psi}_T \boldsymbol{\Delta}_T^{-1/2} \boldsymbol{\Delta}_1^{-1} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\mathbf{V}_t^* \eta_t + \mathbf{Q}_t^{\intercal} \boldsymbol{\varepsilon}_t \right) \right]. \tag{B.30}$$

By the central limit theorem and using the condition that $\boldsymbol{\Psi}_T \boldsymbol{\Psi}_T^{\intercal} \to \boldsymbol{\Psi}$, we can complete the proof of Theorem 4.2. ∎

References

[1] Back, K. E., 2010. *Asset Pricing and Portfolio Choice Theory*. Oxford University Press.

[2] Bosq, D., 1998. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction* (2nd Edition), Springer.

[3] Brandt, M. W., 1999. Estimating portfolio and consumption choice: a conditional Euler equations approach. *Journal of Finance* **54**, 1609–1645.

[4] Brandt, M. W., 2010. Portfolio choice problems. *Handbook of Financial Econometrics, Volume 1: Tools and Techniques* (Editors: Aït-Sahalia Y. and Hansen L. P.), 269-336.

[5] Campbell, J. Y., 1996. Understanding risk and return. *Journal of Political Economy* **104**, 298–345.

[6] Fama, E. F., 1970. Multiperiod consumption-investment decisions. *American Economic Review* **60**, 163–174.

[7] Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 928–961.

[8] Fan, J., Yao, Q., 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

[9] Györfi, L., Ottucsák, G., Urbán, A., 2011. Empirical log-optimal portfolio selections: a survey. *World Scientific Review*, 79-116.

[10] Hansen, B. E., 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24**, 726–748.

[11] Hansen, L. P., Singleton, K. J., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* **50**, 1269–1286.

[12] Kristensen, D., 2009. Uniform convergence rates of kernel estimators with heterogenous dependent data. *Econometric Theory* **25**, 1433–1445.

[13] Li, D., Linton, O., Lu, Z., 2015. A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187**, 345–357.

[14] Li, D., Lu, Z., Linton, O., 2012. Local linear fitting under near epoch dependence: uniform consistency with convergence rates. *Econometric Theory* **28**, 935–958.

[15] Markowitz, H. M., 1952. Portfolio selection. *Journal of Finance* **7**, 77–91.

[16] Merton, R. C., 1969. Lifetime portfolio selection under uncertainty: the continuous time case. *Review of Economics and Statistics* **51**, 247–257.

[17] Robinson, P. M., 1983. Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185–207.

[18] Samuelson, P.A., 1969. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* **51**, 239–246

[19] Sharpe,W., 1963. A simplified model for portfolio analysis. *Management Science* **9**, 277–293.