Econometrica, Vol. 80, No. 2 (March, 2012), 713–754

EFFICIENT SEMIPARAMETRIC ESTIMATION OF THE FAMA–FRENCH MODEL AND EXTENSIONS

BY GREGORY CONNOR, MATTHIAS HAGMANN, AND OLIVER LINTON[1]

This paper develops a new estimation procedure for characteristic-based factor models of stock returns. We treat the factor model as a weighted additive nonparametric regression model, with the factor returns serving as time-varying weights and a set of univariate nonparametric functions relating security characteristic to the associated factor betas. We use a time-series and cross-sectional pooled weighted additive nonparametric regression methodology to simultaneously estimate the factor returns and characteristic-beta functions. By avoiding the curse of dimensionality, our methodology allows for a larger number of factors than existing semiparametric methods. We apply the technique to the three-factor Fama–French model, Carhart's four-factor extension of it that adds a momentum factor, and a five-factor extension that adds an own-volatility factor. We find that momentum and own-volatility factors are at least as important, if not more important, than size and value in explaining equity return comovements. We test the multifactor beta pricing theory against a general alternative using a new nonparametric test.

KEYWORDS: Additive models, arbitrage pricing theory, characteristic-based factor model, kernel estimation, nonparametric regression.

1. INTRODUCTION

INDIVIDUAL STOCK RETURNS have strong common movements, and these common movements can be related to individual security characteristics such as market capitalization and book-to-price ratios. Rosenberg (1974) developed a factor model of stock returns in which the factor betas of stocks are linear functions of observable security characteristics. Rosenberg's approach requires the strong assumption of linearity. Fama and French (1993) used portfolio grouping to estimate a characteristic-based factor model without assuming linearity. They estimated a three-factor model: a market factor, a size factor, and a value factor. The market factor return is proxied by the excess return to a value-weighted market index. The size factor return is proxied by the difference in return between a portfolio of low-capitalization stocks and a portfolio of high-capitalization stocks, adjusted to have roughly equal book-to-price ratios. The value factor is proxied by the difference in return between a portfolio of high book-to-price stocks and a portfolio of low book-to-price stocks, adjusted to have roughly equal capitalization. Using these factor returns, the factor betas are estimated via time-series regression.

Connor and Linton (2007) used a semiparametric method which combines elements of the Rosenberg and Fama–French approaches. They described a

[1] We would like to thank Harold Uhlig, Enno Mammen, and three referees for helpful comments. We also would like to thank the Science Foundation of Ireland (Connor), ESRC (Linton), and ERC (Linton) for financial support.

DOI: 10.3982/ECTA7432

characteristic-based factor model like Rosenberg's, but replaced Rosenberg's assumption that factor betas are linear in the characteristics with an assumption that factor betas are smooth nonlinear functions of the characteristics. In a model with two characteristics—size and value—plus a market factor, they formed a grid of equally spaced characteristic pairs. They used multivariate kernel methods to form factor-mimicking portfolios for the characteristic pairs from each point on the grid. Then they estimated factor returns and factor betas simultaneously using bilinear regression applied to the set of factor-mimicking portfolio returns.

A weakness of the Connor–Linton methodology is the reliance on multivariate kernel methods to create factor-mimicking portfolios. These multivariate kernel methods severely restrict the number of factors which can be estimated well using their technique, due to the curse of dimensionality (Stone (1980)). The same problem appears in a different guise in the Fama–French methodology. To create their size and value factor returns, Fama and French double-sort assets into size and value categories. Adding a third characteristic with this method requires triple sorting and adding a fourth requires quadruple sorting; as in Connor and Linton, the method quickly becomes unreliable for typical sample sizes and more than two characteristic-based factors.

In this paper, we develop a new estimation methodology that does not require any portfolio grouping or multivariate kernels. Instead, we estimate the factor returns and the characteristic-beta functions using weighted additive nonparametric regression. This relies on the fact that in each time period, the characteristic-based factor model proposed in Connor and Linton is a weighted additive sum of univariate characteristic-based functions. The nonparametric part of the estimation problem is made univariate by decomposing the full problem into an iterative set of subproblems in each characteristic singly, a standard trick in weighted additive nonparametric regression. We modify the standard weighted additive nonparametric regression methodology to account for our model's feature that the weights vary each time period while the characteristic-beta functions stay constant. The theoretical basis for our estimation method has been developed in a series of papers: Mammen, Linton, and Nielsen (1999), Linton, Nielsen, and van de Geer (2003), and Linton and Mammen (2005, 2008). See also Carrasco, Florens, and Renault (2006) for a review of the theory and a discussion of applications to other areas in economics.

Our model falls into the class of semiparametric panel data models for large cross section and long time series. There has been some work on semiparametric models for panel data (see, for example, Kyriazidou (1997)), and nonparametric additive models (see, for example, Porter (1996) and, more recently, Mammen, Støve, and Tjøstheim (2009)). Most of this work is in the context of short time series. More recently, there has been work on panel data with large cross-section and time-series dimension, especially in finance, where the data sets can be large along both dimensions, and in macroeconomics, where there are cross-sectional panels of many related series (such as business conditions survey data) with quite long time-series length. Some recent papers include Phillips and Moon (1999), Bai and Ng (2002), Bai (2003, 2004), and Pesaran (2006). These authors addressed a variety of issues including nonstationarity, estimation of unobserved factors, and model selection. They all worked with essentially parametric models. Our semiparametric model takes full advantage of the information provided by large cross-section and time-series dimensions. We establish pointwise asymptotic normality of the functional components of our model at what appears to be an optimal rate. We also establish the asymptotic normality of our estimated factors. We allow for general temporal and cross-sectional dependence in the error terms.

Our model allows for any number of factors with no theoretical loss of efficiency, and we exploit this in our application. In addition to the market, size, and value factors of the standard Fama–French model, we add a momentum factor, as suggested by Jagadeesh and Titman (1993) and Carhart (1997), and an own-volatility factor, a choice influenced by the recent work of Goyal and Santa Clara (2003) and Ang, Hodrick, Xing, and Zhang (2006, 2009). This reflects the feature that our methodology allows us to estimate a model with more factors. We find that the two added factors—momentum and volatility—are as important or more important than size and value in explaining equity return comovements. Hence, the improved data efficiency of our new method has real empirical value.

We develop a new nonparametric test for multifactor pricing models as part of our estimation methodology. To implement the test, we assume that mispricing is a smooth multivariate function of observable security characteristics. We estimate this mispricing function simultaneously with the factor model of returns and test whether the mispricing function is the null function. We find that the five-factor model does a good job of explaining asset return premia; the α function differs only negligibly from a null function, at least for the four security characteristics that we consider.

In the working paper version of this paper, we evaluated various time-series models for the risk factors. We estimated vector autoregressions both for the levels of the factor returns and the factor returns squared (to explore factor volatility dynamics).

We proceed as follows: Section 2 presents the model. Section 3 describes the estimation algorithm in the balanced and unbalanced panel case. Section 4 develops the distribution theory. Section 5 presents an empirical application to the cross section of monthly U.S. stock returns. Section 6 summarizes the findings and concludes.

2. THE MODEL

We assume that there is a large number of securities, indexed by $i = 1, \dots, n$. Asset excess returns (returns minus the risk-free rate) are observed for a number of time periods $t = 1, \dots, T$. We assume that the following characteristic-based factor model generates excess returns:

$$y_{it} = f_{ut} + \sum_{j=1}^{J} g_j(X_{ji}) f_{jt} + \varepsilon_{it}, \tag{1}$$

where y_{it} is the excess return to security i at time t, f_{ut} and f_{jt} are the factor returns, $g_j(X_{ji})$ denotes the factor betas, X_{ji} are observable security characteristics, and ε_{it} are the mean-zero asset-specific returns. The factor returns f_{jt} are linked to the security characteristics by the characteristic-beta functions $g_j(\cdot)$, which map characteristics to the associated factor betas. We assume that each $g_j(\cdot)$ is a smooth time-invariant function of a continuously distributed characteristic j, but we do not assume a particular functional form. This is the same type of factor model used by Connor and Linton (2007). To simplify the exposition, we assume that the characteristics X_{ji} are time invariant. We discuss later on the case where characteristics are allowed to vary over time. We also consider the case where some of the characteristics are discrete (like industry membership), in which case it is appropriate to replace the corresponding unknown function g_j by a known linear function of the discrete variable.

The market factor f_{ut} captures that part of common return not related to the security characteristics; all assets have unit beta to this factor. This factor captures the tendency of all equities to move together, irrespective of their characteristics. It is a common element in panel data models; see Hsiao (2003, Section 3.6.2). In applications to returns data, it is convenient to exclude own-effect intercept terms from (1), since they provide little benefit in terms of explanatory power and necessitate an additional time-series estimation step; see Connor and Korajczyk (1988, 1993) and Connor and Linton (2007).

In a straightforward extension of the model, we consider the case in which the unit constant is replaced with a set of industry-based zero–one dummy variables, essentially allowing f_{ut} to differ across industries. This is done to capture any industry-specific factors in returns. We also allow for the case where one or more of the factors is directly observed and need not be estimated.

Note that for fixed t, equation (1) constitutes a weighted additive nonparametric regression model for panel data, where the factor returns f_{jt} are "parametric weights" and the characteristic-beta functions $g_j(\cdot)$ are univariate nonparametric functions. Some discussion of additive nonparametric models can be found in Linton and Nielsen (1995). The situation here is somewhat nonstandard, since the same regression equation (1) holds each time period, with parametric weights varying each time period and the characteristic-beta functions being time invariant. We extend the weighted nonparametric regression methodology to account for this feature of time-varying weights in a pooled time-series, cross-sectional model.

Our model can be thought of as a special case of the usual statistical factor model

$$y_{it} = \sum_{j=1}^{J} \beta_{ij} f_{jt} + \varepsilon_{it}, \tag{2}$$

where the factor loadings β_{ij} are unrestricted (Ross (1976)). Connor and Koracyzk (1993) developed the asymptotic principal component method for estimation of the factors in the case where the cross section is large but the time series is fixed. Recent work of Bai and Ng (2002) and Bai (2003, 2004) provided analysis for this method for the case where both n and T are large. Bai (2003) established pointwise asymptotic normality for estimates of the factors (at rate $\sqrt{n}$) and the loadings (at rate $\sqrt{T}$) under weak assumptions regarding cross-sectional and temporal dependence.[2]

2.1. *Factor Scale Identification Conditions*

In the case in which both characteristic-beta functions and factor returns are estimated from the data, there is an obvious scale indeterminacy in the model. We make our identifying restrictions on the beta functions rather than on the factors; in particular, we impose that for each factor, the cross-sectional average beta equals 0 and the cross-sectional variance of beta equals 1, that is, $E^* g_j = 0$ and $E^* g_j^2 = 1$, where E^* denotes expectation with respect to some distribution P_j^* (i.e., $E^* g_j = \int g_j(x)\, dP_j^*(x)$), which could be the objective covariate probability distribution or another related distribution. Note that this does not restrict the return model since the additive semiparametric model (1) is invariant to this rescaling. The choice of distribution to use in the normalization affects the interpretation of the factors. The condition $E^* g_j^2 = 1$ sets the magnitude of factor return j; the conditions $E^* g_j = 0$ affect the interpretation of the intercept. If we use the population distribution, then $E^* g_j = 0$ means that the intercept can be interpreted as the return to the average asset in the infinite population of assets; if we use a capitalization-weighted population distribution, then $E^* g_j = 0$ means that the intercept can be interpreted as the return to the capitalization-weighted average asset.

2.2. *Additive Nonparametric Mispricing Functions*

A central concern in the asset pricing literature is the determination of the expected returns on assets and their relationship to the risk exposures of the assets. In this subsection, it is important to note that the time $t-1$ information

[2] Bai (2003) assumed that the loadings β_{ij} are fixed in repeated samples, but treats f_{jt} as random.

set of investors includes the characteristics X. Taking investors' expectation of excess returns y_{it} using (1) gives

$$E[y_{it}] = E[f_{ut}] + \sum_{j=1}^{J} g_j(X_{ji}) E[f_{jt}], \tag{3}$$

which is the standard multifactor asset pricing model: expected excess returns are linear in factor betas. Hence our model as developed so far imposes the standard multifactor pricing condition on expected excess returns.

Our methodology provides a new asset pricing test against a general nonparametric pricing alternative. Fama and French (1993) created factor-mimicking portfolios from size- and value-sorted portfolios, and then estimated characteristic-related mispricing based on a finer grid of value- and size-sorted portfolios. This two-stage procedure leaves open the question whether there is a hidden "identification condition" when using the same characteristics to create mimicking portfolios and to test for mispricing.

Adapting the Fama–French mispricing test to our additive nonparametric framework generates an explicit identification condition. The characteristic-mispricing functions are only identified up to an orthogonality condition relative to the characteristic-beta functions. This is because the same characteristics are used to identify the factor risk premia and factor model mispricing.

We assume that there are mispricing inefficiencies given by a smooth additive univariate nonparametric function $\alpha_j(X_{ij})$ using the same characteristics X_{ij} as in the factor model.[3] The return generating process becomes

$$y_{it} = f_{ut} + \sum_{j=1}^{J} \alpha_j(X_{ij}) + g_j(X_{ji}) f_{jt} + \varepsilon_{it}. \tag{4}$$

For the functions $\alpha_j(Z_{ij})$ to be identified, we must impose

$$E[\alpha_j(X_{ij})] = 0, \tag{5}$$

$$E[\alpha_j(X_{ij}) g_j(X_{ji})] = 0. \tag{6}$$

The mean-zero condition (5) is standard in additive nonparametric models, so that the intercept f_{ut} can be identified. The condition (6) is necessary for the risk premia of each factor return to be identified. To see why this is so, suppose that we relax the identification condition. Then for any constant a, we can replace $\alpha_j(X_{ij})$ with $\alpha_j^*(X_{ij}) = \alpha_j(X_{ij}) + a g_j(X_{ji})$ and $f_{jt}^* = f_{jt} - a$, and the fit of the model is exactly the same. This indeterminacy is only eliminated by

[3] It is possible to include additional nonparametric functions based on other observed variates strictly exogenous relative to y_{it}; this does not require any additional identification conditions beyond the mean-zero condition.

imposing (6). The intuition for the condition is clear: mean return which is in the linear span of the characteristic-beta function must be treated as "factor risk premia" rather than "mispricing."

It is interesting to note that an identification condition analogous to (6) is hidden implicitly in the seminal results of Fama and French (1993). Consider, for example, their Table 6, which shows the results from time-series regressions of each of a collection of 25 size- and value-sorted portfolio returns on an intercept, and the three Fama–French factor portfolios RMRF, SMB, and HML defined in Section 5.4. Letting r_h denote the return to the characteristic-sorted portfolio h for $h = 1, 25$, and letting r_{RMRF}, r_{SMB}, r_{HML} denote the factor portfolio returns, each time-series regression has the form

$$r_h = \alpha_h + \beta_{h,\mathrm{RMRF}} r_{\mathrm{RMRF}} + \beta_{h,\mathrm{SMB}} r_{\mathrm{SMB}} + \beta_{h,\mathrm{HML}} r_{\mathrm{HML}} + \varepsilon_h. \tag{7}$$

The hidden identification conditions imposed on α_h become clear when we note that the three-factor portfolio returns are themselves linear combinations of the characteristic-sorted portfolios: $r_{\mathrm{RMRF}} = \sum_{h=1}^{25} w_{h,\mathrm{RMRF}} r_h$, $r_{\mathrm{SMB}} = \sum_{h=1}^{25} w_{h,\mathrm{SMB}} r_h$, and $r_{\mathrm{HML}} = \sum_{h=1}^{25} w_{h,\mathrm{HML}} r_h$ (see Fama and French (1993) for details). Substituting into (7), taking weighted sums, imposing that ε_h are conditionally mean zero, and rearranging gives

$$\begin{aligned}\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \alpha_h = {} & \left(1 - \sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \beta_{h,\mathrm{RMRF}}\right) E[r_{\mathrm{RMRF}}] \\ & + \left(\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \beta_{h,\mathrm{SMB}}\right) E[r_{\mathrm{SMB}}] \\ & + \left(\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \beta_{h,\mathrm{HML}}\right) E[r_{\mathrm{HML}}],\end{aligned}$$

and exactly analogous equations (not shown to preserve space) for $\sum_{h=1}^{25} w_{h,\mathrm{SMB}} \alpha_h$ and $\sum_{h=1}^{25} w_{h,\mathrm{HML}} \alpha_h$. To simplify, consider the canonical special case in which each cross-weighted exposure equals 0 and each direct-weighted exposure equals 1: $\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \beta_{h,\mathrm{RMRF}} = 1$, $\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \beta_{h,\mathrm{SMB}} = 0$, and so on. Then the identification conditions imposed on the α functions simplify to

$$\sum_{h=1}^{25} w_{h,\mathrm{RMRF}} \alpha_h = \sum_{h=1}^{25} w_{h,\mathrm{SMB}} \alpha_h = \sum_{h=1}^{25} w_{h,HML} \alpha_h = 0.$$

3. ESTIMATION STRATEGY

For simplicity of exposition, we focus on the case in which all the factors are estimated and there are no mispricing functions $\alpha_j(Z_{ij})$ included in the factor model. Connor and Linton (2007) proposed to estimate the period-by-period conditional expectation of y_{it} given the characteristics $X_{1i}, \dots, X_{Ji}$ at a grid of points, and then to estimate the factors and the beta functions at the same grid of points using an iterative algorithm based on bilinear regression. This approach works well enough when the cross section is very large and when J is small, like two in their case. However, it is inefficient in general and works poorly in practice when J is larger than 2. For this reason, we develop an alternative estimation strategy that makes efficient use of the restrictions embodied in (1).

To describe the statistical properties of our estimators, we make some assumptions about the data generating process. For notational convenience, we treat in detail the case of a *fully balanced panel*, where the set of assets and the characteristics of each asset do not vary through time. (In Section 3.2.3 below we describe the modifications necessary for the case of an unbalanced panel.)

3.1. *Identification*

We first establish the identification of the quantities of interest through a population least squares criterion. This is one way to define the quantities f and g consistent with (1); it has the advantage of usually implying an efficient estimation procedure under independent and identically distributed (i.i.d.) normal error terms. The solution to this population problem is characterized by first-order conditions; to derive estimators, we mimic this population first-order condition by a sample equivalent. For clarity, we just treat the case where all the factors are unknown: if some factors are known, then they do not need to be chosen in the optimization below.

Consider the population criterion

$$Q_T(f, g) = \frac{1}{T}\sum_{t=1}^{T} E\left[\left\{y_{it} - f_{ut} - \sum_{j=1}^{J} g_j(X_{ji}) f_{jt}\right\}^2\right]. \tag{8}$$

In this criterion, the expectation is taken over the distribution of returns and characteristics, treating the factors as fixed parameters that are to be chosen (we are thinking of the factors as an exogenous stochastic process). Under some conditions, a limiting (as $T \to \infty$) criterion function $Q(f, g)$ may exist, but we do not require this. We minimize $Q_T(f, g)$ with respect to the factors f (which contain f_{ut} and f_{jt} for all j, t) and the functions $g = (g_1, \dots, g_J)$ subject to the identifying restrictions $E^* g_j = 0$ and $E^* g_j^2 = 1$.

This minimization problem can be characterized by a set of first-order conditions for f and g. For expositional purposes, we divide the problem in two: an equation characterizing f given known g and an equation characterizing g given f.

3.1.1. *The Factor Returns*

First we solve for the minimization of (8) over f_{ut} and f_{jt} for all j, t given $g(\cdot)$ is known. Note that if the population of assets is treated as fixed rather than random, then (8) simply amounts to a collection of unrelated cross-sectional regression problems, one per time period. In this case, the solution to the minimization problem is obviously period-by-period least squares regression. We now show that this intuition extends to our environment with a random population of assets rather than a fixed cross section.

Taking the first derivatives of (8) with respect to f_{ut} and f_{jt}, and setting to zero, the first-order conditions are (for each $t = 1, \dots, T$)

$$E\left[\left\{y_{it} - f_{ut} - \sum_{j=1}^{J} g_j(X_{ji}) f_{jt}\right\}\right] = 0, \tag{9}$$

$$E\left[\left\{y_{it} - f_{ut} - \sum_{k=1}^{J} g_k(X_{ki}) f_{kt}\right\} g_j(X_{ji})\right] = 0, \quad j = 1, \dots, J. \tag{10}$$

These equations are linear in f given g. This delivers a linear system of $J + 1$ equations in $J + 1$ unknowns for each time period t. Letting $f_t = [f_{ut}, f_{1t}, \dots, f_{Jt}]^\top$, $y_t = [y_{1t}, \dots, y_{nt}]^\top$, and $G(X_i) = [1, g_1(X_{1i}), \dots, g_J(X_{Ji})]^\top$, we have $E[G(X_i)G(X_i)^\top] f_t = E[G(X_i) y_t]$. It follows that there is a unique solution given by

$$f_t = E[G(X_i)G(X_i)^\top]^{-1} E[G(X_i) y_t],$$

provided $A = E[G(X_i)G(X_i)^\top]$ is nonsingular, which we assume to be the case.

3.1.2. *The Characteristic-Beta Functions*

Next we turn to the characterization of g given f. Consider the Gateaux pointwise derivative of (8) at $g_j(\cdot)$ in the direction of the function $\psi(\cdot)$:

$$\frac{\partial}{\partial \epsilon}\left(\frac{1}{T}\sum_{t=1}^{T} E\left[\left\{y_{it} - f_{ut} - \{g_j(X_{ji}) + \epsilon \psi(X_{ji})\} f_{jt} - \sum_{k \neq j} g_k(X_{ki}) f_{kt}\right\}^2\right]\right)_{\epsilon=0}.$$

Taking $\psi(\cdot)$ to be the point mass at x_j, we obtain a first-order condition defining the criterion-minimizing function $g_j(x)$ at the value x_j:

$$\begin{aligned}\text{(11)}\qquad \frac{1}{T}\sum_{t=1}^{T} f_{jt} E[y_{it}|X_{ji}=x_j] &= \frac{1}{T}\sum_{t=1}^{T} f_{jt} f_{ut} + g_j(x_j)\frac{1}{T}\sum_{t=1}^{T} f_{jt}^2 \\ &\quad + \frac{1}{T}\sum_{t=1}^{T}\sum_{k\neq j} f_{jt} f_{kt} E[g_k(X_{ki})|X_{ji}=x_j].\end{aligned}$$

Doing this for each $j=1,\dots,J$, we obtain a system of implicit linear equations for g given f, that is, a system of integral equations (of type 2) in the functional parameter g; see Mammen, Linton, and Nielsen (1999) and Linton and Mammen (2005). We next argue that a unique solution to these equations exists. Define

$$m_j(x_j) = \frac{\sum_{t=1}^{T} f_{jt} E[(y_{it} - f_{ut})|X_{ji}=x_j]}{\sum_{t=1}^{T} f_{jt}^2},$$

$$\mathcal{H}_{jk}(x_j, x_k) = \frac{p_{j,k}(x_j, x_k)}{p_j(x_j) p_k(x_k)}, \qquad \beta_{jk} = \frac{\sum_{t=1}^{T} f_{jt} f_{kt}}{\sum_{t=1}^{T} f_{jt}^2},$$

where $p_{j,k}$ is the joint density of (X_{ji}, X_{ki}). We drop the dependency on T in the notation for simplicity. Then for $j=1,\dots,J$, we have the system of linear equations in the space $L_2(p)$,

$$\text{(12)}\qquad m_j = g_j + \sum_{k\neq j} \beta_{jk}\mathcal{H}_j g_k,$$

where $(\mathcal{H}_j g_k)(x_j) = \int \mathcal{H}_{jk}(x_j, x_k) g_k(x_k) p_k(x_k)\, dx_k$. In the absence of β_{jk} and with a different m_j, (12) is the system of equations that define the additive nonparametric regression model (Mammen, Linton, and Nielsen (1999)). We can write the system of equations (12) as

$$\begin{pmatrix} I & \beta_{12}\mathcal{H}_1 & \cdots & \beta_{1J}\mathcal{H}_1 \\ \beta_{21}\mathcal{H}_2 & I & \beta_{23}\mathcal{H}_2 & \cdots \\ \vdots & & \ddots & \\ \beta_{J1}\mathcal{H}_J & \cdots & & I \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_J \end{pmatrix} = \mathcal{H}(\beta) g = m = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_J \end{pmatrix}$$

(cf. Hastie and Tibshirani (1990, equations 5.5 and 5.6)). The question is whether a unique solution to this system of equations exists such that we can write $g = \mathcal{H}(\beta)^{-1}m$. The associated system with $\beta_{ij} = 1$ for all i, j has been well studied in the literature and the system $\mathcal{H}g = m$ has a unique solution.

Consider the case $J = 2$. By substitution, we obtain the equations

$$(I - \beta_{12}\beta_{21}\mathcal{H}_1\mathcal{H}_2)g_1 = m_1 + \beta_{12}\mathcal{H}_1 m_2, \tag{13}$$

$$(I - \beta_{12}\beta_{21}\mathcal{H}_2\mathcal{H}_1)g_2 = m_2 + \beta_{21}\mathcal{H}_2 m_1. \tag{14}$$

For a solution to these equations to exist, it suffices that $I - \beta_{12}\beta_{21}\mathcal{H}_1\mathcal{H}_2$ and $I - \beta_{12}\beta_{21}\mathcal{H}_2\mathcal{H}_1$ be invertible. Suppose that the Hilbert–Schmidt condition holds:

$$\int \frac{p_{k,j}(x, x')^2}{p_j(x)p_k(x')}\, dx\, dx' < \infty \quad \text{for all } j, k. \tag{15}$$

This is satisfied under our Assumption A2 below. Then it holds that the operator norm of the composition of the operators satisfies $\|\mathcal{H}_1\mathcal{H}_2\| < 1$ and $\|\mathcal{H}_2\mathcal{H}_1\| < 1$. Then, since

$$\beta_{12}\beta_{21} = \frac{\left(\sum_{t=1}^{T} f_{1t}f_{2t}\right)^2}{\sum_{t=1}^{T} f_{1t}^2 \sum_{t=1}^{T} f_{2t}^2} \in [0, 1],$$

it follows that $\|\beta_{12}\beta_{21}\mathcal{H}_1\mathcal{H}_2\| < 1$ and $\|\beta_{12}\beta_{21}\mathcal{H}_2\mathcal{H}_1\| < 1$ so that $I - \beta_{12}\beta_{21} \times \mathcal{H}_1\mathcal{H}_2$ and $I - \beta_{12}\beta_{21}\mathcal{H}_2\mathcal{H}_1$ are invertible and there exists a unique solution to (13), $g_1 = (I - \beta_{12}\beta_{21}\mathcal{H}_1\mathcal{H}_2)^{-1}(m_1 + \beta_{12}\mathcal{H}_1 m_2)$, and to (14), $g_2 = (I - \beta_{12}\beta_{21}\mathcal{H}_2\mathcal{H}_1)^{-1}(m_2 + \beta_{21}\mathcal{H}_2 m_1)$.[4]

We now turn to the full problem where the factors and the characteristic functions are unknown. In this case, the unrestricted solution to (8) is not unique and the set of solutions forms a vector space (Breiman and Friedman (1985)). To proceed, we impose the identification conditions on $g_j(\cdot)$. These restrictions can be imposed by considering the constrained optimization problem and manipulating the first-order condition of the associated Lagrangian as in

[4] Furthermore, we can write

$$g_1 = \sum_{k=0}^{\infty} (\beta_{12}\beta_{21}\mathcal{H}_1\mathcal{H}_2)^k (m_1 + \beta_{12}\mathcal{H}_1 m_2).$$

The sum converges geometrically fast, which suggests that iterative methods (which amount to taking a finite truncation of the infinite sum) will converge rapidly to the solution and be independent of starting values (cf. Hastie and Tibshirani (1990, pp. 118–120)).

Lewbel and Linton (2007). An equivalent approach is to take any unrestricted solution f or g and replace $g_j(x_j)$ by

$$\overline{g}_j(x_j) = \frac{g_j(x_j) - \int g_j(x_j)\,dP_j^*(x_j)}{\sqrt{\int g_j^2(x_j)\,dP_j^*(x_j)}}, \tag{16}$$

where P_j^* is the probability distribution associated with the characteristic j.

3.2. *The Estimation Method*

Motivated by the above characterization, we next define our estimation method. The strategy involves solving empirical versions of (9), (10), and (11), which we do in a sequential manner.

The conditional expectations in (11) are unknown and we replace them by consistent estimators. We use the boundary adjusted cross-sectional kernel regression estimate. Thus we estimate the conditional expectation $E[y_{it}|X_{ji} = x]$ by

$$\widehat{E}[y_{it}|X_{ji} = x] = \frac{\sum_{i=1}^{n} K_h(X_{ji}, x) y_{it}}{\sum_{i=1}^{n} K_h(X_{ji}, x)},$$

where for each x in the support of X, $K_h(x, y) = K_h^x(x - y)$ for some kernel K^x such that $K_h^x(u) = h^{-1}K^x(h^{-1}u)$ and $K_h^x(u) = K_h(u)$ for all x in the interior of the support of X_{ji}. Here $K_h(\cdot) = K(\cdot/h)/h$ and K is a kernel while h is a bandwidth. We assume that each covariate is supported on $[\underline{x}, \overline{x}]$ for some known $\underline{x}, \overline{x}$ and that the covariate density is bounded away from zero on this support. We need to make a boundary adjustment to the kernel K to ensure that the bias is the same magnitude everywhere.

3.2.1. *Estimation of Factor Returns and Characteristic-Beta Functions*

We replace the unknown quantities in A, b_t, and equation (11) by estimated values, denoted by carets, and iterate between the factor return f and characteristic-beta function $g(\cdot)$ estimation problems. The solution for f depends on $g(\cdot)$, and the solution for $g_j(\cdot)$ depends both on f and $g_k(\cdot)$, $k \neq j$. We use the Gauss–Seidel iteration to reconcile these component solutions. We give the estimation algorithm below:

Step 1. Let $\widehat{f}^{[0]}$ and $\widehat{g}^{[0]}(\cdot)$ be initial estimates.

Step 2. Then let, for all x,

$$\widehat{g}_j^{[i+1]}(x) = \frac{\sum_{t=1}^{T} \widehat{f}_{jt}^{[i]}\big(\widehat{E}[y_{it}|X_{ji}=x] - \widehat{f}_{ut}^{[i]}\big)}{\sum_{t=1}^{T} \widehat{f}_{jt}^{[i]2}} - \frac{\sum_{t=1}^{T}\sum_{k>j} \widehat{f}_{jt}^{[i]}\widehat{f}_{kt}^{[i]}\widehat{E}\big[g_k^{[i]}(X_{ki})|X_{ji}=x\big]}{\sum_{t=1}^{T} \widehat{f}_{jt}^{[i]2}} - \frac{\sum_{t=1}^{T}\sum_{k<j} \widehat{f}_{jt}^{[i]}\widehat{f}_{kt}^{[i]}\widehat{E}\big[g_k^{[i+1]}(X_{ki})|X_{ji}=x\big]}{\sum_{t=1}^{T} \widehat{f}_{jt}^{[i]2}}, \tag{17}$$

$$\widehat{g}_j^{[i+1]}(x) = \frac{\widehat{g}_j^{[i+1]}(x) - \int \widehat{g}_j^{[i+1]}(x)\,dP_j^*(x)}{\sqrt{\int \widehat{g}_j^{[i+1]}(x)^2\,dP_j^*(x)}}. \tag{18}$$

Step 3. Then given estimates $\widehat{g}_j^{[i]}(X_{ji})$ from the previous iteration on g given f, we compute the least squares regression for each t:

$$\widehat{f}_t^{[i+1]} = \left[\sum_{i=1}^{n} \widehat{G}^{[i]}(X_i)\widehat{G}^{[i]}(X_i)^\top\right]^{-1} \sum_{i=1}^{n} \widehat{G}^{[i]}(X_i)y_{it}, \tag{19}$$

where $\widehat{G}^{[i]}(X_i) = [1, \widehat{g}_1^{[i]}(X_{1i}), \dots, \widehat{g}_J^{[i]}(X_{Ji})]^\top$.

Step 4. Continue until some stopping criterion is met.

The convergence properties of this algorithm are not studied here, but our discussions above, the arguments of Mammen, Linton, and Nielsen (1999), and our experience in the application suggests that the method is likely to converge rapidly. If one has consistent initial starting values, then one can stop after a finite number of iterations, which is what we do in practice. We turn to this issue next.

3.2.2. *Initial Consistent Estimators*

We describe a general way to obtain consistent starting values. Connor and Linton (2007) already proposed consistent estimators here, but with suboptimal rates of convergence. We propose an alternative method that takes into account the additive structure and is based on time averaging the data. This approach has similarities to the averaging method proposed by Pesaran (2006) except that our averaging is over time rather than cross sectional. In particular, let $\{w_{Tt}\}$ be some deterministic triangular array with $\sum_{t=1}^{T} w_{Tt} \leq \overline{w} < \infty$ and let $\overline{f}_{Tu} = \sum_{t=1}^{T} w_{Tt} f_{ut}$, $\overline{f}_{Tj} = \sum_{t=1}^{T} w_{Tt} f_{jt}$. We require that $\overline{f}_{Tu}$ and $\overline{f}_{Tj}$, $j = 1, \ldots, J$, are nonzero for all T larger than some fixed value (which is consistent with the factors being random with nonzero population mean). For example, $w_{Tt} = 1/T$ works in the case that the factors do not have mean zero. In other cases, a different weighting sequence is needed, for example, $w_{Tt} = 1(t = s)$ works in the case that $f_{js} \neq 0$. What happens when either $f_{jt} = 0$ for all t for some j or $\overline{f}_{Tj} \simeq 0$ is an important issue and we discuss this in footnote 5 below.

Letting $\overline{y}_i = \sum_{t=1}^{T} w_{Tt} y_{it}$ and $\overline{\varepsilon}_i = \sum_{t=1}^{T} w_{Tt} \varepsilon_{it}$, where we drop the T subscript for convenience, we have

$$\overline{y}_i = \overline{f}_u + \sum_{j=1}^{J} \overline{g}_j(X_{ji}) + \overline{\varepsilon}_i, \tag{20}$$

where $\overline{g}_j(\cdot) = g_j(\cdot)\overline{f}_j$. This constitutes an additive nonparametric regression with components $\overline{g}_j$ that are mean zero, that is, $E[\overline{g}_j(X_{ji})] = 0$, $j = 1, \ldots, J$, and so fit into the framework of Mammen, Linton, and Nielsen (MLN) (1999). Therefore, we estimate the functions $\overline{g}_j(\cdot)$ by their smooth backfitting method using the cross-sectional data set $\{\overline{y}_i, X_i, i = 1, \ldots, n\}$, or a subset thereof, and denote these estimates by $\widetilde{g}_j(\cdot)$. Then note that $g_j(\cdot) = \overline{g}_j(\cdot)/\sqrt{\int \overline{g}_j(x_j)^2 \, dP_j^*(x_j)}$, so a corresponding renormalization of the estimated $\overline{g}_j$ yields an estimate of g_j. The theory of (MLN) (1999) can be directly applied to $\widetilde{g}_j(\cdot)$, except that the error term $\overline{\varepsilon}_i$ is $O_p(\delta_T)$, where $\delta_T^2 = \sum_{t=1}^{T} w_{Tt}^2$ under our conditions, which makes the convergence rate of the estimated functions faster by this magnitude. When $w_{Tt} = 1/T$, the error term $\overline{\varepsilon}_i = O_p(T^{-1/2})$. Denote these initial consistent estimators by $\widehat{g}_j^{[0]}(x_j) = \widetilde{g}_j(x_j)/\sqrt{\int \widetilde{g}_j(x_j)^2 \, dP_j^*(x_j)}$, $j = 1, \ldots, J$. To estimate the time-series factors f_t themselves, we cross sectionally regress y_{it} on a constant and $\widehat{g}_1^{[0]}(X_{1i}), \ldots, \widehat{g}_J^{[0]}(X_{Ji})$ for each time period t; denote this estimator by $\widetilde{f}_t$. Let $\widehat{f}^{[0]} = \widetilde{f}$.[5]

[5] We discuss here what happens when the condition $\overline{f}_{Tj} \neq 0$ is violated. Consider the special case $y_{it} = g(X_i)f_t + \varepsilon_{it}$, where f_t is random and satisfies $Ef_t = 0$. Taking $w_{Tt} = 1/T$, we obtain $\overline{y}_i = g(X_i)\overline{f} + \overline{\varepsilon}_i$, where, under some further conditions, $\sqrt{T} \times \overline{f} \Rightarrow Z$ for some nor-

The backfitting algorithm of MLN itself requires starting values. We propose to use a variant of Rosenberg's (1974) linear model:

$$g_j(X_{ji}) = X_{ji}. \tag{21}$$

In this linear case, it is simple to rescale the characteristics so that the identification constraints hold using (21). We scale the mean and variance of the characteristics so that $E^*[X_{ji}] = 0$ and $\text{var}^*[X_{ji}] = 1$ for each j; for each characteristic, this just requires subtracting the weighted cross-sectional mean and dividing by the weighted cross-sectional standard deviation each time period. The simple linear model for $g(\cdot)$ gives rise to a linear cross-sectional regression model to estimate f_{jt}:

$$y_{it} = f_{ut} + \sum_{j=1}^{J} X_{ji} f_{jt} + \varepsilon_{it}. \tag{22}$$

We begin with ordinary least squares estimation of (22). These estimates of f_{ut} and f_{jt} serve merely as starting values and have no consistency properties. Connor and Linton (2007) found that this linear model provides quite a reasonable first approximation.

3.2.3. *Unbalanced, Time-Varying Panel Data*

The notation used so far assumes a fully balanced panel data set. The set of observed assets is assumed to be constant over time, with each asset having a fixed vector of characteristic betas. The only time variation in this fully balanced panel comes through the random factor realizations and random asset-specific returns. In applications, the set of assets must be allowed to vary over the time sample, since the set of equities with full records over a reasonably long sample period is a small subset of the full data set. Also, the characteristics of the assets in some cases should be allowed to vary through time. We may assume that the observations are unbalanced in the sense that in time period t, we only observe n_t firms (for simplicity labelled $i = 1, \dots, n_t$). Also, we assume that the characteristics are time varying but stationary over time for each i and are i.i.d. over i. This yields first-order conditions for f and g that are similar to the balanced case. We give some explicit details.

Regarding starting values, proceed as follows: Perform cross-sectional smooth backfitting for each time period, renormalize, and then average the

mal random variable Z. Consider the kernel estimator $\widetilde{g}(x)$ of $g(x)$ based on smoothing $\overline{y}_i$ against X_i. Then $\widetilde{g}(x) = Zg(x)/\sqrt{T} +$ smaller terms, and so the renormalized estimator $\widehat{g}(x) = \widetilde{g}(x)/\sqrt{\int \widetilde{g}(x_j)^2\, dP_j^*(x_j)}$ is consistent (up to a random sign). It follows that the estimated factors from the least squares regression of y_{it} on $\widehat{g}(X_i)$ are also consistent for each t.

estimates over time. We discuss implementation in more detail in Sections 4 and 5.

Note that instead of averaging the data, one could do period-by-period estimation and then average the estimates obtained for each period. This method would work in the case where the covariates are time varying or the data are unbalanced.

4. DISTRIBUTION THEORY

In this section, we provide the distribution theory for our estimates of the factors and of the characteristic functions based on a finite number of iterations from the consistent initial values we proposed above, namely $\widehat{g}_j(x_j) = \widehat{g}_j^{[k]}(x_j)$ and $\widehat{f}_t = \widehat{f}_t^{[k]}$ defined in (18) and (19). Following earlier work, we expect this limit theory to also well represent what happens for the "iterate to convergence from an arbitrary starting value" method. We also work in the balanced case. The general approach uses the methods developed in Mammen, Linton, and Nielsen (1999) and Linton and Mammen (2005) to treat estimators defined as the solutions of type 2 linear integral equations. The novelty here is due to the weighting by the factors and the fact that we wish to allow both the cross section and the time dimension to grow. We allow the error terms to be both temporally and cross-sectional weakly dependent; with regard to the cross-sectional dependence, we assume that there is a known clustering structure (Wooldridge (2006)). Specifically, we write $\varepsilon_{it} = \varepsilon_{\ell m t}$, with $\ell = 1, \dots, L$ and $m = 1, \dots, M$ such that $n = L \cdot M$; independence prevails, across clusters, while within clusters, an arbitrary amount of dependence is permitted. Regarding the asymptotics, we take pathwise limits as $L, T \to \infty$ (and possibly $M \to \infty$) as described in Phillips and Moon (1999, Definition 2(b)).

Let $\mathcal{F}_a^b$ be the σ-algebra of events generated by the vector random variable $\{U_t; a \leq t \leq b\}$. The processes $\{U_t\}$ is called strongly mixing (Rosenblatt (1956)) if

$$\sup_{1 \leq t} \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+k}^{\infty}} |\Pr(A \cap B) - \Pr(A)\Pr(B)| \equiv \alpha(k) \to 0 \quad \text{as} \quad k \to \infty.$$

We make the following assumptions.

ASSUMPTIONS A:

A1. *We suppose that* $\varepsilon_{it} = \sigma_t(X_i)\eta_{it}$ *with* $\{\eta_{it}\}$ *independent of* $\{X_i\}$. *Furthermore, for each* i, η_{it} *are stationary martingale difference sequences and are geometrically strongly mixing across* t (*i.e.*, $\alpha_i(k) \leq c\rho^k$ *for some* $\rho \in (0, 1)$) *with* $E\eta_{it}^2 = 1$. *The random variables* $(X_{\ell 1}, \dots, X_{\ell M})$ *and* $(\eta_{\ell 1 t}, \dots, \eta_{\ell M t})$ *are independent and identically distributed across* $\ell = 1, \dots, L$. *Furthermore, for some* $\kappa > 4$, $E[|\eta_{it}|^{\kappa}] < \infty$.

A2. *The covariate* $X_i = (X_{1i}, \dots, X_{Ji})^\top$ *has absolutely continuous density* p *supported on* $\mathcal{X} = [\underline{x}, \overline{x}]^J$ *for some* $-\infty < \underline{x} < \overline{x} < \infty$. *The functions* $g_j(\cdot)$ *together with the density* $p(\cdot)$ *are twice continuously differentiable over the interior of* $\mathcal{X}$ *and are bounded on* $\mathcal{X}$. *The density function* $p(x)$ *is strictly positive at each* $x \in \mathcal{X}$. *Denote by* $p_j(x)$ *the marginal probability density for characteristic* j *with support* $\mathcal{X}_j = [\underline{x}, \overline{x}]$. *The matrix* $E[G(X_i)G(X_i)^\top]$ *is strictly positive definite. The function* $\sigma_t^2(x) = E[\varepsilon_{it}^2 | X_i = x]$ *is continuous on* $\mathcal{X}$.

A3. *For each* $x \in [\underline{x}, \overline{x}]$, *the kernel function* K^x *has support* $[-1, 1]$, *and satisfies* $\int K^x(u)\,du = 1$ *and* $\int K^x(u)u\,du = 0$ *such that for some constant* C, $\sup_{x \in [\underline{x}, \overline{x}]} |K^x(u) - K^x(v)| \le C|u - v|$ *for all* $u, v \in [-1, 1]$. *Define* $\mu_j(K) = \int u^j K(u)\,du$ *and* $\|K\|_2^2 = \int K^2(u)\,du$. *The kernel* K *is bounded, has compact support* ($[-c_1, c_1]$, *say*), *is symmetric about zero, and is Lipschitz continuous, that is, a positive finite constant* C_2 *exists such that* $|K(u) - K(v)| \le C_2|u - v|$.

A4. *In some pathwise fashion,* $L, T \to \infty$.

A5. *We assume that* $\sum_{t=1}^T w_{Tt}^2 = \delta_T^2 \to 0$ *as* $T \to \infty$. *For* $j = u, 1, \dots, J$, $\sum_{t=1}^T w_{Tt} f_{jt} \to \overline{f}_j > 0$; *for some* $a > 2$, *the quantities* $\sup_{T \ge 1} \sum_{t=1}^T |f_{jt}|^a / T < \infty$. *The quantities* $\Phi_j = \lim_{T\to\infty} T^{-1} \sum_{t=1}^T f_{jt}^2$ *and* $\Psi_j(x_j) = \lim_{T\to\infty} T^{-1} \sum_{t=1}^T f_{jt}^2 \times \sigma_{jt}^2(x_j)$ *exist and* $\Phi_j > 0$.

A6. *The bandwidth sequence* $h(L, T)$ *satisfies* $h \to 0$, $MTh\delta_T^2 \to 0$, *and* $nTh^2 \to \infty$ *as* $L, T \to \infty$.

A7. *For* $j = u, 1, \dots, J$, *there exists* $\rho' > 0$ *such that* $\max_{1 \le t \le T} |f_{jt}| = O((\log T)^{\rho'})$.

In A1, we allow a general form of time-series and cross-sectional conditional heteroskedasticity in the errors ε_{it}, and time-series and cross-sectional dependence, although we assume that η_{it} is a martingale difference sequence, which seems like a natural assumption to make in this context. Assumption A5 embodies an assumption about the magnitudes of the factors; here we assume that they behave like the outcome of a stationary process with finite moments of order a. This could be relaxed to either faster growth in $\sum_{t=1}^T f_{jt}^2$, reflecting nonstationary factors, or slower growth, reflecting perhaps many zero values in the factors, but we do not do this here as the data seem to support this assumption. Assumption A7 is needed for the uniform convergence rates below. This assumption is consistent with the factors being realizations from a stationary Gaussian process. Again, this condition could be weakened to allow faster growth in $\max_{1 \le t \le T} |f_{jt}|$ at the expense of further restrictions elsewhere.

Define for each $j = 1, \dots, J$ and $t = 1, \dots, T$,

$$\Omega_j(x_j) = \frac{\Psi_j(x_j)}{p_j(x_j)\Phi_j^2} \|K\|_2^2, \tag{23}$$

$$\Gamma_t = E\left[\left(\frac{1}{M}\sum_{m=1}^M \sigma_t(X_{\ell m})G(X_{\ell m})\right)\left(\frac{1}{M}\sum_{m=1}^M \sigma_t(X_{\ell m})G(X_{\ell m})^\top\right)\right],$$

$$V_{t,t} = E[G(X_i)G(X_i)^\top]^{-1}\Gamma_t E[G(X_i)G(X_i)^\top]^{-1}. \tag{24}$$

THEOREM 1: *Suppose that Assumptions* A1–A6 *hold and that* $Lh^4 \to 0$. *Then for each* t,

$$\sqrt{L}(\widehat{f}_t - f_t) \Longrightarrow N(0, V_{t,t}). \tag{25}$$

Suppose that A1–A7 *hold. Then for some* $\rho > 0$,

$$\max_{1\leq t\leq T} |\widehat{f}_t - f_t| = O_p\big((L^{-1/2} + h^2)(\log T)^\rho\big). \tag{26}$$

Furthermore, a bounded continuous function $\beta_j(\cdot)$ *exists such that for each* $x_j \in (\underline{x}, \overline{x})$,

$$\sqrt{nTh}(\widehat{g}_j(x_j) - g_j(x_j) - h^2\beta_j(x_j)) \Longrightarrow N(0, \Omega_j(x_j)). \tag{27}$$

Furthermore, $\widehat{g}_j(x_j)$ *and* $\widehat{g}_k(x_k)$ *are asymptotically independent for any* x_j *and* x_k.

REMARK 1: The clustering dependence does not affect the rate of convergence of the nonparametric estimates, provided $MTh\delta_T^2 \to 0$ (which amounts to $Mh \to 0$ when the full sample is used for averaging), and also does not affect the asymptotic variance. However, it does effect the rate of convergence of the factor estimates and their asymptotic variance matrix. It may be that $\widehat{g}_j(x_j)$ converges to $g_j(x_j)$ faster than $\widehat{f}_t$ converges to f_t; this happens when $MTh \to \infty$. This is because of the extra pooling over time in the specification of g_j. Note that the asymptotic variance of the characteristic function estimates is as if the factors were known. The estimators $\widehat{g}_j(x_j)$ are consistent at rate $(nT)^{-2/5}$ provided a bandwidth of order $(nT)^{-1/5}$ is chosen and under some restrictions on the rates at which T, L, and M increase. This should be the optimal rate (Stone (1980)). Note that the asymptotic variance of the factor estimates is as if the characteristic functions were known and least squares were applied.

REMARK 2: When ε_{it} is i.i.d., the asymptotic variance of $\widehat{g}_j(x_j)$ simplifies to $\Omega_j(x_j) = (\sigma_\varepsilon^2/p_j(x_j)\Phi_j)\|K\|_2^2$, where $\sigma_\varepsilon^2 = \sigma_t^2(x)$. We argue that this is a natural "oracle" bound along the lines of Linton (1997). Suppose that we could observe the partial residuals $U_{jit} = y_{it} - f_{ut} - \sum_{k\neq j} f_{kt}g_k(X_{ki})$. Then we can compute the pooled regression smoother

$$\widehat{g}_j^{\text{oracle}}(x) = \frac{\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{n} K_h(X_{ji}, x_j) f_{jt} U_{jit}}{\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{n} K_h(X_{ji}, x_j) f_{jt}^2}, \tag{28}$$

which can be interpreted as a local likelihood estimator (Tibshirani (1984)), for the model $y_{it} = f_{jt}g_j(X_{ji}) + \varepsilon_{it}$ with i.i.d. normal errors. This shares the asymptotic variance of our estimator and since it uses more information than we have available, it is comforting that our estimator performs as well. In fact, one can argue further that the asymptotic variance of the oracle (local likelihood) estimator is the same as the asymptotic variance of the maximum likelihood estimator (MLE) of a correctly specified parametric model at the point of interest, $y_{it} = f_{jt}\theta + \varepsilon_{it}$, where ε_{it} is i.i.d. normally distributed, that uses an equivalent number of observations (Tibshirani (1984, Chap. 5)).

REMARK 3: Standard errors can be obtained in an obvious way by plugging in estimated quantities. In particular, valid "clustered" standard errors for the factors can be obtained from the final stage least squares regression of returns on the characteristic functions (Wooldridge (2006)). We recommend computing standard errors for $\widehat{g}_j(x_j)$ from

$$\frac{\widehat{\Omega_j(x_j)}}{nTh} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n} K_h^2(X_{ji}, x_j)\widehat{f}_{jt}^2\widehat{\varepsilon}_{it}^2}{\left(\sum_{t=1}^{T}\sum_{i=1}^{n} K_h(X_{ji}, x_j)\widehat{f}_{jt}^2\right)^2}, \tag{29}$$

where $\widehat{\varepsilon}_{it} = y_{it} - \widehat{f}_{ut} - \sum_{j=1}^{J} \widehat{f}_{jt}\widehat{g}_j(X_{ji})$ are residuals computed from the estimated factors and characteristic functions (see Fan and Yao (1998) for a discussion of nonparametric standard errors).

REMARK 4: Bandwidth and factor selection can both be handled in the framework of penalized least squares. We recommend the bandwidth selection method developed by Mammen and Park (2005); see Connor, Hagmann, and Linton (2012) for more details.

REMARK 5: The results (25)–(27) follow also for the unbalanced case with a time-varying covariate with suitable generalizations. The case where the covariate process is stationary is particularly simple because then one only needs to replace n by n_t in (25), n by $\min_{1\le t\le T} n_t$ in (26), and nT by $\sum_{t=1}^{T} n_t$ in (27).

REMARK 6: An alternative approach to cross-sectional dependence is to follow Connor and Koraczyk (1993) and assume that some ordering of the observations exists such that there is strong mixing of η_{it} across i with some fairly rapid rate of decay. Robinson (2007) described a more explicit framework through moving average processes. In both cases, the cross-sectional dependence washes out of the distribution for the nonparametric estimates, but does affect the distribution of the factor estimates. Although it does not affect the rate of convergence of the factor estimates (like our clustering assumption may), it does affects the limiting variance in a complicated way and in a way that precludes consistent inference without further assumptions.

REMARK 7—Specification Testing: One can test the underlying specification in a number of ways. Given our sampling scheme, there are two general restrictions on the conditional expectation $E[y_{it}|X_i] = m_t(X_i)$: poolability, and additivity. Baltagi, Hidalgo, and Li (1996) proposed a test of poolability that can be adapted to our framework. Gozalo and Linton (2001) proposed tests of additivity in a cross-sectional setting based on comparing restricted with unrestricted estimators that work with marginal integration estimators (Linton and Nielsen (1995)). The working paper version gives more details.

5. EMPIRICAL ANALYSIS

5.1. *Data*

We follow Fama and French (1993) in the construction of the size and value characteristics. For each separate 12-month period July–June from 1970 to 2007, we find all securities which have complete Center for Research in Security Prices (CRSP) return records over this 12-month period and the previous 12-month period, and have two-digit Standard Industrial Classification code (from CRSP), market capitalization (from Compustat) and book value (from Compustat) records for the previous June. Throughout the empirical analysis, we use returns in excess of the risk-free return, treating the monthly Treasury bill return from CRSP as the risk-free return. The raw size characteristic each month equals the logarithm of the previous June's market value of equity. The raw value characteristic equals the ratio of the market value of equity to the book value of equity in the previous June. In addition to the Fama–French size and value characteristics, we derive from the same return data set a momentum characteristic as in Carhart (1997). This variable is measured as the cumulative 12-month return up to and including the previous month. Finally we add an own-volatility characteristic, a choice inspired by the recent work of Goyal and Santa Clara (2003) and Ang et al. (2006, 2009). We define raw volatility as the standard deviation of the individual security return over 12 months up to and including the previous month. The characteristics equal the raw characteristics except they are standardized each month to have zero mean and unit variance. The size and value characteristics are held constant from July to June, whereas the momentum and own-volatility characteristics change each month. Table I reports some descriptive statistics for the data: the number of securities in the annual cross section and the first four cross-sectional moments of the four characteristics. To save space, the table just shows nine representative dates (July at 5-year intervals), as well as time-series medians over the full 37 year period, using July data.

TABLE I

SAMPLE STATISTICS[a] OF RAW SECURITY CHARACTERISTICS

		Mean				Standard Deviation				Skewness				Excess Kurtosis			
Year	Firms	Size	Value	Mom	Vol	Size	Value	Mom	Vol	Size	Value	Mom	Vol	Size	Value	Mom	Vol
1970	1554	3.71	0.81	−0.44	0.11	1.58	0.51	0.36	0.04	0.34	0.89	−0.12	0.53	−0.39	0.19	−0.26	−0.26
1975	2475	4.01	1.33	0.40	0.17	1.55	0.80	0.36	0.07	0.28	0.89	0.46	0.78	−0.40	0.31	0.24	0.09
1980	2349	4.50	0.93	0.38	0.11	1.74	0.54	0.36	0.05	0.10	0.67	0.49	0.76	−0.35	−0.11	0.23	0.01
1985	4249	3.58	0.73	0.19	0.12	2.06	0.46	0.42	0.06	0.24	0.89	−0.07	0.86	−0.38	0.40	0.31	0.21
1990	4543	3.70	0.80	0.04	0.12	2.21	0.61	0.45	0.07	0.14	1.18	0.14	1.07	−0.32	0.90	0.55	0.60
1995	6037	4.09	0.61	0.18	0.11	1.93	0.40	0.41	0.06	0.29	0.87	0.26	0.96	−0.13	0.23	0.75	0.22
2000	5635	4.68	0.69	0.22	0.18	1.92	0.56	0.74	0.11	0.29	1.03	0.89	1.16	−0.19	0.45	0.66	0.62
2005	4808	5.33	0.50	0.12	0.10	1.99	0.30	0.36	0.06	0.21	0.80	0.02	0.93	−0.19	0.17	0.60	0.20
Med	4518	4.08	0.73	0.17	0.12	1.93	0.48	0.41	0.06	0.21	0.89	0.12	0.92	−0.33	0.33	0.50	0.23

[a] Some descriptive statistics of the cross-sectional data for July at 5-year intervals: the number of securities the annual cross section and the first four cross-sectional moments of the four raw characteristics. Separately provided are the time-series medians over the full 42-year period, also using July data.

Four notes on the interpretation of these characteristics in terms of our econometric theory are in order.

NOTE 1: We treat all four characteristics as observed without error. Informally, we think of momentum and own volatility as behaviorally generated sources of return comovement. Investors observe momentum and own volatility over the previous 12 months (along with the most recent observations of size and value), and adjust their portfolio and pricing behavior to account for the observed values; this in turn accounts (for some unspecified reasons) for the subsequent return comovements associated with these characteristics. Understanding more fundamentally the sources of the characteristic-related comovements is an important topic which we do not address here.

NOTE 2: The cited references Ang et al. (2006, 2009) and Goyal and Santa Clara (2003) used idiosyncratic volatility rather than total volatility as a characteristic. From our perspective, total volatility is preferable, since it does not require a previous estimation step to remove market-related return from each asset's total return.

NOTE 3: In our econometric theory, we allow all the characteristics to vary freely over time. Since size and value change annually, whereas momentum and own volatility change monthly, another approach would be to modify the econometric theory to allow some characteristics to change only at a lower frequency. We do not pursue this alternative approach here.

NOTE 4: In the theoretical development of our estimators, we describe the cluster-based standard errors for the factor estimates for generality, but in our application, we impose unit-asset clusters. This is standard practice in the empirical finance literature and is unlikely to have much impact on the results (particularly since with 1.8 million observations, the standard errors are extremely small).

A useful descriptive statistic is the correlation matrix of the explanatory variables. This is complicated in our model by the time-varying nature of the characteristics which serve as our explanatory variables. Figure 1 shows, for each pair of characteristics, the time-series evolution of the cross-sectional correlation between them, using the cross section each July. It is clear that these correlations are not constant over time. The correlation between size and value exhibits slow and persistent swings, with a negative average. Size and momentum, on the other hand, are, on average, uncorrelated. Most interesting is the relationship between own volatility and momentum, taking large swings from high positive correlation of 0.6 to negative correlation of -0.31. None of the correlations is large enough in magnitude to be worrisome in terms of accurate identification of the model.

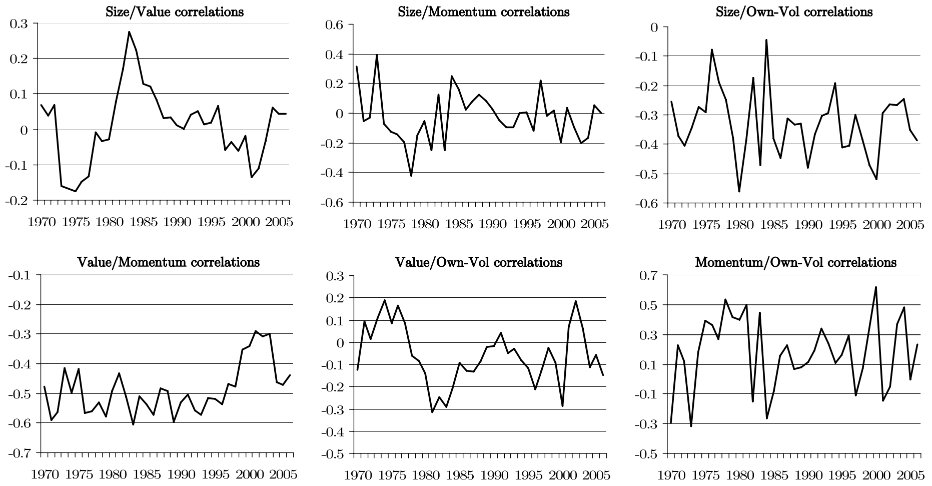

FIGURE 1.—Time series plots of cross-sectional correlations between the characteristics. The time-varying nature of the correlations between security characteristics is depicted by showing the cross-sectional correlation for each pair of characteristics each July.

5.2. *Implementation*

In the case of a fully balanced panel, it would be straightforward to estimate the characteristic-beta function at each data point in the sample. However, in the presence of time-varying characteristics, this is not feasible since the number of asset returns (each with a unique vector of characteristics) equals 1,793,844 in our sample. To make the algorithm described in Section 3 computationally feasible, we concentrate estimation of the characteristic functions on 61 equally spaced grid points between −3 and 3, which corresponds to a distance of 0.1 between contiguous grid points. We use linear interpolation between the values at these grid points to compute the characteristic-beta function at all 1,793,844 sample points. Then we use the full sample of 1,793,844 asset returns and associated factor betas to estimate the factor returns. This procedure greatly improves the speed of our algorithm while sacrificing little accuracy, since the characteristic-beta functions are reasonably linear between these closely spaced grid points.

We chose a Gaussian kernel throughout to nonparametrically estimate the conditional expectations. The advantage of this kernel is that it is very smooth and produces nice regular estimates, whereas, say, the Epanechnikov kernel produces estimates with discontinuities in the second derivatives. We use a bandwidth of 0.10 throughout, based on a simple Silverman rule.

5.3. *The Characteristic-Beta Functions*

Table II shows the estimates of the characteristic-beta functions at a small selected set of characteristic values and the heteroskedasticity-consistent standard errors from (29) for each of these estimates. To avoid any spurious nonlinearity results due to smoothing in regions where there are no data, we report results for each characteristic only over a support ranging from the empirical 2.5% to the 97.5% quantile. The standard errors tend to be somewhat larger in the tails, where the data are sparser. Given that our procedure is able to use all 1.8 million return observations to estimate the characteristic-beta functions, the standard errors are small.

The characteristic-beta functions over all grid points are displayed in Figure 2. Note that these characteristic-beta functions satisfy the equally weighted zero-mean/unit variance identification conditions described in Section 3. We estimate the model with and without industry factors, but find little difference in the estimated characteristic-beta functions. Except where stated otherwise, we refer to the model without industry factors in our empirical analysis. The characteristic-beta functions are monotonically increasing for all four characteristics. Size and value show strongly nonlinear characteristic-beta functions, both with concave shapes. The observed shapes for momentum and own volatility are closer to linear. Not surprisingly, given 1.8 million observations, we can reject linearity in all four cases. The economic (as opposed to statistical) significance of the finding seems strongest for size and value, as illustrated in Figure 2.

TABLE II

SECURITY CHARACTERISTIC-BETA FUNCTIONS AND STANDARD ERRORS

	Size		Value		Momentum		Volatility	
Grid	Value	SE	Value	SE	Value	SE	Value	SE
−2.00	n.c.	n.c.	n.c.	n.c.	−2.51	0.03	n.c.	n.c.
−1.50	−1.93	0.03	n.c.	n.c.	−1.44	0.03	n.c.	n.c.
−1.00	−0.99	0.03	−1.13	0.05	−0.86	0.03	−1.07	0.03
−0.50	−0.24	0.02	−0.14	0.05	−0.36	0.03	−0.54	0.02
0.00	0.32	0.02	0.39	0.04	0.10	0.03	0.13	0.02
0.50	0.74	0.02	0.66	0.04	0.54	0.03	0.64	0.02
1.00	0.94	0.02	0.91	0.03	0.90	0.02	1.06	0.02
1.50	1.08	0.02	1.15	0.04	1.25	0.02	1.51	0.02
2.00	1.10	0.03	1.44	0.04	1.60	0.03	1.86	0.02
2.50	n.c.	n.c.	1.65	0.09	1.96	0.04	2.25	0.04

[a] The time-series medians are over the full 44-year period, also using July data, and the heteroskedasticity-consistent standard errors are given for each of these estimates. Results are reported for each characteristic over a support ranging from the empirical 2.5% to the 97.5% quantile. n.c. denotes not computed.

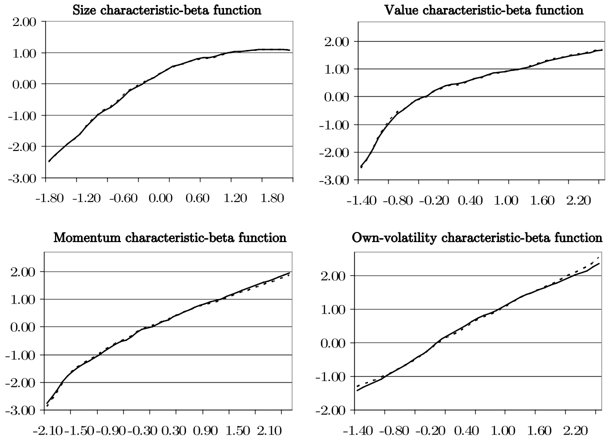

FIGURE 2.—The characteristic-beta functions. The dashed (solid) lines display the estimated characteristic-beta functions when industry effects are (are not) included in the model estimation algorithm. The functions are estimated over a grid ranging from the 2.5% to the 97.5% empirical quantile of the respective security characteristic.

5.4. *Explanatory Power of Each Factor*

Note that at each step of the iterative estimation, the factor returns are the coefficients from period-by-period unconstrained cross-sectional regression of returns on the previous iteration's factor betas. To measure the explanatory power of the factors, we take the final-step estimates of factor betas and perform the set of cross-sectional regressions with all the factors, each factor singly, and all the factors except each one. Table III shows the time-series averages of uncentered R^2 (UR2) statistic in all these cases: all five factors, each

TABLE III

UNCENTERED R^2 STATISTICS (UR2)[a]

	Marginal UR2 Statistics When Adding Individual Factors to the Model				
	Market	Size	Value	Momentum	Volatility
Adding first	10.63%	1.20%	3.41%	2.30%	4.17%
Adding last	n.m.	0.16%	2.88%	1.29%	2.51%
	UR2 With Industry Factors			UR2 Without Industry Factors	
UR2	21.27%			19.81%	

[a]The time-series averages of uncentered cross-sectional R^2 (UR2) statistics are given as a measure of the explanatory power of the factor model. The upper part of the table shows average UR2 statistics from cross-sectional regressions of excess returns on each characteristic-beta function singly as well as their marginal contribution given the other four. The lower part of the table shows the UR2 statistic with all five characteristic beta functions, with and without industry factors.

single factor, and each subset of four factors. The market factor is dominant in terms of explanatory power—a well known result. The own-volatility factor is the strongest of the characteristic-based factors, followed by value, momentum, and size. The ordering of importance is mostly the same (own volatility and value switch places) if we consider the marginal contribution of each given the other four. Table III also shows UR2 when the model is supplemented with the industry factors; they increase explanatory power by 1.45%.

We test for the statistical significance of each factor by calculating, for each cross-sectional regression, the t-statistic for each estimated coefficient, based on Hansen–White heteroskedasticity-consistent standard errors. Then for each factor, we find the average number of cross-sectional regression t-statistics that are significant at a 95% confidence level across the 444 time periods. The resulting count statistic has an exact binomial distribution under the null hypothesis that the factor return is zero each period. Table IVa shows the annualized means and standard deviations of the factor returns, the percentage of significant t-statistics for each factor, and the aggregate p-value. All five factors are highly significant.

Table IVb displays the correlations of the estimated factors, along with the three Fama–French factors, RMRF, SMB, and HML. RMRF is the Fama–French market factor that denotes the return to the value-weighted market index; SMB is the return to a small-capitalization portfolio minus the return to a large-capitalization portfolio; HML is the return to a high book-to-price portfolio minus the return to a low book-to-price portfolio, where all returns are measured as excess to the Treasury bill return. See Fama and French (1993) for a detailed discussion of their portfolio formation rules. We also include a momentum factor created by French—the return to a portfolio with high cumulative returns over the past 12 months minus the return to a portfolio with low cumulative returns over the past 12 months, adjusted to have roughly equal average capitalization; see French's website[6] for details, where all the Fama–French data are freely available. Except for size, our factors and the analogous Fama–French factors are highly correlated. Note that the size characteristic is defined inversely in the two models, hence the negative correlation between them. The Fama–French factors are based on capitalization-weighted portfolios, whereas our factors are statistically generated, treating all assets equally. Since the cross section of securities is dominated, in terms of the number of securities, by low-capitalization firms, this induces a strong positive correlation between our market factor and the Fama–French size (SMB) factor. Our volatility factor has strong positive correlation with the market factor (either version). This corroborates the finding in Ang et al. (2009, Table 10), which shows high covariance between their idiosyncratic-volatility factor returns and the Fama–French market factor returns. It also seems theoretically consistent

[6] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_mom_factor.html.

TABLE IV

FACTOR RETURN STATISTICS AND COMPARISON TO FAMA–FRENCH FACTOR-MIMICKING PORTFOLIOS[a]

a. Factor Return Statistics

	Market	Size	Value	Momentum	Volatility
Annualized mean	11.06%	−4.47%	4.22%	−1.18%	−1.01%
Annualized volatility	19.73%	5.65%	4.29%	6.61%	7.85%
% Periods significant[b]	93.47%	68.02%	66.22%	59.46%	68.92%
Overall p-value	0.00%	0.00%	0.00%	0.00%	0.00%

b. Empirical Factor Return and Fama–French Return Correlations

	Market	Size	Value	Momentum	Volatility	RMRF	SMB	HML	FF_MOM
Market	1.00	−0.01	−0.16	−0.47	0.78	0.85	0.64	−0.32	−0.21
Size		1.00	−0.59	0.13	−0.04	0.35	−0.32	−0.15	−0.11
Value			1.00	−0.09	−0.25	−0.46	0.06	0.71	−0.07
Momentum				1.00	−0.52	−0.24	−0.23	0.04	0.76
Volatility					1.00	0.62	0.61	−0.49	−0.19
RMRF						1.00	0.27	−0.44	−0.09
SMB							1.00	−0.29	−0.01
HML								1.00	−0.10
FF_MOM									1.00

[a] Table IVa shows the mean, volatility, and statistical significance of each factor. The statistical significance is calculated as the percentage of significant t-values (95% confidence level) for each factor by computing for each cross-sectional regression, the t-statistic for each estimated coefficient based on Hansen–White heteroskedasticity-consistent standard errors. The aggregate p-value is also provided. Table IVb displays the correlations between the estimated factors, along with the three Fama–French factors (RMRF, SMB, and HML) and a momentum factor FF-MOM created by French. RMRF is the Fama–French market factor that denotes the return to the value-weighted market index minus the risk-free return; SMB is the return to a small-capitalization portfolio minus the return to a large-capitalization portfolio; HML is the return to a high book-to-price portfolio minus the return to a low book-to-price portfolio. FF-MOM is the return to a portfolio with high cumulative returns over the past 12 months minus the returns to a portfolio with low cumulative returns over the past 12 months, adjusted to have roughly equal average capitalization.

[b] Defined as abs(t-value) $>$ 1.96.

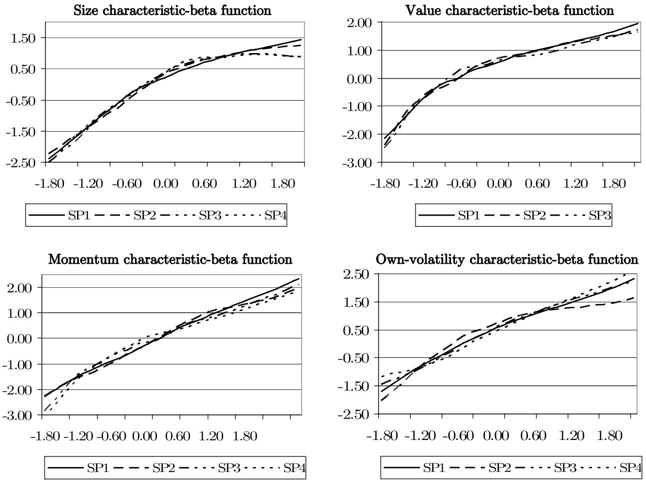

FIGURE 3.—The characteristic-beta functions estimated on four subperiods. The SP1 line is a function for the 1970 July–1980 June period and SP2 is for the 1980 July–1990 June period, while SP3 and SP4 display functions for the 1990 July–2000 June and 2000 July–June 2007 periods, respectively. The functions are estimated over a grid ranging from the 2.5% to the 97.5% empirical quantile of the respective security characteristic.

with the finding in Ang et al. (2006) that the market factor return is negatively correlated with changes in VIX, a forward-looking index of market volatility. Essentially, the positive correlation between the own-volatility factor and market factor means that high own-volatility stocks outperform when the overall market rises and underperform when the overall market falls. There is also a strong negative correlation between the own-volatility and momentum factor returns.

Figure 3 shows the characteristic-beta functions reestimated on the four nonoverlapping 111-month subintervals in the data set. The functions seem stable over time although we do not attempt a formal test.

5.5. *Including a Mispricing Function in Model Estimation*

Next we add an additive nonparametric mispricing function (α function) to the model and reestimate. We use the same set of four characteristics for the α function as in the factor model and an additive form, so that the α function is the sum of four component nonparametric functions. The estimated component α functions can be interpreted as a nonparametric version of the collection of estimated α coefficients for characteristic-sorted portfolios shown in Fama and French (1993, e.g., Table 6). Figure 4 shows the results. There is little evidence against the five-factor asset pricing model: all four mispricing functions differ only negligibly from zero. As a caveat, one should not casually examine these graphs looking for an upward or downward linear trend,

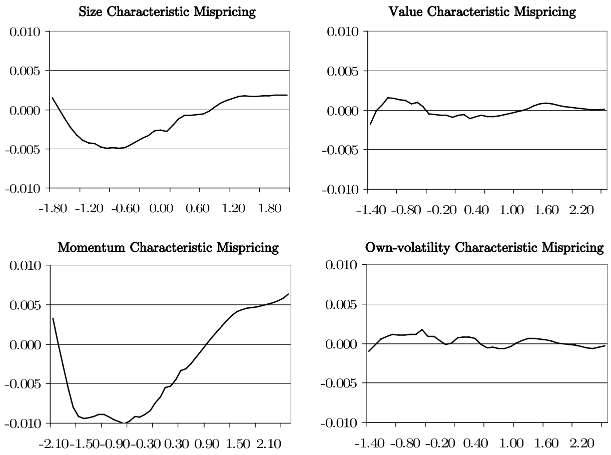

FIGURE 4.—Characteristic-based mispricing functions. The four additive nonparametric characteristic-based mispricing functions are estimated over a grid ranging from the 2.5% to the 97.5% empirical quantile of the respective security characteristic.

and not take the absence of such a trend as empirical confirmation of no mispricing. Since the functions are required to be orthogonal to the estimated characteristic-exposure functions, all of which have near linear shapes, they cannot have such a shape. The same caveat applies when examining the panels of estimated αs in Fama and French (1993, Table 6) since essentially the same identification condition is imposed implicitly there.

6. SUMMARY AND CONCLUSION

Following the pioneering work of Rosenberg (1974), Fama and French (1993) and others, characteristic-based factor models have played a leading role in explaining the comovements of individual equity returns. This paper applies a new weighted additive nonparametric estimation procedure to estimate characteristic-based factor models more data efficiently than existing nonparametric methods. The methodology we have developed extends existing results to the large cross section large time-series setting. We obtain a variety of statistical results that are useful for conducting inference. We think this methodology can be useful elsewhere.

We estimate a characteristic-based factor model with five factors: a market factor, size factor, value factor, momentum factor, and own-volatility factor. Although much of the existing literature has focused on the three-factor Fama–French model (market, size, and value), we find that the momentum and own-volatility factors are at least as important, if not more, than, size and value in explaining return comovements. The univariate functions that map characteristics to factor betas are monotonic but not linear; the deviation from linearity is particularly strong for size and value, and less so for momentum and own volatility.

Our methodology provides a new nonparametric test of multi-beta asset pricing theory. We estimate nonparametrically a set of additive mispricing functions based on the four security characteristics. We show that the estimation problem necessitates that an identification condition be imposed on the mispricing functions: for each characteristic, the mispricing function must be orthogonal to the factor exposure function. We show that the same identification condition is imposed implicitly in the traditional Fama–French portfolio-sort-based method. We find little evidence against the five-factor asset pricing model, but note that the imposition of the identification condition limits the power of this empirical finding. Since the factor exposure functions tend to be fairly smooth monotonic functions, a mispricing function is only identified to the extent that it is orthogonal to that shape.

APPENDIX: PROOF OF RESULTS

Let E_ℓ denote expectation conditional on all cluster ℓ information.

PROOF OF THEOREM 1: For simplicity and without loss of generality, we only take one observation per cluster; specifically, we use the sample $\{y_{\ell mt}, X_{\ell m}, \ell = 1, \dots, L, t = 1, \dots, T\}$ for some specific $m, = 1$, say. Under our conditions, we have L independent cross-sectional units and T weakly dependent time series units. Furthermore, $\overline{\varepsilon}_{\ell m} = \sum_{t=1}^{T} w_{Tt}\varepsilon_{\ell mt} = O_p(\delta_T)$ under the moment and mixing conditions we assume. We first establish the expansion for the initial estimator $\widehat{g}_j^{[0]}(x_j)$. We extend a little the result of Mammen, Linton, and Nielsen (1999). Specifically, their main result contains a stochastic expansion with an error of order $o_p(n^{-2/5})$, which is suitable for a single cross-section of size n. The main modification we make is with regard to the magnitude of the error Δ_n in their Assumptions A6 and A7, which they take to be $o_p(n^{-2/5})$. We note that

$$\begin{aligned}
&\sup_{x_k \in S_k} \left| \int \frac{\widehat{p}_{j,k}(x_j, x_k)}{\widehat{p}_k(x_k)} \widehat{m}_j^A(x_j)\, dx_j \right| \\
&\quad = \sup_{x_k \in S_k} \left| \int \frac{\widehat{p}_{j,k}(x_j, x_k)}{\widehat{p}_k(x_k)\widehat{p}_j(x_j)} \frac{1}{L} \sum_{\ell=1}^{L} K_h(X_{j\ell m}, x_j)\overline{\varepsilon}_{\ell m}\, dx_j \right| \\
&\quad = \sup_{x_k \in S_k} \left| \frac{1}{L} \sum_{\ell=1}^{L} \overline{\varepsilon}_{\ell m} \int \frac{\widehat{p}_{j,k}(X_{j\ell m} + uh, x_k)}{\widehat{p}_k(x_k)\widehat{p}_j(X_{j\ell m} + uh)}\, du \right| \\
&\quad = O_P\big(L^{-1/2}\delta_T\big),
\end{aligned}$$

because of the properties of the kernel density estimators $\widehat{p}_j$ and $\widehat{p}_{j,k}$, and independence across ℓ. A similar result holds for the L_2 error magnitude. As in their A7, the bias term error is $o_p(h^2)$.

Therefore, for any point x_j,

$$\begin{aligned}(30)\qquad \widetilde{g}_j(x_j)-\overline{g}_j(x_j) &= \frac{1}{Lp_j(x_j)}\sum_{\ell=1}^{L}K_h(X_{j\ell m},x_j)\overline{\varepsilon}_{\ell m}+h^2\beta_j(x_j)\\ &\quad + O_P\bigl(L^{-1/2}\delta_T\bigr)+o_p(h^2),\end{aligned}$$

where $\beta_j(x_j)$ is a deterministic bounded continuous function. The error in (30) is uniform over x_j. The first leading term in the expansion is $O_p(L^{-1/2}h^{-1/2}\delta_T)$. It follows that $\widetilde{g}_j(x_j)$ is $\sqrt{Lh}\delta_T$ consistent and asymptotically normal, and asymptotically independent of $\widetilde{g}_k(x_k)$. Then by standard arguments, $\int\widetilde{g}_j(x_j)^2\,dP^*(x_j)=\int\overline{g}_j(x_j)^2\,dP^*(x_j)+2\int\overline{g}_j(x_j)\beta_j(x_j)\,dP^*(x_j)+o_p(h^2)+O_p(L^{-1/2}\delta_T)$, so that

$$\begin{aligned}\widehat{g}_j^{[0]}(x_j)-g_j(x_j) &= \frac{\dfrac{1}{Lp_j(x_j)}\displaystyle\sum_{\ell=1}^{L}K_h(X_{j\ell m},x_j)\overline{\varepsilon}_{\ell m}+h^2\beta_j^{[0]}(x_j)}{\sqrt{\displaystyle\int\overline{g}_j(x_j)^2\,dP^*(x_j)}}\\ &\quad + O_P\bigl(L^{-1/2}\delta_T\bigr)+o_p(h^2),\end{aligned}$$

where $\beta_j^{[0]}(x_j)=\beta_j(x_j)-\overline{g}_j(x_j)\int\overline{g}_j(x_j)\beta_j(x_j)\,dP^*(x_j)$.

The proof of our main result is given in the following lemmas. Lemmas 1 and 2 give the pointwise performance of the initial factor estimator, denoted $\widetilde{f}_t$, while Lemmas 3 and 4 give the uniform (over t) performance of $\widetilde{f}_t$. Lemmas 5 and 6 give the pointwise expansion of the update $\widetilde{g}_j^{[1]}(x_j)$ of $\widetilde{g}_j(x_j)$.

Consider the infeasible estimator $f_t^\dagger$ that is the solution of the system of linear equations $A^\dagger f_t^\dagger=b_t^\dagger$, where

$$b_t^\dagger=\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}G(X_{\ell m})y_{\ell mt},$$

$$A^\dagger=\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}G(X_{\ell m})G(X_{\ell m})^\top.$$

This is unique with probability 1.

LEMMA 1: *Under our assumptions, for any t,*

$$\sqrt{L}(f_t^\dagger - f_t) \Longrightarrow N(0, V_{t,t}).$$

PROOF: We have

$$\sqrt{L}(f_t^\dagger - f_t) = (A^\dagger)^{-1}\frac{\sqrt{L}}{n}\sum_{i=1}^{n} G(X_i)\varepsilon_{it} = (A^\dagger)^{-1}\frac{1}{\sqrt{L}}\sum_{\ell=1}^{L}\zeta_{\ell t},$$

where $\zeta_{\ell t} = M^{-1}\sum_{m=1}^{M} G(X_{\ell m})\sigma_t(X_{\ell m})\eta_{\ell m t}$ satisfies $E\zeta_{\ell t} = 0$ and

$$\operatorname{var}(\zeta_{\ell t}) = E\left[\left(\frac{1}{M}\sum_{m=1}^{M} G(X_{\ell m})\sigma_t(X_{\ell m})\right)\right.$$
$$\left.\times\left(\frac{1}{M}\sum_{m=1}^{M} G(X_{\ell m})\sigma_t(X_{\ell m})\right)^{\top}\right].$$

Furthermore, by the independence over ℓ, the central limit theorem (CLT) applies and $\frac{1}{\sqrt{L}}\sum_{\ell=1}^{L}\zeta_{\ell t}$ is asymptotically normal. Furthermore, $A^\dagger = A + o_p(1)$ using a law of large numbers over the independent clusters. The result then follows by the Slutsky theorem. Note that $f_t^\dagger$ and $f_s^\dagger$ are correlated through the error term. *Q.E.D.*

Now consider the feasible factor estimator based on the initial estimator $\widetilde{f}_t = \widetilde{A}^{-1}\widetilde{b}_t$, where $\widetilde{b}_t = n^{-1}\sum_{i=1}^{n}\widetilde{G}(X_i)y_{it}$ and $\widetilde{A} = n^{-1}\sum_{i=1}^{n}\widetilde{G}(X_i)\widetilde{G}(X_i)^\top$, where $\widetilde{G}(X_i) = [1, \widetilde{g}_1(X_{1i}), \dots, \widetilde{g}_J(X_{Ji})]^\top$. Actually, this should be $\widehat{g}_j^{[0]}$ in place of $\widetilde{g}_j$, but we ignore the distinction for notational compactness.

LEMMA 2: *Under our assumptions, for any t, there is a stochastically bounded sequence $\delta_{n,t}$ such that*

$$\sqrt{L}(\widetilde{f}_t - f_t^\dagger - h^2\delta_{n,t}) = o_p(1).$$

PROOF: We use the matrix expansion $(I + \Delta)^{-1} = I - \Delta + (I + \Delta)^{-1}\Delta^2$ to obtain

$$\begin{aligned}
\widetilde{f}_t - f_t^\dagger &= \widetilde{A}^{-1}\widetilde{b}_t - A^{\dagger-1}b_t^\dagger \\
&= A^{\dagger-1}(\widetilde{b}_t - b_t^\dagger) - A^{\dagger-1}(\widetilde{A} - A^\dagger)A^{\dagger-1}b_t^\dagger \\
&\quad - A^{\dagger-1}(\widetilde{A} - A^\dagger)A^{\dagger-1}(\widetilde{b}_t - b_t^\dagger) \\
&\quad + A^{\dagger-1/2}\big[I + A^{\dagger-1/2}(\widetilde{A} - A^\dagger)A^{\dagger-1/2}\big]^{-1} \\
&\quad \times \widetilde{A}^{-1/2}(\widetilde{A} - A^\dagger)A^{\dagger-1}(\widetilde{A} - A^\dagger)A^{\dagger-1}\widetilde{b}_t.
\end{aligned} \tag{31}$$

The error $\|\widetilde{f}_t - f_t^\dagger\|$ is majorized by the errors $\|\widetilde{b}_t - b_t^\dagger\|$ and $\|\widetilde{A} - A^\dagger\|$ times constants due to the invertibility of $A^\dagger$. For example,

$$\begin{aligned}\|A^{\dagger-1}(\widetilde{b}_t - b_t^\dagger)\| &\le \lambda_{\max}(A^{\dagger-1})\|\widetilde{b}_t - b_t^\dagger\| \\ &= \frac{1}{\lambda_{\min}(A^\dagger)}\left(\sum_{j=0}^{J}(\widetilde{b}_{jt} - b_{jt}^\dagger)^2\right)^{1/2},\end{aligned}$$

$$\begin{aligned}\|A^{\dagger-1}(\widetilde{A} - A^\dagger)A^{\dagger-1}b_t^\dagger\| &\le \frac{\lambda_{\max}^{1/2}((\widetilde{A} - A^\dagger)^\top(\widetilde{A} - A^\dagger))}{\lambda_{\min}^2(A^\dagger)}\|b_t^\dagger\| \\ &\le \frac{\left(\sum_{j,k=0}^{J}(\widetilde{A}_{jk} - A_{jk}^\dagger)^2\right)^{1/2}}{\lambda_{\min}^2(A^\dagger)}\|b_t^\dagger\|,\end{aligned}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest (respectively) eigenvalues of a square symmetric matrix. Furthermore, $\lambda_{\min}(A^\dagger) \ge \lambda_{\min}(A) - o_p(1)$, where, by assumption, $\lambda_{\min}(A) > 0$. We establish the order in probability of the terms $\widetilde{b}_{jt} - b_{jt}^\dagger$ and $\widetilde{A}_{jk} - A_{jk}^\dagger$.

Consider the typical element in $\widetilde{b}_t - b_t^\dagger$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} y_{it}[\widetilde{g}_j(X_{ji}) - g_j(X_{ji})] \\ &\quad = \frac{1}{n}\sum_{i=1}^{n} q_{it}[\widetilde{g}_j(X_{ji}) - g_j(X_{ji})] + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{it}[\widetilde{g}_j(X_{ji}) - g_j(X_{ji})] \\ &\quad = T_{n1} + T_{n2},\end{aligned}$$

where $q_{it} = f_{ut} + \sum_{j=1}^{J} g_j(X_{ji})f_{jt}$. We consider the term T_{n1}. From (30), we have

$$\begin{aligned}T_{n1} &= \frac{1}{n}\sum_{i=1}^{n} q_{it}[\widetilde{g}_j(X_{ji}) - g_j(X_{ji})] \\ &= \frac{1}{n}\sum_{i=1}^{n} q_{it}\frac{1}{np_j(X_{ji})}\sum_{i'=1}^{n} K_h(X_{ji'}, X_{ji})\overline{\varepsilon}_{i'} + h^2\frac{1}{n}\sum_{i=1}^{n} q_{it}\beta_{n,j}(X_{ji}) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} q_{it}\frac{1}{n}\sum_{i'=1}^{n} s_n(X_{i'}, x_j)\overline{\varepsilon}_{i'} + \frac{1}{n}\sum_{i=1}^{n} q_{it}R_{nj}(X_{ji}) \\ &= T_{n11} + T_{n12} + T_{n13} + T_{n14},\end{aligned}$$

where R_{nj} is the remainder term in (30). The first term T_{n11} is essentially a degenerate U-statistic across clusters (Powell, Stock, and Stoker (1989)), that is, $T_{n11} = \sum_{\ell=1}^{L} \sum_{\ell'=1}^{L} \varphi_{n\ell,\ell'}$ with $\varphi_{n\ell,\ell'} = n^{-2} \sum_{m=1}^{M} \sum_{m'=1}^{M} q_{\ell m t} K_h(X_{j\ell' m'}, X_{j\ell m}) \overline{\varepsilon}_{\ell' m'} / p_j(X_{j\ell m})$. Then $E_\ell[\varphi_{n\ell,\ell'}] = 0$. Therefore, we can write $T_{n11} = \sum_{\ell=1}^{L} \varphi_{n\ell,\ell} + \sum_{\ell'=1}^{L} E_{\ell'}[\varphi_{n\ell,\ell'}] + \sum\sum_{\ell \neq \ell'} \widetilde{\varphi}_{n\ell,\ell'}$, where $\widetilde{\varphi}_{n\ell,\ell'} = \varphi_{n\ell,\ell'} - E_{\ell'}[\varphi_{n\ell,\ell'}]$ and, by construction, $E_\ell[\widetilde{\varphi}_{n\ell,\ell'}] = E_{\ell'}[\widetilde{\varphi}_{n\ell,\ell'}] = 0$. By straightforward moment calculations, it can be shown that $\sum_{\ell=1}^{L} \varphi_{n\ell,\ell} = O_p(n^{-3/2} h^{-1} \delta_T) = o_p(L^{-1/2})$. Specifically, we have

$$\begin{aligned}
\operatorname{var}(\varphi_{n\ell,\ell}) &= \frac{1}{n^4} E\Biggl[\sum_{m=1}^{M} \sum_{m'=1}^{M} \sum_{m''=1}^{M} \frac{q_{\ell m t} q_{\ell m'' t}}{p_j(X_{j\ell m}) p_j(X_{j\ell m'})} \\
&\qquad \times K_h(X_{j\ell m'}, X_{j\ell m}) K_h(X_{j\ell m'}, X_{j\ell m''}) \overline{\varepsilon}^2_{\ell m'}\Biggr] \\
&= O(M^3 n^{-4} \delta_T^2) + O(M^2 h^{-1} n^{-4} \delta_T^2) + O(M h^{-2} \delta_T^2 n^{-4}),
\end{aligned}$$

and then we use that $Mh \to 0$ to conclude that the third term in the above display is the largest. Furthermore, we have

$$\begin{aligned}
E_{\ell'}[\varphi_{n\ell,\ell'}] &= \frac{M}{n^2} \sum_{m'=1}^{M} \overline{\varepsilon}_{\ell' m'} E_{\ell'}\left[\frac{q_{\ell m t} K_h(X_{j\ell' m'}, X_{j\ell m})}{p_j(X_{j\ell m})}\right] \\
&\simeq \frac{M^2}{n^2} \frac{1}{M} \sum_{m'=1}^{M} \overline{\varepsilon}_{\ell' m'} \overline{q}_{\ell' m' t},
\end{aligned}$$

where $\overline{q}_{\ell' m' t} = E[(f_{ut} + \sum_{j'=1}^{J} g_{j'}(X_{j'\ell m}) f_{j't}) / p_j(X_{j\ell m}) | X_{j\ell' m'}]$. It follows that $\operatorname{var}(E_{\ell'}[\varphi_{n\ell,\ell'}]) = O(n^{-4} M^4 \delta_T^2)$ and so $\sum_{\ell'=1}^{L} E_{\ell'}[\varphi_{n\ell,\ell'}] = O_p(L^{-3/2} \delta_T)$. Furthermore, $\operatorname{var}(\widetilde{\varphi}_{n\ell,\ell'}) = O(M h^{-2} \delta_T^2 n^{-4})$ so that $\sum\sum_{\ell \neq \ell'} \widetilde{\varphi}_{n\ell,\ell'} = O_p((L^2 M h^{-2} \delta_T^2 \times n^{-4})^{1/2}) = o_p(L^{-1/2} \delta_T)$. Furthermore, $T_{n12} = O_p(h^2)$ and $T_{n13}, T_{n14} = o_p(L^{-1/2} \times \delta_T)$. Therefore, $\widetilde{b}_t - b_t^\dagger = O_p(L^{-1/2} \delta_T) + O_p(h^2)$.

Likewise, the typical element in $\widetilde{A} - A^\dagger$ satisfies

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} [\widetilde{g}_j(X_{ji}) \widetilde{g}_k(X_{ki}) - g_j(X_{ji}) g_k(X_{ki})] \\
&\quad = O_p(h^2) + O_p(L^{-1/2} \delta_T) + O_p(n^{-1} T^{-1} h^{-1}).
\end{aligned}$$

It follows that provided $Lh^4 \to 0$, $\sqrt{L}(\widetilde{f}_t - f_t^\dagger) = o_p(1)$. More generally, letting

$$\delta_{n,t} = \frac{1}{n} \sum_{i=1}^{n} \left[f_{ut} + \sum_{j=1}^{J} g_j(X_{ji}) f_{jt} \right] \beta_{n,j}(X_{ji}),$$

we have $\sqrt{L}(\widetilde{f}_t - f_t^{\dagger} - h^2\delta_{n,t}) = o_p(1)$. *Q.E.D.*

We now turn to the uniform over t properties, (26). By the triangle inequality, $\max_{1\leq t\leq T}\|\widetilde{f}_t - f_t\| \leq \max_{1\leq t\leq T}\|\widetilde{f}_t - f_t^{\dagger}\| + \max_{1\leq t\leq T}\|f_t^{\dagger} - f_t\|$. We first examine $\max_{1\leq t\leq T}\|f_t^{\dagger} - f_t\|$.

LEMMA 3: *Under our assumptions*,

$$\max_{1\leq t\leq T}\|f_t^{\dagger} - f_t\| = O_p\big(L^{-1/2}(\log T)^{\rho}\big).$$

PROOF: By extending the argument of Lemma 1, there is a finite constant C such that

$$\max_{1\leq t\leq T}\|f_t^{\dagger} - f_t\| \leq (C + o_p(1)) \max_{j=u,1,\ldots,J}\max_{1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{it}g_j(X_{ji})\right|$$

with $g_u = 1$. Since J is finite, it suffices to show that for each j,

$$\max_{1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{it}g_j(X_{ji})\right| = \max_{1\leq t\leq T}\left|\frac{1}{L}\sum_{\ell=1}^{L}\varepsilon_{j\ell t}^{\blacktriangledown}\right| = O_p\big(L^{-1/2}(\log T)^{\rho}\big) \tag{32}$$

for some $\rho > 0$, where $\varepsilon_{j\ell t}^{\blacktriangledown} = M^{-1}\sum_{m=1}^{M}\varepsilon_{\ell mt}g_j(X_{j\ell m})$.

Let $\varepsilon_{j\ell t}^{+} = \varepsilon_{j\ell t}^{\blacktriangledown}1(|\varepsilon_{j\ell t}^{\blacktriangledown}| \leq (LT)^{1/\kappa}) - E[\varepsilon_{j\ell t}^{\blacktriangledown}1(|\varepsilon_{j\ell t}^{\blacktriangledown}| \leq (LT)^{1/\kappa})]$. Then $1 - \Pr[|\varepsilon_{j\ell t}^{\blacktriangledown}| \leq (LT)^{1/\kappa}$ for $1 \leq t \leq T$ and $1 \leq \ell \leq L] \leq LT\Pr[|\varepsilon_{j\ell t}^{\blacktriangledown}| > (LT)^{1/\kappa}] \leq E[|\varepsilon_{j\ell t}^{\blacktriangledown}|^{\kappa}1(|\varepsilon_{j\ell t}^{\blacktriangledown}| > (LT)^{1/\kappa})] \to 0$. We now apply the Bonferroni and exponential inequalities to $\max_{1\leq t\leq T}|\frac{1}{L}\sum_{\ell=1}^{L}\varepsilon_{j\ell t}^{+}|$. In particular, letting $\tau_{LT}^2 = \inf_{1\leq t\leq T}\operatorname{var}[\sum_{\ell=1}^{L}\varepsilon_{j\ell t}^{+}]$, we have

$$\begin{aligned}\Pr\left[\max_{1\leq t\leq T}\left|\sum_{\ell=1}^{L}\varepsilon_{j\ell t}^{+}\right| > KL^{1/2}\right] &\leq \sum_{t=1}^{T}\Pr\left[\left|\sum_{\ell=1}^{L}\varepsilon_{j\ell t}^{+}\right| > KL^{1/2}\right]\\ &\leq 2T\exp\left(-\frac{LK^2}{2\tau_{LT}^2 + 2(LT)^{1/\kappa}L^{1/2}K/3}\right).\end{aligned}$$

By taking $K = (\log T)^{\rho}$, the right-hand side is $o(1)$ provided $\kappa > 4$. *Q.E.D.*

LEMMA 4: *Under our assumptions*,

$$\max_{1\leq t\leq T}\|\widetilde{f}_t - f_t^{\dagger}\| = O_p\big(L^{-1/2}\delta_T(\log T)^{\rho}\big) + O_p(h^2(\log T)^{\rho'}).$$

PROOF: As before, we apply the triangle inequality again to each term in (31). We have $\max_{1\le t\le T}\|b_t^\dagger\| = O_p((\log T)^{\rho\prime})$, so it suffices to bound the terms $\max_{1\le t\le T}|\widetilde{b}_{jt} - b_{jt}^\dagger|$ and $\max_{1\le t\le T}|\widetilde{A}_{jk} - A_{jk}^\dagger|$. We just show that

$$\max_{1\le t\le T}\left|\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}\left[f_{ut} + \sum_{j=1}^{J} g_j(X_{j\ell m})f_{jt}\right][\widetilde{g}_j(X_{j\ell m}) - g_j(X_{j\ell m})]\right| \tag{33}$$
$$= O_p\big(L^{-1/2}(\log T)^{\rho}\big) + O_p(h^2(\log T)^{\rho\prime}),$$

$$\max_{1\le t\le T}\left|\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}\varepsilon_{\ell mt}[\widetilde{g}_j(X_{j\ell m}) - g_j(X_{j\ell m})]\right| = O_p\big(L^{-1/2}(\log T)^{\rho}\big). \tag{34}$$

This uses the same type of techniques as above. In particular, we have

$$\begin{aligned}
&\max_{1\le t\le T}\left|h^2\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}\left[f_{ut} + \sum_{j=1}^{J} g_j(X_{j\ell m})f_{jt}\right]\beta_j(X_{j\ell m})\right| \\
&\le h^2\Bigg(\max_{1\le t\le T}|f_{ut}|\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}|\beta_j(X_{j\ell m})| \\
&\qquad + \sum_{j=1}^{J}\max_{1\le t\le T}|f_{jt}|\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}|g_j(X_{j\ell m})\|\beta_j(X_{j\ell m})|\Bigg) \\
&= O_p(h^2(\log T)^{\rho'}).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\max_{1\le t\le T}\left|\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{M}\sum_{m=1}^{M}\overline{\varepsilon}_{\ell m}\left[f_{ut} + \sum_{j'=1}^{J} E[g_{j'}(X_{j'\ell m})|X_{j\ell m}]f_{j't}\right]\right| \\
&= O_p\big(L^{-1/2}\delta_T(\log T)^{\rho'}\big).
\end{aligned}$$

In conclusion we have shown $\max_{1\le t\le T}\|\widetilde{f}_t - f_t\| = O_p((L^{-1/2}\delta_T(\log T)^{\rho}) + O_p(h^2(\log T)^{\rho'})$. *Q.E.D.*

Finally, we establish the asymptotic distribution of $\widehat{g}_j(x_j)$. Consider the one-step estimator

$$\widehat{g}_j^{[1]}(x_j)=\frac{\displaystyle\sum_{t=1}^{T}\widetilde{f}_{jt}\left[\widehat{\lambda}_{1t}(j,x_j)-\widetilde{f}_{ut}-\sum_{k\neq j}\widetilde{f}_{kt}\widetilde{\lambda}_2(j,k,x_j)\right]}{\displaystyle\sum_{t=1}^{T}\widetilde{f}_{jt}^2},$$

$$\widehat{\lambda}_{1t}(j,x_j)=\frac{\displaystyle\sum_{i=1}^{n}K_h(X_{ji},x_j)y_{it}}{\displaystyle\sum_{i=1}^{n}K_h(X_{ji},x_j)},$$

$$\widetilde{\lambda}_2(j,k,x_j)=\frac{\displaystyle\sum_{i=1}^{n}K_h(X_{ji},x_j)\widetilde{g}_k(X_{ki})}{\displaystyle\sum_{i=1}^{n}K_h(X_{ji},x_j)}.$$

Define the infeasible estimator

$$\widetilde{g}_j^{[1]}(x_j)=\frac{\displaystyle\sum_{t=1}^{T}f_{jt}\left[\widehat{\lambda}_{1t}(j,x_j)-f_{ut}-\sum_{k\neq j}f_{kt}\widetilde{\lambda}_2(j,k,x_j)\right]}{\displaystyle\sum_{t=1}^{T}f_{jt}^2}.$$

LEMMA 5: *Under our assumptions*,

$$\widehat{g}_j^{[1]}(x_j)-g_j(x_j)=\widetilde{g}_j^{[1]}(x_j)-g_j(x_j)+O_p\big(L^{-1/2}\delta_T\big)+O_p(h^2).$$

PROOF: We expand $\widehat{g}_j^{[1]}(x_j)$ about $\widetilde{g}_j^{[1]}(x_j)$ in a Taylor expansion in $\widetilde{f}_{jt}-f_{jt}$ and $\widetilde{g}_k(X_{ki})-g_k(X_{ki})$ and obtain many terms. A typical term is $\sum_{t=1}^{T}(\widetilde{f}_{jt}-f_{jt})f_{jt}g_j(x_j)/\sum_{t=1}^{T}f_{jt}^2$. Then

$$\frac{1}{T}\sum_{t=1}^{T}(\widetilde{f}_{jt}-f_{jt})f_{jt}=\frac{1}{T}\sum_{t=1}^{T}(\widetilde{f}_{jt}-f_{jt}^{\dagger})f_{jt}+\frac{1}{T}\sum_{t=1}^{T}(f_{jt}^{\dagger}-f_{jt})f_{jt},\tag{35}$$

where $T^{-1}\sum_{t=1}^{T}(f_{jt}^{\dagger}-f_{jt})f_{jt}=(A^{\dagger})^{-1}T^{-1}\sum_{t=1}^{T}f_{jt}n^{-1}\sum_{i=1}^{n}G(X_i)\varepsilon_{it}=O_p(L^{-1/2}T^{-1/2})$. The expansion for $T^{-1}\sum_{t=1}^{T}(\widetilde{f}_{jt}-f_{jt}^{\dagger})f_{jt}$ is more complicated, but one obtains terms like

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T} f_{jt}\frac{1}{n}\sum_{i=1}^{n} y_{it}[\widetilde{g}_j(X_{ji})-g_j(X_{ji})]\\&\quad=\frac{1}{T}\sum_{t=1}^{T} f_{jt}\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{it}[\widetilde{g}_j(X_{ji})-g_j(X_{ji})]\\&\qquad+\frac{1}{T}\sum_{t=1}^{T} f_{jt}\frac{1}{n}\sum_{i=1}^{n}\left[f_{ut}+\sum_{j=1}^{J} g_j(X_{ji})f_{jt}\right][\widetilde{g}_j(X_{ji})-g_j(X_{ji})].\end{aligned}$$

The double averaging makes the leading stochastic terms $O_p(L^{-1/2}T^{-1/2})$ the bias terms are always $O_p(h^2)$, and we cannot eliminate the remainder term of $O_p(L^{-1/2}\delta_T)$. *Q.E.D.*

Define

$$\widetilde{U}_{n1j}=\frac{1}{p_j(x_j)}\frac{1}{n}\sum_{i=1}^{n} K_h(X_{ji},x_j)\widetilde{\varepsilon}_{ji},\quad \widetilde{\varepsilon}_{ji}=\frac{\sum_{t=1}^{T} f_{jt}\varepsilon_{it}}{\sum_{t=1}^{T} f_{jt}^2}. \tag{36}$$

LEMMA 6: *Under our assumptions*,

$$\begin{aligned}&\widetilde{g}_j^{[1]}(x_j)-g_j(x_j)=\widetilde{U}_{n1j}+O_p(h^2)+o_p\big(n^{-1/2}T^{-1/2}h^{-1/2}\big),\\&\sqrt{nTh}\,\widetilde{U}_{n1j}\Longrightarrow N(0,\Omega_j(x_j)).\end{aligned}$$

PROOF: The term $\widetilde{U}_{n1j}$ is a sum of independent random variables, and is $O_p(n^{-1/2}h^{-1/2}T^{-1/2})$ with mean zero and variance as stated in Theorem 1. Specifically, write

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_{ji},x_j)\widetilde{\varepsilon}_{ji}=\frac{1}{L}\sum_{\ell=1}^{L}\xi_{j\ell},\quad \xi_{j\ell}=\frac{1}{M}\sum_{m=1}^{M} K_h(X_{j\ell m},x_j)\widetilde{\varepsilon}_{j\ell m}$$

with $\xi_{j\ell}$ independent across ℓ. We have

$$\begin{aligned}\operatorname{var}(\xi_{j\ell})&=\frac{1}{M^2}\sum_{m=1}^{M} E[K_h(X_{j\ell m},x_j)^2\widetilde{\varepsilon}_{j\ell m}^2]\\&\quad+\frac{1}{M^2}\sum_{\substack{m=1\\ m\neq m'}}^{M}\sum_{m=1}^{M} E[K_h(X_{j\ell m},x_j)K_h(X_{j\ell m'},x_j)\widetilde{\varepsilon}_{j\ell m}\widetilde{\varepsilon}_{j\ell m'}]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M} E[K_h(X_{j\ell m}, x_j)^2 \widetilde{\varepsilon}^2_{j\ell m}] \\
&\quad + \frac{M(M-1)}{M^2} E[K_h(X_{j\ell m}, x_j) K_h(X_{j\ell m'}, x_j) \widetilde{\varepsilon}_{j\ell m} \widetilde{\varepsilon}_{j\ell m'}].
\end{aligned}
$$

Note that $E[\widetilde{\varepsilon}_{j\ell m}|X_{j\ell m} = x_j] = 0$ and

$$
\operatorname{var}[\widetilde{\varepsilon}_{j\ell m}|X_{j\ell m} = x_j] = \frac{\sum_{t=1}^{T} f_{jt}^2 \sigma_{jt}^2(x_j)}{\left(\sum_{t=1}^{T} f_{jt}^2\right)^2} \leq \frac{C}{T}
$$

for some $C < \infty$ for large enough T. Therefore, $\operatorname{var}(\xi_{j\ell}) = O(M^{-1}T^{-1}h^{-1}) + O(T^{-1})$. Since we assumed that $Mh \to 0$, it is the first term that dominates and $\widetilde{U}_{n1j} = O_p(L^{-1/2}M^{-1/2}T^{-1/2}h^{-1/2})$. Furthermore, we can apply the Lindeberg's CLT for independent random variables. The quantities $\widetilde{U}_{n1j}(x_j)$ and $\widetilde{U}_{n1k}(x_k)$ are asymptotically independent by standard arguments for kernels.

Consider

$$
\begin{aligned}
&\frac{\sum_{t=1}^{T} f_{jt}\left[\widehat{\lambda}_{1t}(j, x_j) - f_{ut} - \sum_{k\neq j} f_{kt}\widetilde{\lambda}_2(j, k, x_j)\right]}{\sum_{t=1}^{T} f_{jt}^2} - g_j(x_j) \\
&= \frac{1}{\sum_{t=1}^{T} f_{jt}^2} \sum_{t=1}^{T} f_{jt} \\
&\quad \times \left\{ \frac{\sum_{i=1}^{n} K_h(X_{ji}, x_j)[y_{it} - f_{ut} - \sum_{k\neq j} f_{kt}\widetilde{g}_k(X_{ki})]}{\sum_{i=1}^{n} K_h(X_{ji}, x_j)} - f_{jt}g_j(x_j) \right\} \\
&= \frac{1}{\sum_{t=1}^{T} f_{jt}^2} \sum_{t=1}^{T} f_{jt} \frac{\sum_{i=1}^{n} K_h(X_{ji}, x_j)\varepsilon_{it}}{\sum_{i=1}^{n} K_h(X_{ji}, x_j)}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{\sum_{t=1}^{T} f_{jt}^2}\sum_{t=1}^{T} f_{jt}\frac{\sum_{i=1}^{n} K_h(X_{ji},x_j)[g_j(X_{ji})-g_j(x_j)]}{\sum_{i=1}^{n} K_h(X_{ji},x_j)}\\
&-\frac{1}{\sum_{t=1}^{T} f_{jt}^2}\sum_{t=1}^{T} f_{jt}\frac{\sum_{i=1}^{n} K_h(X_{ji},x_j)\sum_{k\neq j} f_{kt}[\widetilde{g}_k(X_{ki})-g_k(X_{ki})]}{\sum_{i=1}^{n} K_h(X_{ji},x_j)}\\
=\;& U_{nj1}+U_{nj2}+U_{nj3}.
\end{aligned}
$$

By the same arguments as above, $\widehat{p}_j(x_j)=n^{-1}\sum_{i=1}^{n}K_h(X_{ji},x_j)=p_j(x_j)+O(h^2)+O_p(L^{-1/2})+O_p(n^{-1/2}h^{-1/2})$. Therefore,

$$
U_{nj1}=\frac{1}{n}\sum_{i=1}^{n}K_h(X_{ji},x_j)\widetilde{\varepsilon}_{ji}\frac{1}{p_j(x_j)(1+o_p(1))}.
$$

The term U_{nj2} is a bias term of order h^2. The term U_{nj3} can be shown to be $O_p(h^2)+o_p(n^{-1/2}T^{-1/2}h^{-1/2})$. *Q.E.D.*

Now define $\widehat{f}_t^{[1]}$ as in (19). Using the above expansion it can be shown that the results of Lemmas 2 and 4 continue to hold with $\widehat{f}_t^{[1]}$ replacing $\widetilde{f}_t$ with a difference sequence $\delta_{n,t}$. Then we can show that the conclusion of Lemmas 5 and 6 continue to hold with $\widehat{g}_j^{[2]}$ replacing $\widehat{g}_j^{[1]}$. This process can be continued for any finite number of iterations; see Linton, Nielsen, and Van der Geer (2003). The only thing that changes in the expansions is the bias function, although it can still be approximated by some bounded continuous function. If the initial estimator has a slower initial rate, then more iterations are needed to guarantee the same result. *Q.E.D.*

REFERENCES

ANG, A., R. HODRICK, Y. XING, AND X. ZHANG (2006): "The Cross-Section of Volatility and Expected Returns," *Journal of Finance*, 61, 259–299. [715,732,734,740]

——— (2009): "High Idiosyncratic Risk and Low Returns: International and Further U.S. Evidence," *Journal of Financial Economics*, 91, 1–23. [715,732,734,738]

BAI, J. (2003): "Inferential Theory for Factor Models of Large Dimension," *Econometrica*, 71 (1), 135–171. [715,717]

——— (2004): "Estimating Cross-Section Common Stochastic Trends in Nonstationary Panel Data," *Journal of Econometrics*, 122 (1), 137–183. [715,717]

BAI, J., AND S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [715,717]

BALTAGI, B. H., J. HIDALGO, AND Q. LI (1996): "A Nonparametric Test for Poolability Using Panel Data," *Journal of Econometrics*, 75, 345–367. [732]

BREIMAN, L., AND J. H. FRIEDMAN (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598. [723]

CARHART, M. M. (1997): "On Persistence of Mutual Fund Performance," *Journal of Finance*, 52, 57–82. [715,732]

CARRASCO, M., J. P. FLORENS, AND E. RENAULT (2006), "Linear Inverse Problems in Structural Econometrics," in *The Handbook of Econometrics*, Vol. 6, ed. by J. J. Heckman and E. Leamer. Amsterdam: North Holland. [714]

CONNOR, G., AND R. A. KORAJCZYK (1988): "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics*, 21, 255–289. [716]

——— (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48, 1263–1288. [716,717,731]

CONNOR, G., AND O. B. LINTON (2007): "Semiparametric Estimation of a Characteristic-Based Factor Model of Stock Returns," *Journal of Empirical Finance*, 14, 694–717. [713,716,720,726, 727]

CONNOR, G., M. HAGMANN, AND O. LINTON (2012): "Supplement to 'Efficient Semiparametric Estimation of the Fama–French Model Extensions'," *Econometrica Supplemental Material*, 80, http://www.econometricsociety.org/ecta/Supmat/7432_data and programs.zip. [731]

FAMA, E. F., AND K. R. FRENCH (1993): "Common Risk Factors in the Returns to Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56. [713,718,719,732,738,740,741]

FAN, J., AND Q. YAO (1998): "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrika*, 85, 645–660. [731]

GOYAL, A., AND P. SANTA-CLARA (2003): "Idiosyncratic Risk Matters!" *Journal of Finance*, 58, 975–1008. [715,734]

GOZALO, P. L., AND O. B. LINTON (2001): "Testing Additivity in Generalized Nonparametric Regression Models With Estimated Parameters," *Journal of Econometrics*, 104, 1–48. [732]

HASTIE, T., AND R. J. TIBSHIRANI (1990): *Generalized Additive Models*. London: Chapman and Hall. [723]

HSIAO, C. (2003), *Analysis of Panel Data* (Second ed.). Econometric Society Monograph, Vol. 34. Cambridge: CUP. [716]

JAGADEESH, N., AND S. TITMAN (1993): "Returns to Buying Winners and Selling Losers: Implications of Stock Market Efficiency," *Journal of Finance*, 48, 65–92. [715]

KYRIAZIDOU, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364. [714]

LEWBEL, A., AND O. B. LINTON (2007): "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," *Econometrica*, 75, 1209–1227. [724]

LINTON, O. B. (1997): "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika*, 84, 469–473. [730]

LINTON, O. B., AND E. MAMMEN (2005): "Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods," *Econometrica*, 73, 771–836. [714,722,728]

——— (2008): "Nonparametric Transformation to White Noise," *Journal of Econometrics*, 142, 241–264. [714]

LINTON, O. B., AND J. P. NIELSEN (1995): "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–100. [716,732]

LINTON, O. B., J. P. NIELSEN, AND S. VAN DE GEER (2003): "Estimating Multiplicative and Additive Hazard Functions by Kernel Methods," *The Annals of Statistics*, 31 (2), 464–492. [714,752]

MAMMEN, E., AND B. PARK (2005): "Bandwidth Selection for Smooth Backfitting in Additive Models," *The Annals of Statistics*, 33, 1260–1294. [731]

MAMMEN, E., O. LINTON, AND J. P. NIELSEN (1999): "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *The Annals of Statistics*, 27, 1443–1490. [714,722,725,726,728,742]

MAMMEN, E., B. STØVE, AND D. TJØSTHEIM (2009): "Nonparametric Additive Models for Panels of Time Series," *Econometric Theory*, 25, 442–481. [714]

PESARAN, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels With a Multifactor Error Structure," *Econometrica*, 74, 967–1012. [715,726]

PHILLIPS, P. C. B., AND H. R. MOON (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111. [715,728]

PORTER, J. (1996): "Nonparametric Regression Estimation for a Flexible Panel Data Model," Ph.D. Thesis, Department of Economics, MIT. [714]

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. [746]

ROBINSON, P. M. (2007): "Nonparametric Regression With Spatial Data," Working Paper, STICERD. [731]

ROSENBERG, B. (1974): "Extra-Market Components of Covariance in Security Prices," *Journal of Financial and Quantitative Analysis*, 9, 263–274. [713,727,741]

ROSENBLATT, M. (1956): "A Central Limit Theorem and Strong Mixing Conditions," *Proceedings of the National Academy of Science U.S.A.*, 42, 43–47. [728]

ROSS, S. A. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360. [717]

STONE, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360. [714,730]

TIBSHIRANI, R. (1984): "Local Likelihood Estimation," Ph.D. Thesis, Stanford University. [731]

WOOLDRIDGE, J. M. (2006): "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," Working Paper, MSU. [728,731]

Dept. of Economics, National University of Ireland, Maynooth, Ireland; gregory.connor@nuim.ie,

Oxford-Man Institute, University of Oxford, Eagle House, Oxford, OX2 65D U.K. and AHL; mhagmann@ahl.com,

and

Faculty of Economics, Cambridge University, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD U.K.; obl20@cam.ac.uk.

Manuscript received September, 2007; final revision received August, 2010.