Dimension Reduction Using
Inverse Regression and Nonparametric Factors

July, 2008

Pian Chen  
School of Economics and Finance  
Victoria University of Wellington  
pian.chen@vuw.ac.nz

and

Aaron Smith  
Department of Agricultural and Resource Economics  
University of California, Davis  
adsmith@ucdavis.edu

Abstract: We develop a nonparametric method to reduce the dimension of econometric models that contain many explanatory variables and many regression functions. To enable nonparametric regression, we reduce the many explanatory variables to a few inverse regression variates. We then reduce the many regression functions to a few nonparametric factors. Not only may these nonparametric factors represent economically meaningful objects, but they also facilitate high dimensional data visualization and parsimonious model specification. To estimate the number of nonparametric factors, we propose a nonparametric rank test that is root-$n$ consistent and applies to near epoch dependent data. We apply this method to financial asset pricing by studying the relationship between the excess returns on 25 stock portfolios and 18 macroeconomic variables. Our method reduces the 18 explanatory variables to one inverse regression variate, and it reduces the 25 regression functions to two nonparametric factors. The first factor captures the essence of the CAPM, and the second factor is a nonlinear term that represents downside risk.

Key words: curse of dimensionality, variable space, function space, $U$-statistic, near-epoch dependent process, $\alpha$-mixing, local power, data visualization, model specification, financial asset pricing
1. Introduction

High dimensionality hampers econometric analysis of models with many explanatory variables and many regression functions. To make useful inference in such settings, econometricians aim to reduce the dimension of the data to a parsimonious level while preserving relevant information. This dimension reduction is often carried out by imposing parametric structures, excluding variables, or aggregating data. Rather than imposing such restrictions a priori, we propose a data-driven method to reduce the dimension in two spaces: the variable space spanned by the explanatory variables and the function space spanned by the regression functions. We reduce the variable space to a few inverse regression variates and reduce the function space to a few nonparametric factors.

We demonstrate how to visualize high dimensional data using estimated inverse regression variates and nonparametric factors. Data visualization can often identify interesting relationships among variables, but a high dimensional system with an unknown joint distribution cannot be fully represented by arbitrary bivariate plots (Carroll and Li 1992, Cook 1994). However, by reducing the system to a set of orthogonal nonparametric factors and a set of uncorrelated inverse regression variates, we can fully characterize a high dimensional system using a minimum number of bivariate plots. These plots can reveal underlying structure previously hidden in high dimensional data and provide guidance for parsimonious model specification.

Specifically, we study the nonparametric model

$$Y_i = F_0(X_i) + U_i \quad (i = 1, 2, \ldots, n),$$

where $Y_i$ represents a $G$-vector of dependent variables, $X_i$ denotes $J$ explanatory variables, $F_0(X_i)$ is a $G$-vector of unknown functions of $X_i$, $E(U_i|X_i) = 0$, and $E(U_i'U_i'|X_i) = \Sigma$. We reduce the dimension of the function space by decomposing $F_0(X_i)$ as

$$F_0(X_i) = AH(X_i),$$

where $A$ is a $G \times L$ constant matrix and $H(X_i)$ is an $L$-vector of unknown functions with $L \leq G$. This decomposition implies that $F_0(X_i)$ falls into a function space spanned by $L$ basis functions $(H_1, \ldots, H_L)$. Following Donald (1997), we call these unknown basis functions nonparametric factors.
Equation (2) arises naturally in economics. For example, in asset pricing models, $Y_i$ denotes a vector of excess stock returns, $X_i$ contains variables that capture aggregate market risk, and each nonparametric function represents an aggregate risk factor. The capital asset pricing model (CAPM, Sharpe 1964, Lintner 1965) specifies one factor that is a linear function of the market portfolio; the arbitrage pricing model (APT, Ross 1976) specifies multiple linear factors; Fama and French (1992, 1993) specify size and book-to-market factors in addition to the market portfolio; Chen, Roll, and Ross (1986) specify factors that are linear functions of several macroeconomic variables. Equation (2) generalizes these linear parametric models by allowing the aggregate risk factors to be nonparametric functions of a potentially large set of observed $X$ variables. Other examples of the model in (2) include demand systems and nonparametric instrumental variables (Donald 1997).

To estimate $(H_1, \ldots, H_L)$, we need to estimate $F_0(X_i)$ nonparametrically. In applications with many $X$ variables, such estimation is hindered by the curse of dimensionality. To overcome this problem, we use inverse regression, which is based on the premise that a small number of linear combinations of $X$ can capture the information in $X$ that is relevant for $Y$ (Li 1991). Therefore, we can replace $J$-dimensional $X$ with $d$ linear combinations of $X$, i.e., $F_0(X_i) = F_1(X_iB)$, and redefine the nonparametric factors as unknown functions of $X_iB$, i.e.,

\begin{align*}
F_1(X_iB) = AH(X_iB),
\end{align*}

where $B = (\beta_1, \ldots, \beta_d)$ is a $J \times d$ constant matrix with $d \leq J$. Replacement of $X_i$ with $X_iB$ entails an orthogonal projection that transforms a large set of correlated $X$ variables into a small set of uncorrelated inverse regression variates $XB$.

The matrix $B$ cannot be uniquely identified from the data, but the space spanned by $B$ is identifiable. Li (1991) proposes the sliced inverse regression method, which generates root-$n$ consistent estimates of this dimension reduction space without knowledge of the link function. Because the inverse regression estimates converge at a faster rate than standard nonparametric statistics, substituting the estimated inverse regression variates for the observed $X$ variables does not affect the asymptotic distribution of nonparametric statistics in general. Consequently, in addition to the function space reduction problem we treat in this paper, inverse regression can be applied to many other testing and estimation problems in nonparametric econometrics.

---

1 For notational convenience, we define $X_i$ as the $1 \times J$ row vector that inhabits the rows of the $n \times J$ matrix $X$. 

The rest of the paper is organized as follows. In Section 2, we introduce inverse regression for variable space reduction and compare it with other dimension reduction methods such as principal components and least squares regression. We present a new result that, under the assumption of a linear link function, inverse regression is identical to OLS and canonical correlation. In Section 3, we discuss function space reduction and data visualization for parsimonious model specification. In Section 4, we present a nonparametric rank test for the dimension of the reduced function space. Our test has nontrivial power against local alternatives that converge to the null at rate $O(n^{-1/2})$. Allowing for general dependence in the data, we derive the limiting null distribution of our test statistic. Our test statistic is not asymptotically pivotal, so we propose a resampling strategy to generate critical values. We then demonstrate the finite sample performance of the test via Monte Carlo simulations.

In Section 5, we apply our dimension-reduction method to estimate the aggregate risk factors needed to explain 25 excess stock returns, where each risk factor is a nonparametric function of 18 macroeconomic variables. We identify two risk factors. The first is linear and in line with the CAPM, while the second is a nonlinear downside risk factor that accounts for the size- and value-premium puzzles. Section 6 concludes the paper, and the appendices contain the proofs of all theorems.

2. Variable Space Reduction Using Inverse Regression

Inverse regression is a dimension reduction method first proposed by Li (1991). In this section, we introduce the fundamental theory, outline the estimation procedure, and compare inverse regression to principal component analysis, linear regression, and canonical correlations.

2.1 Inverse Regression: Theory and Estimation

Inverse regression methods assume that the distribution of $Y_i$ is determined only through $d$ linear combinations of $X_i$, i.e.,

$$Y_i = F(X_i, \beta_1, X_i, \beta_2, \cdots, X_i, \beta_d, U_i),$$

where $d \leq J$, $B = (\beta_1, \beta_2, \cdots, \beta_d)$ are $J$-dimensional column vectors known as “effective dimension reduction” (e.d.r.) directions, the link function $F$ is unknown, and $X_i B$ are called inverse regression variates. This multi-index model nests many models widely used in economics.
For example, if \( d = 1 \) and \( F \) is linear, then it is simply a linear regression model, in which one linear combination of \( X \) is sufficient to explain \( Y \). On the other hand, if \( d = 1 \) and \( F \) is nonlinear, it is a single index model. Logit and probit models can be viewed as special cases in which \( Y \) is binary and \( F \) takes certain parametric functional forms. If \( d = J \), then variable space reduction is impossible and \( B \) can be an identity matrix.

Li (1991) shows how to estimate \( B \) using inverse regressions of the explanatory variables on \( Y \). For easy exposition, we standardize \( X_i \) to \( Z_i = \Sigma_{XX}^{-1/2} [X_i - E(X_i)] \), where \( \Sigma_{XX} \) denotes the covariance matrix of \( X_i \). Li (1991) shows that under the model defined in (4), the centered inverse regression curve \( E(Z_i | Y_i) \) is contained in the linear subspace spanned by \( \eta_k = \Sigma_{XX}^{1/2} \beta_k \) \((k = 1, 2, \ldots, d)\). This result implies that the covariance matrix \( \text{Cov}[E(Z_i | Y_i)] \) is degenerate in any direction orthogonal to \((\eta_1, \eta_2, \ldots, \eta_d)\). Therefore, \((\eta_1, \eta_2, \ldots, \eta_d)\) are the eigenvectors associated with the \( d \) nonzero eigenvalues of \( \text{Cov}[E(Z_i | Y_i)] \), and the \( e.d.r. \) directions are \( \beta_k = \Sigma_{XX}^{-1/2} \eta_k \) \((k = 1, 2, \ldots, d)\).

The eigen decomposition of \( \text{Cov}[E(Z_i | Y_i)] \) requires an estimate of \( E(Z_i | Y_i) \). For univariate \( Y_i \), Li (1991) suggests using the mean values of \( Z_i \) within several intervals or slices determined by \( Y_i \), which leads to the following sliced inverse regression (SIR) algorithm:

**Step 1:** Standardize \( X_i \) by an affine transformation to yield \( \hat{Z}_i = \hat{\Sigma}_{XX}^{-1/2} (X_i - \bar{X}) \) \((i = 1, 2, \ldots, n)\), where \( \hat{\Sigma}_{XX} \) and \( \bar{X} \) are the sample covariance and sample mean of \( X \), respectively.

**Step 2:** Divide the range of \( Y \) into \( S \) slices, denoted by \((I_1, \ldots, I_S)\). Let the proportion of the \( Y_i \)'s that fall in slice \( s \) be \( \hat{p}_s \) \((s = 1, 2, \ldots, S)\).

**Step 3:** Within each slice, compute the sample mean of \( \hat{Z}_i \), denoted by \( \hat{m}_s \), such that the sliced mean \( \hat{m}_s = (n \hat{p}_s)^{-1} \sum_{Y_i \in I_s} \hat{Z}_i \).

**Step 4:** Calculate the weighted variance-covariance matrix of the sliced means \( \hat{M}_{SIR} = \sum_{s=1}^{S} \hat{p}_s \hat{m}_s \hat{m}_s' \), and then find the eigenvalues and eigenvectors of \( \hat{M}_{SIR} \).

**Step 5:** Let \( \hat{\eta}_k \) \((k = 1, 2, \ldots, d)\) be the \( d \) largest eigenvectors of \( \hat{M}_{SIR} \). The outputs, \( \hat{\beta}_k = \hat{\Sigma}_{XX}^{-1/2} \hat{\eta}_k \) \((k = 1, 2, \ldots, d)\), are the estimates of the \( e.d.r. \) directions.
Chen and Li (1998) show that $\hat{\beta}_k$ from SIR is consistent and asymptotically normal with a covariance matrix that can be approximated by $n^{-1}\hat{\lambda}_k^{-1}(1-\hat{\lambda}_k)\hat{S}^{-1}_{XX}$, where $\hat{\lambda}_k$ is the $k^{th}$ largest eigenvalue of the matrix $\hat{M}_{\text{SIR}}$ in step 4. The order of this covariance matrix illustrates the root-$n$ consistency of each $\hat{\beta}_k$. For normally distributed $X_i$, Li (1991) derives the asymptotic distribution for the summation of the smallest $J-k$ eigenvalues of $\hat{M}_{\text{SIR}}$:

$$(5) \quad n \sum_{j=k+1}^{J} \hat{\lambda}_j \sim \chi^2((J-k)(S-k-1)).$$

This asymptotic distribution enables estimation of the number of significant e.d.r. directions by testing $H_0: d = k$ v.s. $H_1: d = k+1$. Specifically, if $n \sum_{j=k+1}^{J} \hat{\lambda}_j$ exceeds the corresponding $\chi^2((J-k)(S-k-1))$ critical value, then there are at least $k+1$ significant e.d.r. directions.

Applied users can conduct these tests sequentially for $k = 0, 1, \cdots$, until the null hypothesis is not rejected. For non-normal $X$, Bura and Cook (2001a) develop a weighted chi-square test.

SIR is easy to implement if the dependent variable $Y_i$ is univariate. When $Y_i$ is of high dimension, the slicing step of SIR becomes impractical. To resolve this problem, Hsing (1999) defines the conditioning set by nearest neighbors. Let $i \in \{1, 2, \cdots, n\}$ and $i^* \in \{1, 2, \cdots, n\} - \{i\}$ be the indices for which

$$(6) \quad \gamma(Y_i, Y_{i^*}) = \min_{j \neq i, 1 \leq j \leq n} \gamma(Y_i, Y_j),$$

where $\gamma(\cdot, \cdot)$ is a metric such as Euclidean distance, i.e., $Y_{i^*}$ is the nearest neighbor of $Y_i$. Let $X_{i^*}$ be the concomitant of $Y_{i^*}$. Hsing (1999) estimates

$$(7) \quad \text{Cov}[E(Z_i | Y_i)] = 0.5E[Z_i' E(Z_i | Y_i)] + 0.5E[E(Z_i' | Y_i)Z_i]$$

using

$$(8) \quad \hat{M}_{\text{NNIR}} = (2n)^{-1} \sum_{i=1}^{n} (\hat{Z}_i' \hat{Z}_i + \hat{Z}_{i^*}' \hat{Z}_{i^*}),$$

where $\hat{Z}_i$ and $\hat{Z}_{i^*}$ are standardized $X_i$ and $X_{i^*}$ respectively. Although $\hat{Z}_{i^*}$ provides a noisy estimate of $E(Z_i | Y_i)$, the space spanned by the first $d$ eigenvectors of $\hat{M}_{\text{NNIR}}$ is root-$n$ consistent for the e.d.r. space under general conditions.
The eigenvectors of $\hat{M}_{NNIR}$ provide estimates of the e.d.r. space but, to our knowledge, the asymptotic distribution of the $d$ smallest eigenvalues of $\hat{M}_{NNIR}$ has not been derived. Thus, hypothesis tests based on a statistic like (5) cannot be used to estimate the number of significant e.d.r. directions. However, cross-validation and permutation tests (Cook and Weisberg 1991, Cook and Yin 2001) can be useful for determining $d$ in practice.

We close this section with two remarks about inverse regression methods.

**Remark 1:** The eigenvectors $\eta = (\eta_1, \eta_2, \ldots, \eta_d)$ satisfy $\eta' \eta = I$, which implies that $Z_i \eta$ is an orthogonal projection of $Z_i$. This orthogonal projection transforms a large set of correlated $X_i$ into a small set of uncorrelated $X_i B$ because

$$\text{cov}(X_i, B) = E \left[ (X_i B - E(X_i) B)' (X_i B - E(X_i) B) \right] = \eta' E(Z_i' Z_i) \eta = I_d.$$  

The intuition is that the $X_i$ variables may be correlated with each other and share some common information. Therefore, we can condense the variable space using an orthogonal projection to purge the redundant information. After this transformation, the resulting inverse regression variates $X_i B$ are uncorrelated and span a vector space of smaller dimension $d$.

**Remark 2:** To prove that $E(Z_i | Y_i)$ is contained in the linear subspace spanned by $(\eta_1, \eta_2, \ldots, \eta_d)$, Li requires the linear design condition, i.e., $E(X_i b | X_i \beta_1, X_i \beta_2, \ldots, X_i \beta_d)$ is linear in $X_i \beta_1, X_i \beta_2, \ldots, X_i \beta_d$ for any vector $b \in \mathbb{R}^d$. A sufficient but not necessary assumption for this condition to hold is that the distribution of $X_i$ is elliptically symmetric (e.g., the normal distribution). However, Hall and Li (1993) prove under general conditions that deviations from the linear design condition converge to zero as the dimension of $X_i$ increases. Therefore, the linear design condition holds asymptotically for high dimensional data.

### 2.2 Comparison to Other Dimension Reduction Methods

#### 2.2.1 Principal Components Analysis

Rather than using inverse regression, it is common to apply principal component analysis (PCA) on $X$ and keep the first few principal components for modeling the relationship between $Y$ and $X$ (e.g., Stock and Watson 2002b). The main drawback of this approach is that PCA does not take into account the relationship between $Y$ and $X$ when estimating the principal components.
Thus, for any two different $Y$ variables, PCA reduces the data to the same linear combinations of $X$, even if the relationships between $Y$ and $X$ are different across the two $Y$ variables. Consequently, PCA can misspecify the fundamental relationship between $Y$ and $X$ by excluding relevant components and introducing irrelevant components for $Y$.

A simple example illustrates this point. We generate 200 observations using $Y_i = \sin(X_i \beta) + U_i$, where $U_i \sim iid N(0, 0.01^2)$, $\beta_i = [0.25, -0.35, -0.57, 0.70]^T$, $X_i \sim iid N(\mu, \Phi)$ with mean vector $\mu = [5, 5, 5, 5]$ and covariance matrix $\Phi = (\beta_i' \Sigma_{XX} \beta_i)^{-1} \Sigma_{XX}$ with

$$
\Sigma_{XX} = \begin{bmatrix}
1 & \rho & \rho & \rho \\
\rho & 1 & \rho & \rho \\
\rho & \rho & 1 & \rho \\
\rho & \rho & \rho & 1
\end{bmatrix}.
$$

In this example, the sine link function is nonlinear in the four $X$ variables, and the four $X$ variables can be reduced to one linear component denoted by $X_i \beta_i$. We choose the covariance matrix of $X_i$ to be $\Phi = (\beta_i' \Sigma_{XX} \beta_i)^{-1} \Sigma_{XX}$ so that $\text{var}(X_i \beta_i) = 1$.

We experiment with three correlation coefficients $\rho = 0, 0.5, \text{ and } 0.95$ and apply both SIR and PCA to estimate $\beta_i$. We use ten slices to obtain the SIR estimates. In Figure 1, we plot $Y$ against the first SIR variate in the left panel and plot $Y$ against the first principal component in the right panel. It is evident that SIR correctly identifies the nonlinear sine pattern while PCA finds no pattern for all three correlation coefficients. When $\rho$ increases, the $X$ variables become highly correlated and SIR’s ability to identify the e.d.r. direction decreases. Consequently, we observe that the data points are more scattered around the underlying sine curve for $\rho = 0.95$ than for $\rho = 0$.

### 2.2.2 OLS and Canonical Correlations

Under the assumption of a linear link function, inverse regression is identical to OLS and canonical correlation. In the following proposition, we present this new result, which generates intuition about the theory underlying inverse regression. For easy exposition, we standardize $X_i$ to $Z_i = \Sigma_{XX}^{-1/2} [X_i - E(X_i)]$. 


Proposition: Consider the classic linear regression model \( Y_i = Z_i \eta + u_i \) and the linear inverse regression curve \( E(Z_i \mid Y_i) = \Sigma_{YZ}^{-1} \Sigma_{YZ}^{-1} Y_i \), where \( \Sigma_{ZY} = \text{Cov}(Z_i, Y_i) \) and \( \Sigma_{YY} = \text{Cov}(Y_i) \).

(i) For a univariate \( Y_i \), the estimates of \( \eta \) by OLS and by inverse regression are identical in population.

(ii) For a multivariate \( Y_i \), the estimates of \( \eta \) by canonical correlation and by inverse regression are identical in population.

We only prove part (ii) because it includes part (i) as a special case. Consider the canonical correlation between \( Z_i, \eta \), a linear combination of \( Z_i \), and \( Y_i, \alpha \), a linear combination of \( Y_i \). For identification, we use the normalizations \( \eta' \Sigma_{ZZ} \eta = \eta' \eta = 1 \) and \( \alpha' \Sigma_{YY} \alpha = 1 \), which imply that the canonical correlation coefficients \( \eta \) equal the eigenvectors of \( \Sigma_{ZY} \Sigma_{YY}^{-1} \Sigma_{YZ} \) (Hamilton 1994, pp. 630-635). Inverse regression calculates the eigenvectors of \( \text{Cov}[E(Z_i \mid Y_i)] \), which equals \( \Sigma_{ZY} \Sigma_{YY}^{-1} \Sigma_{YZ} \) if \( E(Z_i \mid Y_i) = \Sigma_{ZY} \Sigma_{YY}^{-1} Y_i \). Because both methods perform an eigen decomposition of \( \Sigma_{ZY} \Sigma_{YY}^{-1} \Sigma_{YZ} \), canonical correlation and inverse regression are identical given a linear relationship between \( Y \) and \( Z \).

Moreover, when estimated by maximum likelihood, a reduced rank regression model \( Y_i = Z_i \xi^\prime + \epsilon_i \) with a normally distributed \( \epsilon_i \) has the first order condition \( (\Sigma_{ZY} \Sigma_{YY}^{-1} \Sigma_{YZ} - \lambda^2) \eta = 0 \) (see Camba-Mendez, Kapetanios, Smith, and Weale 2003). This condition is the same as the one for canonical correlation and therefore makes inverse regression identical to maximum likelihood for this case. Although all four methods utilize information from \( Y \), inverse regression generalizes linear regression, canonical correlation, and reduced rank regression because it imposes no structure on link function and error term. The inverse regression curve is estimated nonparametrically and the error term does not need to be additive.

3. Function Space Reduction Using Nonparametric Factors

In this section, we shift our attention from the variable space to the function space. First, we discuss how to estimate regression functions after dimension reduction in variable space. Second, we present function space reduction and visualization. The idea of function space reduction is similar to that of variable space reduction. That is, the information in a large number
of unknown regression functions may be well captured by a small number of orthogonal basis functions, known as nonparametric factors. Identifying nonparametric factors enables function-space visualization, which in turn provides information for parsimonious model specification.

### 3.1 Nonparametric Regression after Inverse Regression

In this section, we suggest using inverse regression variates for nonparametric regression with high dimensional $X_i$. Without variable space reduction, nonparametric estimation of $F_0(X_i)$ is subject to the curse of dimensionality. For example, consider the Nadaraya-Watson kernel estimator

$$
\hat{F}_0(X_i) = \frac{\sum_{j=1}^{n} K \left( h^{-1} (X_j - X_i) \right) Y_j}{\sum_{j=1}^{n} K \left( h^{-1} (X_j - X_i) \right)},
$$

where $K$ represents a kernel weighting function and $X_i$ is of dimension $J$. The optimal bandwidth $h_{opt} = O(n^{-1/(J+4)})$ results in the smallest mean squared error (MSE) of $O(n^{-4/(J+4)})$ under the regularity conditions (A2)-(A7) in Pagan and Ullah (1999, pp. 96-104). It follows that the MSE converges to zero slowly when $J$ is large, creating the curse of dimensionality.

To mitigate the curse of dimensionality, we regress $Y_i$ on the $d$ inverse regression variates $X_iB$, rather than $Y_i$ on the $J$-dimensional $X_i$. Because the reduced variates $X_iB$ capture all the information in $X_i$ that is relevant for $Y_i$, we have $F_0(X_i) = F_i(X_iB)$, which implies that we can estimate $F_0(X_i)$ using

$$
\hat{F}_i(X_iB) = \frac{\sum_{j=1}^{n} K \left( h^{-1} (X_j - X_i)B \right) Y_j}{\sum_{j=1}^{n} K \left( h^{-1} (X_j - X_i)B \right)}.
$$

The MSE of this kernel regression is $O(n^{-4/(d+4)})$, which is smaller than $O(n^{-4/(J+4)})$ if $d \ll J$.

In practice, we do not observe the true e.d.r. directions, so we estimate $B$ using inverse regression. Inverse regression methods produce root-$n$ consistent estimates, whose convergence rate is faster than those of nonparametric estimates in general. Therefore, replacing $B$ with $\hat{B}$ does not affect the asymptotic properties of nonparametric statistics in general.

### 3.2 Estimating Nonparametric Factors

The matrix $A$ and the nonparametric factors in (3) are not separately identifiable because

$$
F_i(X_iB) = AH(X_iB) = A\Pi^{-1} \Pi H(X_iB) = A \hat{H}'(X_iB)
$$

- 9 -
holds for any \( L \times L \) full rank matrix \( \Pi \). Although \( H(X, B) \) and \( H'(X, B) = \Pi H(X, B) \) are different functions, they span the same function space because \( H'(X, B) \) is a linear transformation of \( H(X, B) \). Consequently, we could identify the function space using \( H(X, B) \) or \( H'(X, B) \). This fact implies that we could consistently estimate the function space spanned by the nonparametric factors using

\[
(12) \quad \hat{H}(X, \hat{B}) = \Xi \hat{F}_1(X, \hat{B}),
\]

where \( \Xi \) denotes any \( L \times G \) matrix with full rank \( L \).

To identify the nonparametric factors, we impose the orthonormal restriction

\[
(13) \quad \Xi \left( n^{-1} \sum_{i=1}^{n} \hat{F}_i(X, \hat{B}) \hat{F}_i(X, \hat{B})' \right) \Xi' = I_L.
\]

This restriction generates orthogonal nonparametric factors, and it implies that we can estimate \( \Xi \) using an eigenvalue decomposition. Specifically, we set \( \Xi = \Lambda^{-1/2} C' \), where \( \Lambda \) is a diagonal matrix containing the \( L \) largest eigenvalues of \( n^{-1} \sum_{i=1}^{n} \hat{F}_i(X, \hat{B}) \hat{F}_i(X, \hat{B})' \) and \( C \) is a \( G \times L \) matrix containing the corresponding eigenvectors.

The factor loadings matrix, \( A \), measures the sensitivity of the dependent variables to the nonparametric factors. Because (13) implies \( n^{-1} \sum_{i=1}^{n} \hat{H}(X, \hat{B}) \hat{H}(X, \hat{B})' = I_L \), the resulting \( L \) nonparametric factors are orthogonal and have a unit norm. Consequently, the magnitude of \( A \) indicates the relative contribution of each nonparametric factor to the variation of \( F_i(X, B) \). We can calculate \( A = \Xi' (\Xi \Xi')^{-1} \) and obtain its confidence intervals using the nonparametric bootstrap (see Section 5 for an example).

### 3.3 Function Space Visualization

To visualize the function space, we plot each of the estimated nonparametric factors, \([\hat{H}_1(X, \hat{B}), \ldots, \hat{H}_L(X, \hat{B})]\), against each of the estimated inverse regression variates, \((X_i \hat{\beta}_1, \ldots, X_i \hat{\beta}_L)\). Such bivariate plots can effectively reveal features of the data that would be imperceptible from the bivariate plots of each \( Y \) against each \( X \) or the plots of estimated nonparametric regression functions \( \hat{F}_i(X, \hat{B}) \).
We illustrate function space visualization through an example. Suppose that the data generating process is \( Y_i = A_0 H_0(X_i \beta_i) + U_i \), where \( H_0(X_i \beta_i) = \begin{bmatrix} (X_i \beta_i) & (X_i \beta_i)^3 \end{bmatrix} \), \( X_i \sim \text{i.i.d. } N(0,I_4) \), \( \beta_i = \begin{bmatrix} 0.25 & -0.35 & -0.57 & 0.70 \end{bmatrix} \), \( U_i \sim \text{i.i.d. } N(0, \Sigma) \),

\[
A_0 = \begin{bmatrix}
2 & 1 \\
2 & 2 \\
4 & 1
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
2 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2
\end{bmatrix}.
\]

In this example, the four \( X \) variables can be reduced to a single inverse regression variate, \( X_i \beta_i \), and the three regression functions can be reduced to two basis functions, \( X_i \beta_i \) and \( (X_i \beta_i)^3 \). As discussed in Section 3.2, the basis functions are not identified by the data. However, we can identify the function space spanned by these basis functions, and we can visualize the function space using orthonormal nonparametric factors.

First, we generate a sample of size 400 and apply nearest neighbor inverse regression to estimate \( \hat{\beta}_i \). We obtain \( \hat{\beta}_i = [0.26, -0.32, -0.54, 0.73]' \). Second, using \( \hat{\beta}_i \), we estimate the three regression functions \( F_i(X_i, \hat{\beta}_i) \) using kernel regression. Third, we compute the orthonormal nonparametric factors using the formula \( \hat{H}(X_i, \hat{\beta}_i) = \Xi \hat{F}_i(X_i, \hat{\beta}_i) \), where we obtain \( \Xi \) from an eigenvalue decomposition of \( n^{-1} \sum_{i=1}^{n} \hat{F}_i(X_i, \hat{\beta}) \hat{F}_i(X_i, \hat{\beta})' \) as in Section 3.2. Finally, we estimate the factor loading matrix and obtain

\[
A = \Xi'(\Xi \Xi')^{-1} = \begin{bmatrix}
5.07 & -0.14 \\
8.60 & 0.77 \\
6.79 & -0.88
\end{bmatrix}.
\]

For comparison, we plot the estimated regression functions with the data in Figure 2 and the estimated nonparametric factors in Figure 3. The three regression functions are linear combinations of the two nonparametric factors. The first nonparametric factor is concave for negative values of \( X_i \hat{\beta}_i \) and convex for positive values of \( X_i \hat{\beta}_i \). The three regression functions look similar to the first nonparametric factor because large values in the first column of \( A \) (i.e., large factor loadings) cause this factor to dominate the relationship between \( Y \) and \( X \).

However, the three regression functions exhibit different amounts of curvature because of the second nonparametric factor, which is also concave for negative values of \( X_i \hat{\beta}_i \) and convex...
for positive values of $X_i\hat{\beta}_i$. The first regression function is only slightly affected by the second factor because the factor loading -0.14 is close to zero. For the second regression function, adding 0.77 times the second factor increases the curvature of the function. On the other hand, for the third function, subtracting 0.88 times the second factor reduces the curvature of the function.

The curvature pattern derived from nonparametric factors matches the data generating process because it shows more curvature in the second regression function than the third. This feature arises because for the second regression function the factor loading on the linear term is relatively larger than the factor loading on the cubic term. This evidence shows how we can use function space visualization to uncover the nonlinear relationship hidden in the data even though the underlying basis functions $H_0(X_i\hat{\beta}_i)$ are not identifiable. Function space visualization requires knowledge of the number of nonparametric factors. In the next section we propose a consistent sequential testing procedure for this purpose.

4. A Consistent Nonparametric Rank Test

In this section, we propose a consistent nonparametric rank test for the dimension of the reduced function space. The dimension corresponds to the rank of a covariance matrix, a problem that has been addressed by Gill and Lewbel (1992), Robin and Smith (1995, 2000), Cragg and Donald (1997), and Donald (1997), among others. However, none of these tests are practical if the variable space is of high dimension. For example, Donald (1997) includes only one $X$ variable in both his simulation and application. We extend Donald’s test by allowing for a large variable space and stationary data with limited dependence. In addition, we achieve a faster (root-$n$) convergence rate by using a different matrix that retains the same rank.

4.1 Test Statistic

We aim to estimate $L$, which is the rank of the matrix $A$ in $F_i(X_iB) = AH(X_iB)$. We estimate $L$ by testing the rank of the matrix

\[
\Gamma_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} F_i(X_jB) F_i(X_iB)'
\]

\[
= A \left[ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X_jB) H(X_iB)' \right] A'
\]
where \( K_y = h_1^{-d} K \left( h_1^{-1} (X_i - X_j)B \right) \) is a symmetric kernel weighting function and \( h_1 \) is a bandwidth. The matrix \( \Gamma_n \) is of rank \( L \) as long as \( \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X_jB)H(X_iB)' \) has full rank \( L \) with probability one. To estimate \( \Gamma_n \), we use the second-order \( U \)-statistic

\[
\Gamma = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} Y_j Y_i'.
\]

We use a constant bandwidth \( h_1 \), which is the key to achieving root-\( n \) consistency for the test statistic. Donald (1997) shows that, for \( iid \) data, \( \Gamma \xrightarrow{p} \Gamma_D = E \left[ p(X_iB)F_0(X_iB)F_0(X_iB)' \right] \) if \( h \to 0 \) as \( n \to \infty \), where \( p(X_iB) \) denotes the \( p.d.f. \) of \( X_iB \). Using a constant \( h_1 \), we show below in Corollary 2 that \( \Gamma - \Gamma_n = O_p \left( n^{-1/2} \right) \) for near epoch dependent (NED) explanatory variables (\( iid \) can be regarded as a special case of NED). We obtain a more efficient test statistic based on \( \Gamma_n \) rather than \( \Gamma_D \) because \( \Gamma - \Gamma_n \xrightarrow{p} 0 \) faster than \( \Gamma \xrightarrow{p} \Gamma_D \). Intuitively, we can think of \( \Gamma \) as a biased estimator for Donald’s \( \Gamma_D \), but the bias term does not affect the true rank. Therefore, there is no efficiency gain to eliminating the bias asymptotically.

The efficiency gain can be characterized in terms of local power. Using a constant bandwidth \( h_1 \), we can distinguish the null from local alternatives when the difference between them diminishes at the parametric rate \( n^{-1/2} \). This rate is faster than \( n^{-1/2} h_1^{-d/4} \) with \( h_1 \to 0 \), the rate achieved by Donald’s test.\(^2\) The faster convergence rate of our test statistic generates better power performance in finite samples by permitting a larger bandwidth.

Following Donald (1997), we standardize \( \Gamma \) by \( \Sigma = E(U_iU_i') \) to account for the covariance of the error terms. Because the number of nonzero eigenvalues of \( \Gamma_n \Sigma^{-1} \) equals the rank of \( \Gamma_n \), we construct the test statistic

\[
S(l) = n \sum_{s=1}^{G-l} \lambda_s (\Gamma \Sigma^{-1})
\]

using the sum of \( G-l \) smallest eigenvalues of \( \Gamma \Sigma^{-1} \), rescaled by the sample size \( n \), where \( l \in \{0,1,\cdots,G\} \).

\(^2\) In Donald’s Lemma 3 (b), \( \delta_n = n^{-1/2} h_1^{-d/2} \) is a typographical error.
In the next subsection, we impose restrictions on the data generating process and derive the asymptotic properties of $S(l)$ under both the null hypothesis $H_0: L=l$ and the alternative hypothesis $H_0: L>l$.

4.2 Assumptions and Asymptotic Distributions

Assumptions:

A1: $Y_i = F_i(X_iB) + U_i$ for $i = 1, \cdots, n$, where $B = (\beta_1, \beta_2, \cdots, \beta_d)$ with $d \leq J$ and $U_i$ is an iid random vector independent of $X_i$ and satisfying $E(U_i) = 0$.

A2: $\{X_i\}$ is a strictly stationary, $L_2$-NED sequence of size $-1$ on some strong mixing sequence of size $-2r/(r-2)$ for $r > 2$. Suppose $X_iB$ is continuously distributed and has common c.d.f. $P$ with a compact support on $\chi$, a measurable space for all $i$.

A3: $F_i(xB)$ are Borel measurable functions having finite first and second moments and satisfying the following conditions:

(i) $\{E| F_i(xB)|^p\}^{1/p} < \infty$ as $p \to \infty$ (i.e., $L_\infty$-boundedness);

(ii) $\sup_x |F_i(xB)| \leq C_1$, where $C_1$ is a finite constant (i.e., uniform boundedness);

(iii) $|F_i(x_1B) - F_i(x_2B)| \leq C_2 \sum_{s=1}^{J} |x_1\beta_s - x_2\beta_s|$, where $C_2$ is a finite constant (i.e., uniform Lipschitz condition).

A4: $F_i \in \mathfrak{F}(L)$, where $\mathfrak{F}(L)$ defines a set of functions $\{F : R^d \to R^G\}$ that satisfy the condition $F_i(X_iB) = AH(X_iB)$ with probability one, where $\text{rank}(A) = L \leq G$.

A5: $E(U_iU'_i) = \Sigma$, where $\Sigma$ is a finite and positive definite matrix with full rank $G$.

A6: $\sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij}H(X_iB)H(X_jB)'$ has full rank $L$ with probability one, where $K_{ij}$ is a symmetric kernel function with a constant bandwidth $h_1$.

A7: The kernel function $\{K : \chi \times \chi \to R\}$ satisfies

(i) $\sup_{i,j} E(K_{ij}^{2+\delta}) < \infty$ and $\iint [K(x_iB,x_2B)]^{2+\delta} dP(x_iB)dP(x_2B) < \infty$ for some $\delta > 0$;

(ii) $K_{ij} = \sum_{q=1}^{\infty} \omega_q \tilde{\xi}_q(X_iB)\tilde{\xi}_q(X_jB)$, where $\omega_q$ and $\tilde{\xi}_q(\cdot)$ denote the eigenvalues and corresponding eigenfunctions. The eigenvalues and eigenfunctions satisfy
(a) $\int K(x,B,x_2,B) \xi_q(x_2,B)dP(x_2,B) = \omega_q \xi_q(x_1,B)$ and $\int \xi_q(x_1,B)dP(x_1,B) = 0$;

(b) $\int \tilde{\xi}_q(xB)\xi_q(xB)dP(xB) = \delta_{qq'}$, where $\delta_{qq'}$ equals one if $q = q'$ and zero if $q \neq q'$;

(c) $\int K(x_1,B,x_2,B)dP(x_1,B)dP(x_2,B) = \sum_{q=1}^{\infty} \omega_q^2 < \infty$ and $\sum_{q=1}^{\infty} |\omega_q| < \infty$;

(d) $\sup_x |\xi_q(xB)| \leq \tilde{C}_1$ and $|\xi_q(x_1,B) - \xi_q(x_2,B)| \leq \tilde{C}_2 \sum_{x=1}^{d} |x_1 \beta_s - x_2 \beta_s|$, where both $\tilde{C}_1$ and $\tilde{C}_2$ are finite constants.

Assumption A1 specifies a conditional mean equation $F_i(x_i,B)$ that allows for dimension reduction in variable space and iid error terms. Assumption A2 imposes restrictions on the dependence of stochastic process $\{X_i\}$. The near epoch dependent (NED) process can be a function of the infinite history of a mixing process but depends almost entirely on the “near epoch” of the underlying mixing process.³ We assume the same sizes for the NED process and for the underlying mixing process as in Chen and White (1996, 1998). Assumption A3 imposes moment conditions on $F_i(xB)$ required by the law of large numbers and central limit theorem for NED processes. As functions of $L_2$-NED process of size-1, $F_i(X_i,B)$ are also $L_2$-NED processes of size -1 under the uniform Lipschitz condition (Theorem 17.12 in Davidson 2002). The uniform Lipschitz condition and $\sup_x |F_i(xB)| \leq C_1$ combine with Assumption A7(ii)(d) to guarantee that $F_{i,q}(X_i,B) = \xi_q(X_i,B)F_i(X_i,B)$ also satisfies the uniform Lipschitz condition (see Appendix B.I.).

Assumption A4 ensures that the function space is spanned by $L \leq G$ nonparametric factors. We impose in Assumption A5 a second moment condition on the error term, and we make Assumption A6 to prevent $\Gamma_n$ from degenerating. In addition, we use a symmetric kernel function so that the eigenvalues of $\Gamma_n \Sigma^{-1}$ are all real numbers. Finally, in Assumption A7, we impose boundedness conditions on the kernel function and validate the eigenfunction expansion needed in proving Theorem 1 below.

Next, we present the asymptotic distribution of $S(l)$ under the null hypothesis $H_0$: $L = l$, which we derive using the Fujikoshi (1977) asymptotic expansion and eigenfunction expansions in Hilbert space.

³ See Davidson (2002) for advantages of NED over the mixing concept in time series modeling.
Theorem 1: Given Assumptions A1-A7,

\[ S(l) = n^{\frac{G-L}{2}} \sum_{i=1}^{n} \lambda_i (\Gamma \Sigma^{-1}) \]

\[ = tr \left\{ (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \varepsilon_j \varepsilon_i' \right\} \]

\[ - \left\{ n^{-1/2} (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \varepsilon_j H(X_i, B) \left( n^{-1/2} (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X_i, B) \varepsilon_j' \right) \right\} + o_p(1) \]

\[ \xrightarrow{d} \sum_{q=1}^{\infty} \left[ \theta_q X_{(G-L)q}^2 - \omega_q (G-L) \right] \]

under \( H_0: L=l \), where \( tr(\cdot) \) denotes the trace operator, \( \varepsilon_j \) is a \((G-L) \times 1\) iid random vector with mean zero and covariance matrix \( I_{G-L} \), \( H(X_i, B) \) is normalized such that

\[ n^{-1} (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X_i, B) H(X_i, B)' = I_L, \quad \{X_{(G-L)q}^2, q=1,2,\cdots\} \]

denotes an infinite sequence of independent Chi-square random variables with \((G-L)\) degrees of freedom, and both \( \{\theta_q\} \) and \( \{\omega_q\} \) are sequences of nonzero constants.

Proof: See Appendix A

Corollary 2: Given Assumptions A1-A7, \( \bar{\Gamma} - \Gamma_n = O_p(n^{-1/2}) \)

Proof: See Appendix B

The asymptotic distribution of \( S(L) \) is an infinite weighted sum of Chi-square random variables subtracted by an unknown constant. Because of the constant bandwidth \( h \), this asymptotic distribution is not pivotal but involves an infinite number of nuisance parameters. Specifically, the sequence \( \{\omega_q\} \) derives from the eigenfunction expansion of the kernel function defined in Assumption A7(ii). The sequence \( \{\theta_q\} \) denotes the eigenvalues of the infinite-dimensional matrix \( \Omega^* = \Omega - \Omega \bar{H} \bar{H}' \Omega \), where \( \Omega = diag\{\omega_q, q=1,2,\cdots\} \) and \( \bar{H}' = [\bar{H}_1, \cdots, \bar{H}_q, \cdots] \) with the \( q^{th} \) column \( \bar{H}_q = E\left[ \xi_q(X, B)H(X, B) \right] \).

To estimate \( L \), we also derive the asymptotic properties of \( S(l) \) under the alternative hypothesis \( H_1: L > l \).
Theorem 3: Given the assumptions A1-A7, we have \( S(l) = n \sum_{s=1}^{G-l} \lambda_s (\hat{\Gamma} \Sigma^{-1}) \xrightarrow{p} \infty \) under \( H_1 : L > l \).

The proof for Theorem 3 is straightforward because the eigenvalues \((\lambda_{G-L+1}, \ldots, \lambda_{G-l})\) of \( \Gamma_n \Sigma^{-1} \) are all positive with probability one if \( L > l \), and consequently \( n \sum_{s=1}^{G-l} \lambda_s (\hat{\Gamma} \Sigma^{-1}) \xrightarrow{p} \infty \) as the sample size \( n \to \infty \). Our Theorem 3 resembles the divergence result of Theorem 1 in Donald (1997). Theorem 1 and Theorem 3 together suggest estimating \( L \) via sequential testing of the hypotheses \( H_0: L=0 \) vs \( H_1: L>0 \), \( H_0: L=1 \) vs \( H_1: L>1 \), \( H_0: L=2 \) vs \( H_1: L>2 \), and so on until the null is not rejected. As discussed by Donald, this sequential test is consistent because the power \( \Pr(\text{reject } H_0 : L = l | H_1 : L > l) \to 1 \) as sample size \( n \to \infty \). This consistency prevents us from underestimating the true number of nonparametric factors in large samples.

4.3 Implementation of the Test

To obtain a feasible test statistic, we replace \( B \) with the root-\( n \) consistent estimate \( \hat{B} \) generated by inverse regression and calculate

\[
\hat{\Gamma} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{K}_{ij} Y_j Y_i',
\]

where \( \hat{K}_{ij} = h_1^{-d} K(\hat{h}^{-1}_1 (X_i - X_j) \hat{B}) \) is a kernel weighting function satisfying the assumptions A6 and A7. In addition, we replace \( \Sigma \) by the consistent estimator

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{F}_i(X_i, \hat{B}))(Y_i' - \hat{F}_i(X_i, \hat{B})),
\]

where \( \hat{F}_i(X_i, \hat{B}) = \sum_{j=1}^{n} \hat{W}_{ij} Y_j / \sum_{j=1}^{n} \hat{W}_{ij} \) and \( \hat{W}_{ij} = h_2^{-d} W(h_2^{-1} (X_i - X_j) \hat{B}) \) denotes another kernel weighting function. We use two bandwidths to calculate the statistic. For \( \hat{\Gamma} \), we use the kernel \( \hat{K}_{ij} \), which requires a constant bandwidth \( h_1 \), so that \( \hat{\Gamma} - \Gamma_n = O_p(n^{-1/2}) \). However, for \( \hat{\Sigma} \), we use the kernel \( \hat{W}_{ij} \), which requires a bandwidth \( h_2 \to 0 \) as \( n \to \infty \), so that \( \hat{\Sigma} \) is consistent for \( \Sigma \).

We combine the consistent estimates, \( \hat{\Gamma} \) and \( \hat{\Sigma} \), to form a feasible test statistic

\[
\hat{S}(l) = n \sum_{s=1}^{G-l} \lambda_s (\hat{\Gamma} \hat{\Sigma}^{-1}).
\]
Given the fact that eigenvalues are continuous functions of the elements of a matrix, \( \hat{S}(l) \) and \( S(l) \) are asymptotically equivalent by the Slutsky theorem. Consequently, by Lemma 4.7 in White (2001), \( \hat{S}(l) \) and \( S(l) \) have the same asymptotic properties as in Theorem 1 and Theorem 3.

Theorem 1 shows that \( S(l) \) has a well-defined limiting distribution under the null hypothesis \( H_0: L=l \). However, this limiting distribution is of a complicated form involving an infinite number of unknown quantities, i.e., \( \{\omega_q\} \) and \( \{\theta_q\} \) for \( q=1,2,\cdots \). This problem also occurs in Chen and Fan (1999), who propose a simulation-based resampling method. Here, we use a similar approach. Given \( B \) and \( H(X,B) \), we can simulate the asymptotic distribution of \( S(l) \) using the second expression in Theorem 1.

The simulation proceeds as follows. First, we draw a random sample of the \( n \times (G-l) \) matrix \( \tilde{\epsilon} \) from the standard normal distribution. Second, we obtain an \( n \times J \) matrix \( \tilde{X} \) and an \( n \times G \) matrix \( \tilde{Y} \) by sampling with replacement from the observed pair of \( (X,Y) \). Third, using \( \tilde{\epsilon} \), \( \tilde{X} \), and \( \tilde{Y} \), we calculate the statistic

\[
S(l) = tr \left\{ (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \hat{K}_{ij} \hat{\varepsilon}_i \hat{\varepsilon}_j^t \right. \\
- \left. \left( n^{-1/2} (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \hat{K}_{ij} \hat{\varepsilon}_j^t \hat{H}(\tilde{X}_i) \hat{H}(X,B) \right) \left( n^{-1/2} (n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \hat{K}_{ij} \hat{H}(\tilde{X}_i) \hat{\varepsilon}_j^t \right) \right\}
\]

where \( \hat{K}_{ij} = h^{-1}_i K(h^{-1}_i (\tilde{X}_i - \tilde{X}_j)B) \). Repeating these steps a large number of times generates a set of samples from the asymptotic distribution of \( S(l) \). Because \( \varepsilon_j \) is iid, the summands in the second expression in Theorem 1 are iid, so the serial correlation in \( X \) does not affect the asymptotic distribution. Thus, we do not need to preserve the serial correlation in \( X \) when resampling.

In presenting the asymptotic null distribution in Theorem 1, we define \( H(X,B) \) using the normalization condition \( n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X,B) H(X,B)' = I_L \). Thus, we normalize the estimated nonparametric factors in equation (20) using the condition \( n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \hat{K}_{ij} \hat{H}(X,B) \hat{H}(X,B)' = I_L \). This condition differs from the orthonormal condition \( n^{-1} \sum_{i=1}^{n} \hat{H}(X,B) \hat{H}(X,B)' = I_L \) for function space visualization that we use in Section 3. The data identify only the function space spanned by \( H(X,B) \), so these two normalizations are
equally consistent with the data. In general, applied users should choose a convenient normalization based on the context.

4.4 Monte Carlo Simulations

In this section, we examine the small sample properties of the test. We generate the data \((Y_i, X_i)\) using \(Y_i = \omega A H(X_iB) + U_i\), where \(\omega\) denotes the signal-to-noise ratio, \(A\) is a \(G \times 1\) constant matrix, \(H(X,B) = (X_iB) \cdot \exp(-X_iB/2)\), \(X_i \sim iid \ Unif(-\sqrt{3}, \sqrt{3})\), \(B\) is a \(J \times 1\) constant vector with unit length, and \(U_i \sim iid \ N(0, I_G)\). We begin with a base setting defined by \(\omega=1, G=5, J=5, n=200,\) and \(h_1=1\). We calculate the test statistic using the first NNIR variate and the Epanechnikov kernel. To illustrate the features of the test, we investigate the following four variations from the base setting:

(i) Varying bandwidth \((h_1)\)
(ii) Varying dimension \((G, J)\)
(iii) Varying signal/noise \((\omega)\)
(iv) Varying sample size and signal/noise \((n, \omega)\).

Table 1 reports the size (probability of rejecting \(H_0: L=1\)) and power (probability of rejecting \(H_0: L=0\)) for these settings. To simulate the rejection probabilities, we generate 1000 Monte Carlo samples of size \(n\). For each Monte Carlo sample, we compute the test statistic and compare it to the 95% critical value obtained using the resampling strategy proposed in Section 4.3 with 1000 replications. The simulated rejection probability equals the proportion of Monte Carlo samples in which the test statistic exceeds the critical value.

We report rejection probabilities for the cases where \(B, H(X,B)\), and \(\Sigma\) are known under the columns labeled “Infeasible”, and we report results using the estimated \(B, H(X,B)\), and \(\Sigma\) under the columns labeled “Feasible”. In the base setting, power drops from 0.979 to 0.834 when we replace the unknown quantities with their estimates. This power loss results from imprecision introduced by estimation of the unknown quantities. Estimating the unknown quantities has little effect on size in the baseline setting.

Decreasing the bandwidth \(h_1\) reduces the power of the test. When \(h_1=0.1\), the power in the feasible case equals 0.492, compared to 0.772 for \(h_1=0.5\) and 0.834 for \(h_1=1\). This result shows the benefit of using a larger bandwidth to calculate the statistic rather than a smaller bandwidth.
under the pretext that the bandwidth should tend to zero as the sample size grows. For this data generating process, we find no gain to increasing $h_1$ beyond 1.

Increasing the dimension of $(Y, X)$ while holding sample size constant raises the rejection probability of the test. For a sample of size $n=200$, the size of the feasible test increases from 0.044 for $G=J=5$ to 0.084 for $G=J=10$ to 0.152 for $G=J=20$. This size distortion reflects increasing imprecision from estimating $B$, $H(X; B)$, and $\Sigma$ as the dimension increases and the sample size is held constant.

Increasing the signal-to-noise ratio $\omega$ from 0.5 to 1 improves the power and drives the size toward the 5 percent nominal level. Setting $\omega$ to zero corresponds to $L=0$. For $\omega=0$, the simulated rejection probability for $H_0: L=0$ matches its nominal size (5 percent) for the infeasible case, but it is higher at 11.9 percent for the feasible case. This excess size again reflects the imprecision from estimating the unknown $B$, $H(X; B)$, and $\Sigma$. The simulated rejection probability for $H_0: L=1$ and $\omega=0$ is less than 5 percent. Similarly, a weak signal causes an undersized test. For example, for $\omega=0.5$ the data are closer to $L=0$ than for $\omega=1$, which reduces rejection rates for both $H_0: L=0$ (power) and $H_0: L=1$ (size).

Finally, we increase the sample size and decrease the signal-to-noise ratio simultaneously to demonstrate the local power property of the test. We observe similar rejection probabilities for the three scenarios, $(n=200, \omega=0.5)$, $(n=400, \omega=0.35)$, and $(n=800, \omega=0.25)$, which illustrates that the test possesses nontrivial power when the local alternatives converge to the null at the rate $O(n^{-1/2})$. Overall, these simulations corroborate the asymptotic theory and show that the test can perform well in finite samples.

5. **Application to Financial Asset Pricing**

Asset pricing theory purports that risk premia depend on the sensitivity of each asset to aggregate risk. Popular measures of aggregate risk include returns on the market portfolio as in the capital asset pricing model (CAPM) (Sharpe 1964, Lintner 1965), returns on portfolios based on size or the ratio of book equity to market equity (Fama and French 1992, 1993), and macroeconomic variables (Chen, Roll, and Ross 1986). In this section, we estimate the number of aggregate risk factors needed to explain excess stock returns, where the factors are nonparametric functions of a potentially large set of macroeconomic variables.
For asset returns, we use 25 portfolios of monthly stock returns constructed by Fama and French. These portfolios comprise the intersections of 5 portfolios formed on size and 5 portfolios formed on the ratio of book equity to market equity. We model excess returns, which we calculate by subtracting the one-month Treasury bill rate from each portfolio return. For the explanatory variables, we use a selection of 18 monthly macroeconomic variables measured over the period 1959:3–1998:12 (see Table 2 for details). These variables cover a wide range of real and nominal macroeconomic variables and also include an aggregate stock return index. We assume that each variable affects excess stock returns in the month when its value is released. So, our explanatory variables $X_i$ are composed by $(X_{1t}, X_{2t-1})$, where $X_{1t}$ denotes the information released in the current month and $X_{2t-1}$ the information released one month later. Following Stock and Watson (2002a), we remove trends when appropriate by taking logs and/or differencing.

5.1 Variable and Function Space Reduction

The econometric model is $R_t = F_t(X_t,B) + U_t$, where $R_t$ denotes the vector of excess stock returns, $F_t(X_t,B) = AH(X_t,B)$, and $U_t$ satisfies $E(U_t|X_t) = 0$. In this model, excess stock returns are linear in the nonparametric factors $H$, which in turn are unknown functions of the observed $X$ variables. This model differs from that in Chen, Roll, and Ross (1986) because it includes a large number of $X$ variables, and it allows for the factors to be nonlinear functions of $X$.

To reduce the variable space, we estimate the e.d.r. directions, $B$, using NNIR. We then estimate $E(R_t|X_t,B)$ using kernel regression as in (10) with $B$ replaced by $\hat{B}$ and jointly select the bandwidth, $h_2$, and the number of significant NNIR variates, $d$, by cross validation (see Pagan and Ullah, 1999, pp.119). Cross validation suggests that only the first NNIR variate is important in explaining the excess stock returns and the optimal bandwidth $h_2=0.54$.

In Figure 4, we plot in the left panel the estimated coefficients of the first NNIR variate, where each estimate is normalized by the standard deviation of the corresponding $X$ variable to enable comparison. The stock price index (variable 9 in Table 2) dominates the first NNIR variate, which matches the prediction of the CAPM. Other important variables are housing starts (variable 5) and the 10-year Treasury bond yield (variable 12). The other macroeconomic variables

---

4 http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
contribute little to excess stock returns. This result indicates that stock returns relate to an index that loads positively on the stock price index and housing starts and negatively on interest rates.

Rather than using inverse regression, applied researchers could use principal component analysis to reduce the variable space dimension. In the right panel of Figure 4, we plot the estimated coefficients of the first principal component of $X$. The principal component loads heavily on variables 1-8, 12, and 18 but has a small loading on the most influential variable – the stock price index. Thus, it produces a large estimated prediction error (EPE) of 715.61. Even if we include the first three principal components, the EPE only reduces to 608.76, which is still larger than 190.72, the EPE associated with a single NNIR variate.

Using the first NNIR variate, we perform the rank test and find evidence of two significant nonparametric factors. We report the $p$-values of the sequential testing procedure in Table 3. We show results for $d=1$ and 2 and for the values of $h_2$ chosen by cross validation, although the results are not sensitive to these choices. Using $h_1=1.0$ and 1.5, the $p$-values indicate the presence of two nonparametric factors. The $p$-values increase when we reduce $h_1$ from 1.0 to 0.5 and the $p$-value of $H_0: L=1$ exceeds 0.05 for $d=2$ and $h_1=0.5$. This result is consistent with our simulations in Section 4.4, which show that choosing a small bandwidth $h_1$ reduces the power of the test in finite samples. Moreover, this result illustrates the merit of not requiring $h_1 \to 0$ asymptotically.

The existence of the second nonparametric factor implies that any one-factor model is inadequate for excess stock returns. Including the second factor helps reconcile the following two puzzles in financial asset pricing (see Table 4):

(i) Value stocks have higher average excess returns but lower betas in the CAPM (known as the value premium puzzle);

(ii) In the value 1 category (growth stocks), the large size stocks have higher average excess returns but lower betas in the CAPM (known as the size premium puzzle).

In addition to the CAPM market portfolio, Fama and French (1992, 1993) proposed two linear risk factors to accommodate these puzzles: a size factor and a book-to-market equity ratio factor. However, our result reveals that a single nonlinear risk factor is sufficient to account for both the size- and value-premia. Next, we use function space visualization to show that the second nonparametric factor captures downside risk, which modifies the excess stock returns differently depending on the good or bad states of the economy (Ang, Chen and Xing, 2006).
5.2 Function Space Visualization: Downside Risk

To demonstrate the function space reduction, we plot in Figure 5 the estimated regression functions for the 25 portfolios in the left panel and the two estimated nonparametric factors in the right panel. Although all the regression functions appear linear in the NNIR variate, the rank test reveals that they are linear combinations of two nonparametric factors. The first nonparametric factor (solid curve) captures the essence of CAPM because it is approximately linear in the NNIR variate, which is dominated by the stock price index. The second nonparametric factor (dashed curve) appears to be a nonlinear function of the NNIR variate.

To understand the contribution of the second nonparametric factor, we present in Table 5 the estimated factor loadings, which measure the sensitivity of a portfolio to the aggregate risk factors. We estimate the factor loadings as in Section 3.2 using $A = \Xi'\Xi^{-1}$ and obtain the 95% confidence intervals using 1000 nonparametric bootstrap samples.\(^5\) The large factor loadings on the first nonparametric factor show that it dominates the relationship between the excess stock returns and the NNIR variate. The second nonparametric factor modifies this relationship.

The shape of the second nonparametric factor resembles the net payoff function to the writer of a call option. It is relatively flat for negative values of the NNIR variate and negatively sloped for positive values of the NNIR variate. This option would be written at the money on a portfolio with returns that mimic the NNIR variate. Writing such an option provides insurance against negative values of the NNIR variate, which correspond to bad states of the economy (e.g., low stock market returns, low housing starts, and high interest rates). For this reason, we interpret the second nonparametric factor as a downside risk factor. The existence of this factor is consistent with loss aversion preferences (Kahneman and Tversky, 1979) and the finding of Ang, Chen and Xing (2006) that a downside risk factor is priced in a cross section of stock returns.

The downside risk factor loadings explain the value- and size-premium puzzles. As shown in panel (ii) of Table 5, value stocks tend to have positive factor loadings and growth stocks tend to have negative factor loadings. Positive factor loadings imply that the second factor modifies returns by increasing the downside risk premium. Conversely, negative factor loadings reduce the downside risk premium. Thus, relative to growth stocks, value stock returns include a

\(^5\) We split the sample of size 477 into 40 non-overlapping blocks. The first 39 blocks are of length 12 and the last block contains 9 observations. We sample with replacement from the 40 blocks to generate 40 new blocks, which have different temporal ordering from the original 40 blocks. We then estimate $F_t(X,B)$, $\Xi$, and $A$ using this bootstrap sample.
larger insurance premium against downside risk. Similarly, in the value 1 category (growth stocks), large size stocks have less negative factor loadings than small size stocks, so returns on these large stocks include relatively more of an insurance premium against downside risk. In sum, value stocks and large growth-stocks are relatively more sensitive to downside market movements, and therefore command higher downside risk premia.

6. Conclusions

In this article, we present a new and convenient framework to analyze high dimensional econometric models. Rather than imposing a parametric model \textit{a priori}, we take a nonparametric approach and apply dimension reduction techniques to evade the \textit{curse of dimensionality}. We reduce dimension in both the variable space and the function space. For the variable space, we reduce a large number of explanatory variables to a few inverse regression variates. For the function space, we model a large number of regression functions using a small number of nonparametric factors. To estimate the dimension of function space, we develop a consistent nonparametric rank test that can distinguish between the null and the alternative when their difference is of order $O(n^{-1/2})$. We derive its non-pivotal asymptotic distribution under the null and propose a resampling strategy for simulating the asymptotic distribution.

Our dimension reduction method and rank test provide a practical toolkit for nonparametric estimation, data visualization, and parsimonious model specification, thereby enabling a more detailed understanding of complex economic processes. In the application section, we use these new econometric tools to identify a priced downside risk factor in stock returns. We show that this factor explains the value- and size-premium puzzles.
References


Appendix A

Proof of Theorem 1:

We derive the asymptotic distribution of \( n \sum_{j=1}^{G-L} \lambda_j \left( \Gamma \Sigma^{-1} \right) \). These eigenvalues are invariant to the transformation of \( \Psi' \Gamma \Psi \) and \( \Psi' \Sigma \Psi \), where \( \Psi \) is a nonsingular \( G \times G \) matrix chosen so that \( \Psi' \Sigma \Psi = I_G \) and \( \Psi' A_1 \Psi \) is diagonal and contains the eigenvalues of \( A_1 \Sigma^{-1} \) (Donald 1997).

Partition \( \Psi = (\Psi_1 \; \Psi_2) \) such that \( \Psi_2' A_1 \Psi_2 = 0 \), where \( \Psi_1 \) is \( G \times L \) and \( \Psi_2 \) is \( G \times (G - L) \). Using Lemma 1 of Fujikoshi (1977), the smallest \( G - L \) eigenvalues of \( \Psi' \Gamma \Psi \) are equal to the eigenvalues of

\[
0 \cdot I_{G-L} + \Psi_2' A_2 \Psi_2 + \Psi_2' A_3 \Psi_2 - \Psi_2' A_2 \Psi_1 \left( \Psi_1' A_1 \Psi_1 \right)^{-1} \Psi_1' A_2 \Psi_2 + o_p(1).
\]

Note that \( \Psi_2' A_1 \Psi_2 = 0 \) implies \( \Psi_2' A_2 \Psi_2 = 0 \) and

\[
\Psi_2' A_2 \Psi_1 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \left( \Psi_2' F_i(X, B) U_i' \Psi_1 + \Psi_2' U_j F_i(X, B)' \Psi_1 \right)
\]

\[
= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_j H(X, B)' A' \Psi_1
\]

and

\[
\Psi_1' A_1 \Psi_1 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_1' F_i(X, B) F_i(X, B)' \Psi_1
\]

\[
= \Psi_1' A \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X, B) H(X, B)' \right) A' \Psi_1
\]

\[
= \Psi_1' A A' \Psi_1
\]

using the normalization \( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X, B) H(X, B)' = I_L \). Thus,

\[
\Psi_2' A_2 \Psi_1 \left( \Psi_1' A_1 \Psi_1 \right)^{-1} \Psi_1' A_2 \Psi_2
\]

\[
= \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_j H(X, B)' \right) \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} H(X, B) U_j' \Psi_2 \right).
\]

So, the sum of the smallest \( G - L \) eigenvalues of \( \Gamma \Sigma^{-1} \) is
\[
\sum_{s=1}^{G=L} \lambda_s (\Gamma \Sigma^{-1}) = \sum_{s=1}^{G=L} \lambda_s \left( \Psi^T \Psi (\Psi^T \Psi)^{-1} \right) = \sum_{s=1}^{G=L} \lambda_s (\Psi^T \Psi)
\]

\[
= \text{tr} \left( \Psi'^2 A_2 \Psi'^2 - \Psi'^2 A_2 \Psi' \left( \Psi'^2 A_1 \Psi' \right)^{-1} \Psi'^2 A_2 \Psi' \right) + o_p(1)
\]

\[
= \text{tr} \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} \Psi'^2 U_i U'^2 \Psi' \right)
\]

\[
- \text{tr} \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} \Psi'^2 U_i H(X_i, B)' \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} H(X_i, B) U_i U'^2 \Psi' \right) \right) + o_p(1).
\]

We study the two terms in this expression separately. The first term is a second-order U-statistic because

\[
\text{tr} \left( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} \Psi'^2 U_i U'^2 \Psi' \right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} \text{tr} \left( \Psi'^2 U_i U'^2 \Psi' \right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} U_i U'^2 \Psi' U_j.
\]

We perform an eigenfunction expansion of \( K_{ij} \) as in Assumption A7(ii), i.e.,

\[
K_{ij} = \sum_{q=1}^{\infty} \omega_q \xi_q(X, B) \xi_q(X, B),
\]

where \( \omega_q \) denotes the eigenvalues and \( \xi_q(\cdot) \) denotes the eigenfunctions, to obtain

\[
\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} U_i U'^2 \Psi'^2 U_j \xi_q(X, B).
\]

Define the \((G-L)\)-vector \( \epsilon_i = \Psi'^2 U_i \). We then have

\[
\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_q \xi_q(X, B) U_i U'^2 \Psi'^2 U_j \xi_q(X, B)
\]

\[
= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_q \xi_q(X, B) \epsilon_i \epsilon_j \xi_q(X, B)
\]

\[
= \frac{1}{(n-1)} \sum_{q=1}^{\infty} \omega_q \left[ \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_q(X, B) \epsilon_i \right) \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_q(X, B) \epsilon_i \right) - \frac{1}{n} \sum_{i=1}^{n} \xi_q(X, B)^2 \epsilon_i \epsilon_i \right]
\]

\[
= \frac{1}{(n-1)} \sum_{q=1}^{\infty} \omega_q \left[ \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{qi} \right) \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{qi} \right) - \frac{1}{n} \sum_{i=1}^{n} W_{qi} W_{qi} \right]
\]

where \( W_{qi} = \xi_q(X, B) \epsilon_i \). Using Assumptions A1, A2, and A7, we apply the CLT of Chen and White (1998, Proposition 5.2) to obtain

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{qi} = Z_q + o_p(1)
\]
where $Z_q \sim iid N(0, I_{G-L})$ across $q$ because $\varepsilon_i \sim iid D(0, I_{G-L})$ is independent of $X_i$ by assumption A1, $E\left[\xi_q(X_i, B)^2\right]=1$ and $E\left[\xi_q(X_i, B)\xi_{q'}(X_j, B)\right]=0 \forall q \neq q'$ by assumption A7(ii)(b). By Assumptions A1, A2, A7, and the LLN of Chen and White (1996, Corollary 4.4), we have

$$\frac{1}{n} \sum_{i=1}^{n} W_{qi} W_{qi} \rightarrow^P \sum_{s=1}^{G-L} E\left[\xi_q(X_i, B)^2\xi_{ai}\right] = \sum_{s=1}^{G-L} E\left[\xi_q(X_i, B)^2\right] E(\xi_{ai}^2) = G - L.$$ 

Thus, we have

$$n \text{tr}\left(\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_i U_i'^t \Psi_2\right) \rightarrow^d \sum_{q=1}^{\infty} \omega_q \left(Z_q' Z_q - (G - L)\right).$$

Using the same eigenfunction expansion of $K_{ij}$, consider $\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_j H(X_i, B)'$ in the second term:

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_j H(X_i, B)' = \frac{1}{n(n-1)} \sum_{i=1}^{\infty} \sum_{j \neq i} \omega_q \xi_q(X_i, B)\varepsilon_j H(X_i, B)' \xi_q(X_i, B)$$

$$= \frac{1}{\sqrt{n}} \sum_{q=1}^{\infty} \omega_q \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_q(X_i, B)\varepsilon_i\right) \left(\frac{1}{n-1} \sum_{i=1}^{n} H(X_i, B)' \xi_q(X_i, B)\right)$$

$$- \frac{1}{n-1} \sum_{q=1}^{\infty} \omega_q \left(\frac{1}{n} \sum_{i=1}^{n} \xi_q(X_i, B)\varepsilon_i H(X_i, B)' \xi_q(X_i, B)\right)$$

Using the fact that $F_j(X, B) \equiv AH(X_i, B)$, Appendix B.I shows that we can apply the LLN of Chen and White (1996, Corollary 4.4) to obtain

$$\frac{1}{n-1} \sum_{i=1}^{n} H(X_i, B)\xi_q(X_i, B) \rightarrow^P \Pi_q,$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \xi_q(X_i, B)\varepsilon_i H(X_i, B)' \xi_q(X_i, B) \rightarrow^P 0,$$

which imply

$$\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \Psi_2' U_j H(X_i, B)' = \sum_{q=1}^{\infty} \omega_q Z_q' \Pi_q + o_P(1),$$

where $Z_q$ is as defined above.

Putting it all together, we have
\[ n \sum_{s=1}^{G-L} \lambda_s (\Gamma \Sigma^{-1}) = \sum_{q=1}^{\infty} \omega_q (Z_q' \Sigma_q - (G-L)) - tr \left( \sum_{q=1}^{\infty} \omega_q Z_q' \Sigma_q \cdot \sum_{q=1}^{\infty} \omega_q \Sigma_q Z_q' \right) + o_p(1) \]

\[ = tr(Z' \Omega Z) - (G-L) \sum_{q=1}^{\infty} \omega_q - tr(Z' \Omega \Sigma \cdot \Sigma Z) + o_p(1) \]

\[ = tr \left\{ Z' \left( \Omega - \Omega \Sigma \Omega \right) Z \right\} - (G-L) \sum_{q=1}^{\infty} \omega_q + o_p(1) \]

\[ -d \sum_{q=1}^{\infty} \left[ \theta_q \lambda_q (G-L) - \omega_q (G-L) \right], \]

where \( Z' = [Z_1, Z_2, \ldots] \) with \( Z_q \) being a \((G-L) \times 1\) vector, \( \Sigma_q' = [\Sigma_1, \Sigma_2, \ldots] \) with \( \Sigma_q \) being an \( L \times 1 \) vector, \( \Sigma \) is a diagonal matrix with elements \( \{ \omega_q : q = 1, 2, \ldots \} \), and \( \{ \theta_q : q = 1, 2, \ldots \} \) denotes the eigenvalues of the infinite dimensional matrix \( \Omega^* = A \Sigma A' \). The distribution result holds because \( tr(Z' \Omega' Z) = \sum_{s=1}^{G-L} z_s' \Omega z_s \), where \( z_s \) denotes the \( s^{th} \) column of \( Z \) and \( z_s \sim iid N(0, I_{G-L}) \). Because \( \Omega^* \) is symmetric, we have the spectral decomposition \( \Omega^* = C \Theta C' \), in which \( \Theta \) is a diagonal matrix containing the eigenvalues of \( \Omega^* \) and \( C'C = I_{G-L} \). This decomposition results in

\[ \sum_{s=1}^{G-L} z_s' \Omega' z_s = \sum_{s=1}^{G-L} z_s' C \Theta C' z_s = \sum_{s=1}^{G-L} z_s' \Theta z_s = \sum_{s=1}^{\infty} \sum_{q=1}^{G-L} \theta_q z_{sq}^2 \sum_{s=1}^{G-L} z_{sq}^2 = \sum_{s=1}^{G-L} \theta_q \lambda_{(G-L)q}, \]

where \( z_s^* = C' z_s \) with its \( q^{th} \) element \( z_{sq}^* \sim iid N(0, 1) \).

### Appendix B

We prove \( \Gamma - \Gamma_n = O_p(n^{-1/2}) \) by decomposing \( \Gamma = \Gamma_n + A_2 + A_3 \), where

\[
A_2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} \left( F_i(X_iB)U_j' + U_j F_i(X_iB)' \right), \]

and

\[
A_3 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} U_j U_i', \]

We obtain the following two results:

(a) \( A_2 = O_p(n^{-1/2}) \) (see Appendix B.I. for proof), and

(b) \( A_3 = O_p(n^{-1}) \) (see Appendix B.II. for proof).
Combining results (a) and (b), we have $\bar{\Gamma} = \Gamma_n + O_p(n^{-1/2}) + O_p(n^{-1})$, which implies that $\bar{\Gamma} - \Gamma_n = O_p(n^{-1/2})$.

Appendix B.I.

Applying the eigenfunction expansion of $K_{ij}$, we have

$$
\frac{1}{n^{1/2}(n-1)} \sum_{i,j=1}^{n} K_{ij} F_i(X_j B) U'_i = \frac{1}{n^{1/2}(n-1)} \sum_{i,j=1}^{n} \sum_{q=1}^{\infty} \omega_q \xi_q(x_j B) \xi_q(x_i B) F_i(X_j B) U'_i \\
= \sum_{q=1}^{\infty} \omega_q \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{j=1}^{n} F_i(x_j B) \xi_q(x_i B) U'_i = O_p(1)
$$

given the following three results:

(i) $\frac{1}{n-1} \sum_{j=1}^{n} F_{1q}(X_j B) \xrightarrow{p} E\left[ F_{1q}(X_j B) \right]$;

(ii) $\frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_q(x_i B) U'_i = O_p(1)$ because $U_i$ is iid and independent of $X_i$ under Assumption A1 (The proof is similar to that for $\sqrt{n} \sum_{i=1}^{n} W_{qi}$ in Appendix A);

(iii) $\sum_{q=1}^{\infty} \omega_q < \infty$, which is guaranteed by $\sum_{q=1}^{\infty} |\omega_q| < \infty$ under assumption A7(ii)(c).

The proof for result (i) is as follows. Using Chen and White’s (1996) WLLN for NED processes, we have

$$
\frac{1}{n^{1/2}} \sum_{i=1}^{n} F_{1q}(X_i B) \xrightarrow{p} E\left[ F_{1q}(X_i B) \right],
$$

if we can verify $F_{1q}(X_i B)$ is a NED process. $F_{1q}(X_i B)$ is the product of $F_1(X_i B)$ and $\xi_q(x_i B)$, both of which are functions of $L_2$-NED process of size -1 and therefore are $L_2$-NED processes of size -1 by Theorem 17.12 in Davidson (2002). The uniform Lipschitz condition on $F_{1q}(X_i B)$ required in Theorem 17.12 is satisfied under assumptions A3 and A7(ii)(d) because

$$
\left| F_{1q}(x_i B) - F_{1q}(x_2 B) \right| = \left| F_1(x_i B) \xi_q(x_i B) - F_1(x_2 B) \xi_q(x_2 B) \right| \\
= \left| F_1(x_i B) \xi_q(x_i B) - F_1(x_2 B) \xi_q(x_i B) + F_1(x_2 B) \xi_q(x_2 B) - F_1(x_2 B) \xi_q(x_2 B) \right| \\
\leq \left[ F_1(x_i B) - F_1(x_2 B) \right] \left| \xi_q(x_i B) - \xi_q(x_2 B) \right| + \left| F_1(x_2 B) \right| \left| \xi_q(x_i B) - \xi_q(x_2 B) \right| \\
\leq \left| F_1(x_i B) - F_1(x_2 B) \right| \left| \xi_q(x_i B) \right| + \left| F_1(x_2 B) \right| \left| \xi_q(x_i B) - \xi_q(x_2 B) \right|
$$
\[
\leq C_1 \sum_{j=1}^d |x_j \beta_x - x_2 \beta_s| \cdot \text{Sup}_{x} |\xi_q(xB)| + \text{Sup}_{x} |F_1(xB)| \cdot \tilde{C}_2 \sum_{j=1}^d |x_j \beta_x - x_2 \beta_s|
\]
\[
= (C_1 \tilde{C}_1 + C_1 \tilde{C}_2) \sum_{j=1}^d |x_j \beta_x - x_2 \beta_s|.
\]

Consequently, \( F_{1q}(X,B) \) is an \( L_1 \)-NED processes of size -1 by Theorem 17.9 in Davidson (2002).

**Appendix B.II.**

Applying the eigenfunction expansion of \( K_{ij} \), we have

\[
\frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij} U_i U'_j = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{q=1}^{n} \omega_q \xi_q(X_j B) \xi_q(X_i B) U_i U'_j
\]

\[
= \frac{1}{n-1} \sum_{q=1}^{n} \omega_q \left[ \sum_{i=1}^{n} \xi_q(X_i B) U_i \right] \left[ \sum_{i=1}^{n} \xi_q(X_i B) U'_j \right] - \sum_{i=1}^{n} \xi_q(X_i B)^2 U_i U'_j
\]

\[
= \sum_{q=1}^{n} \omega_q \left[ \frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_q(X_i B) U_i \right] \left[ \frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_q(X_i B) U'_j \right] - \frac{1}{n} \sum_{i=1}^{n} \xi_q(X_i B)^2 U_i U'_j + o_p(1)
\]

\[
= O_p(1)
\]

given the following three results:

(i) Under the assumptions A1, A5, and A7(ii)(b),

\[
\frac{1}{n} \sum_{i=1}^{n} \xi_q(X_i B)^2 U_i U'_j \overset{p}{\longrightarrow} E\left[ \xi_q(X_i B)^2 U_i U'_j \right] = \Sigma
\]

using Chen and White’s WLLN for NED processes;

(ii) Same as Appendix B.I.(ii);

(iii) Same as Appendix B.I.(iii).
Table 1: Simulations

<table>
<thead>
<tr>
<th></th>
<th>Pr{Reject $H_0; L=1$}</th>
<th>Pr{Reject $H_0; L=0$}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feasible</td>
<td>Infeasible</td>
</tr>
<tr>
<td><strong>Baseline:</strong> $G=J=5, n=200, h_1=1, \omega=1$</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td><strong>Varying Bandwidth</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_1=0.1$</td>
<td>0.029</td>
<td>0.027</td>
</tr>
<tr>
<td>$h_1=0.5$</td>
<td>0.047</td>
<td>0.034</td>
</tr>
<tr>
<td>$h_1=1$</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td><strong>Varying Dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G=J=5$</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td>$G=J=10$</td>
<td>0.084</td>
<td>0.090</td>
</tr>
<tr>
<td>$G=J=20$</td>
<td>0.152</td>
<td>0.038</td>
</tr>
<tr>
<td><strong>Varying Signal/Noise Ratio</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega=0$</td>
<td>0.011</td>
<td>0.000</td>
</tr>
<tr>
<td>$\omega=0.5$</td>
<td>0.030</td>
<td>0.014</td>
</tr>
<tr>
<td>$\omega=1$</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td><strong>Varying Sample Size &amp; Signal/Noise Ratio</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=200, \omega=0.5$</td>
<td>0.030</td>
<td>0.014</td>
</tr>
<tr>
<td>$n=400, \omega=0.35$</td>
<td>0.015</td>
<td>0.006</td>
</tr>
<tr>
<td>$n=800, \omega=0.25$</td>
<td>0.015</td>
<td>0.007</td>
</tr>
</tbody>
</table>

**Notes:**

(a) $A_{G=5} = \begin{bmatrix} 0.05 & 0.10 & 0.15 & 0.20 & 0.25 \end{bmatrix}$ for $G=5$; $A_{G=10} = \begin{bmatrix} A_{G=5} & A_{G=5} \end{bmatrix}$ for $G=10$; $A_{G=20} = \begin{bmatrix} A_{G=5} & A_{G=5} & A_{G=5} & A_{G=5} \end{bmatrix}$ for $G=20$.

(b) To consistently estimate $\Sigma$, we use the bandwidths $h_2 = 0.2$ for $n=200$; $h_2 = 0.15$ for $n=400$; $h_2 = 0.1$ for $n=800$. 
Table 2: Macroeconomic Variables Used to Form Factors

<table>
<thead>
<tr>
<th>Order</th>
<th>Variable</th>
<th>Variable Description</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IP</td>
<td>Industrial production: total index (1992=100, sa)</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>GMYXPQ</td>
<td>Personal income less transfer pmts (chained) (bil 92$, saar)</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>LPNAG</td>
<td>Employees on nonag. Payrolls: total (thous.,sa)</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>GMCQ</td>
<td>Personal consumption expend (chained)-total (bil 92$, saar)</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>HSFR</td>
<td>Housing starts: nonfarm (47-58); farm &amp; nonfarm (59-) (sa)</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>PMNV</td>
<td>NAPM inventories index (percent)</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>MO</td>
<td>Mfg new orders: all manufacturing industries, total</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>MU</td>
<td>Mfg unfilled orders: all mfg industries, total (mil$. sa)</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>SI</td>
<td>NYSE, AMEX, NASDAQ stock price index</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>EXRUS</td>
<td>United States effective exchange rate (merm)(index no.)</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>EXRCAN</td>
<td>Foreign exchange rate: Canada (canada $ per U.S.$)</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>FYGT10</td>
<td>Interest rate: US treasury const maturities, 10-yr.</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>SFYGT10</td>
<td>Spread fygt10-fyff (where fyff is Federal funds interest rate)</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>FM2</td>
<td>Money stock: m2</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>FM2DQ</td>
<td>Money supply - m2 in 1992 dollars (bci)</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>PUNEW</td>
<td>CPI-u: all items</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>LEHM</td>
<td>Average hr earnings of prod. workers: manufacturing</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>HHSNTN</td>
<td>U of Michigan index of consumer expectations (bcd-83)</td>
<td>1</td>
</tr>
</tbody>
</table>

**Notes:**
(a) All variables except SI (variable 9) are from Stock and Watson (2002a). SI is from Kenneth French’s data library.
(b) The data transformations are (1) no transformation, (2) first difference, (4) logarithm, (5) first difference of logarithm, and (6) second difference of logarithm.
(c) $x_{1t}$ includes variables 9, 10, 11, 12, 13, 14, 15, and 18; while $x_2_{t-1}$ includes the remaining variables.
Table 3: P-Values of Sequential Testing

<table>
<thead>
<tr>
<th>$d$</th>
<th>$h_2$ (CV)</th>
<th>$h_1$</th>
<th>$L=0$</th>
<th>$L=1$</th>
<th>$L=2$</th>
<th>$L=3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (CV)</td>
<td>0.54</td>
<td>0.5</td>
<td>0.000</td>
<td>0.015</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.77</td>
<td>0.5</td>
<td>0.000</td>
<td>0.276</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.000</td>
<td>0.005</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Note: (CV) indicates the bandwidth and the number of NNIR variates that minimize the cross-validation criterion.

Table 4: Average Excess Return on the 25 Portfolios

<table>
<thead>
<tr>
<th>Size</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size 1</td>
<td>0.189</td>
<td>0.678</td>
<td>0.684</td>
<td>0.922</td>
<td>1.042</td>
</tr>
<tr>
<td>Size 2</td>
<td>0.370</td>
<td>0.622</td>
<td>0.818</td>
<td>0.937</td>
<td>1.010</td>
</tr>
<tr>
<td>Size 3</td>
<td>0.453</td>
<td>0.701</td>
<td>0.670</td>
<td>0.852</td>
<td>0.891</td>
</tr>
<tr>
<td>Size 4</td>
<td>0.482</td>
<td>0.479</td>
<td>0.702</td>
<td>0.808</td>
<td>0.854</td>
</tr>
<tr>
<td>Size 5</td>
<td>0.525</td>
<td>0.495</td>
<td>0.581</td>
<td>0.603</td>
<td>0.727</td>
</tr>
</tbody>
</table>

Note: Sizes 1 - 5 represent small – large stocks; values 1-5 present growth – value stocks.
Table 5: Nonparametric Factor Loadings for the 25 Portfolios

(i) Nonparametric Factor 1

<table>
<thead>
<tr>
<th></th>
<th>value 1</th>
<th>value 2</th>
<th>value 3</th>
<th>value 4</th>
<th>value 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>size 1</td>
<td>5.584</td>
<td>5.066</td>
<td>4.620</td>
<td>4.365</td>
<td>4.497</td>
</tr>
<tr>
<td>size 2</td>
<td>5.725</td>
<td>4.986</td>
<td>4.523</td>
<td>4.328</td>
<td>4.623</td>
</tr>
<tr>
<td></td>
<td>(4.577, 6.427)</td>
<td>(4.033, 5.605)</td>
<td>(3.630, 5.100)</td>
<td>(3.462, 4.844)</td>
<td>(3.673, 5.221)</td>
</tr>
<tr>
<td>size 4</td>
<td>4.957</td>
<td>4.556</td>
<td>4.286</td>
<td>3.991</td>
<td>4.361</td>
</tr>
<tr>
<td>size 5</td>
<td>4.111</td>
<td>4.015</td>
<td>3.651</td>
<td>3.521</td>
<td>3.648</td>
</tr>
</tbody>
</table>

(ii) Nonparametric Factor 2

<table>
<thead>
<tr>
<th></th>
<th>value 1</th>
<th>value 2</th>
<th>value 3</th>
<th>value 4</th>
<th>value 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>size 1</td>
<td>-0.640</td>
<td>0.048</td>
<td>0.131</td>
<td>0.334</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>(-1.067, -0.032)</td>
<td>(-0.566, 0.416)</td>
<td>(-0.470, 0.429)</td>
<td>(-0.568, 0.597)</td>
<td>(-0.726, 0.845)</td>
</tr>
<tr>
<td>size 2</td>
<td>-0.607</td>
<td>-0.109</td>
<td>0.223</td>
<td>0.311</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>(-0.876, 0.243)</td>
<td>(-0.296, 0.092)</td>
<td>(-0.330, 0.418)</td>
<td>(-0.305, 0.491)</td>
<td>(-0.429, 0.743)</td>
</tr>
<tr>
<td>size 3</td>
<td>-0.517</td>
<td>0.040</td>
<td>0.104</td>
<td>0.267</td>
<td>0.355</td>
</tr>
<tr>
<td></td>
<td>(-0.810, 0.411)</td>
<td>(-0.149, 0.122)</td>
<td>(-0.155, 0.224)</td>
<td>(-0.081, 0.426)</td>
<td>(-0.364, 0.586)</td>
</tr>
<tr>
<td>size 4</td>
<td>-0.525</td>
<td>-0.2000</td>
<td>0.065</td>
<td>0.263</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>(-0.979, 0.577)</td>
<td>(-0.400, 0.407)</td>
<td>(-0.117, 0.273)</td>
<td>(-0.105, 0.525)</td>
<td>(-0.089, 0.662)</td>
</tr>
<tr>
<td>size 5</td>
<td>-0.483</td>
<td>-0.281</td>
<td>-0.092</td>
<td>0.117</td>
<td>0.384</td>
</tr>
<tr>
<td></td>
<td>(-1.160, 0.805)</td>
<td>(-0.620, 0.612)</td>
<td>(-0.422, 0.509)</td>
<td>(-0.306, 0.676)</td>
<td>(-0.111, 0.606)</td>
</tr>
</tbody>
</table>

Note: The 95% confidence intervals (in parentheses) are obtained using moving blocks nonparametric bootstrap.
Figure 1: Sliced Inverse Regression versus Principal Component Analysis
Figure 2: Estimated Nonparametric Regression Functions

Figure 3: Estimated Orthonormal Nonparametric Factors
Figure 4: Variable Space Reduction

Figure 5: Function Space Reduction