The Annals of Statistics
1994, Vol. 22, No. 3, 1386–1427

NONPARAMETRIC ESTIMATION OF COMMON REGRESSORS FOR SIMILAR CURVE DATA[1]

BY ALOIS KNEIP

Université Catholique de Louvain

The paper is concerned with data from a collection of different, but related, regression curves $(m_j)_{j=1,\dots,N}$, $N \gg 1$. In statistical practice, analysis of such data is most frequently based on low-dimensional linear models. It is then assumed that each regression curve m_j is a linear combination of a small number $L \ll N$ of common functions $g_1, \dots, g_L$. For example, if all m_j's are straight lines, this holds with $L = 2$, $g_1 \equiv 1$ and $g_2(x) = x$. In this paper the assumption of a prespecified model is dropped. A nonparametric method is presented which allows estimation of the smallest L and corresponding functions $g_1, \dots, g_L$ from the data. The procedure combines smoothing techniques with ideas related to principal component analysis. An asymptotic theory is presented which yields detailed insight into properties of the resulting estimators. An application to household expenditure data illustrates the approach.

1. Introduction. Most work in regression theory refers to a single sample (Y_i, X_i), $i = 1, \dots, n$. Very often, however, statisticians have to deal with more than one single regression in practice. Many studies are initiated to investigate typical features of some regression relationship under different experimental conditions, for different individuals and so on. Examples are psychophysiological studies of EEG curves for different individuals. Further examples are economic studies where household expenditures on certain commodities versus total expenditures are to be analyzed for different time periods with differing prices. Such studies contain observations Y_{ij}, $j = 1, \dots, N$, for N different, but related, subjects (individuals, experimental units, etc.). The data can frequently be represented within the following setup:

There are $n \cdot N$ data points (Y_{ij}, X_i), $i = 1, \dots, n$, $j = 1, \dots, N$. Observations Y_{ij} and design points $X_i \in J \subset \mathbb{R}^d$, $d \in \mathbb{N}$, are connected by a regression relationship:

$$Y_{ij} = m_j(X_i) + \epsilon_{ij}, \qquad i = 1, \dots, n,\ j = 1, \dots, N, \tag{1.1}$$

where ε_{ij} denotes an unknown zero-mean error term, and where the $m_j \colon J \to \mathbb{R}$ are unknown smooth regression functions.

In statistical practice, analysis of this type of regression data is most frequently based on parametric models. The simplest and best interpretable para-

Received December 1991; revised December 1993.

[1] This work has been performed as part of the research program of the Sonderforschungsbereich 303 at the University of Bonn. Financial support by Deutsche Forschungsgemeinschaft is gratefully acknowledged.

AMS 1991 *subject classification*. 62G07.

Keywords and phrases. Regression, curve estimation, linear models, model selection, principal components.

metric models are linear ones. It is then assumed that the regression functions m_j can be described by a linear combination of a number $L \leq N$ of functions $g_1, \ldots, g_L$. In other words,

$$m_j(X_i) = \sum_{r=1}^{L} \theta_{jr} \cdot g_r(X_i), \qquad i = 1, \ldots, n, j = 1, \ldots, N, \tag{1.2}$$

for unknown real parameters $\theta_{1r}, \ldots, \theta_{Nr}$ and smooth, real-valued functions g_r, $r = 1, \ldots, L$. In the following $g_1, g_2, \ldots$ will be called *basis functions.* The *smallest* L such that (1.2) holds for some appropriate basis functions will be called the *dimension* of the model and will be denoted by L_0.

Such linear models are usually postulated with prespecified $L \ll N$ and prespecified basis functions. Frequently a somewhat different notation is used, and the model is written in the equivalent form $\underline{Y}_j = G \cdot \underline{\theta}_j + \underline{\epsilon}_j$, where $G := (g_r(X_i))_{i,r}$ and where $\underline{Y}_j$, $\underline{\theta}_j$ and $\underline{\epsilon}_j$ are the vectors of observations, parameters and error terms of the jth individual. Prespecified linear models are widely used in regression analysis. However, very often the field of application does not provide sufficiently accurate prior knowledge, and model selection relies on trial and error. Then this approach encounters the problem of misspecification. They might lead to seriously deficient models which generate misleading results.

In this paper we drop the assumption that L and $g_1, \ldots, g_L$ are specified a priori. Let us consider the resulting situation:

1. Relation (1.2) is always fulfilled with $L = N, g_1 := m_1, \ldots, g_N := m_N$. It thus does not impose any restriction to ask for the components of *some* model of the form (1.2) which is able to describe the N regression functions.
2. A high-dimensional ($L_0 \approx N, L_0 \approx n$) model is of no use. On the other hand, the availability of a low-dimensional ($L_0 \ll N, n$) model (1.2) is highly desirable. It then provides a large reduction of dimensionality, and any further statistical analysis will benefit from the model. Furthermore, it will also be of interest in the field of application, since analyzing $g_1, \ldots, g_{L_0}$ will lead to insight into common structural properties of $m_1, \ldots, m_N$.
3. There are connections between the mechanisms generating the different regression curves. In many applications one will thus *expect* relations between the regression curves which might, at least approximately, be well described by a low-dimensional model (1.2). In many cases the *hypothesis* that a low-dimensional model (1.2) holds will thus be a reasonable starting point for statistical analysis.

These considerations motivate the approach presented in this paper. In the following it is only assumed that prior knowledge leads to the *hypothesis* that some model of the form (1.2) holds with $L_0 \ll N, n$. A method is then developed which uses the *data* themselves to *estimate* L_0 *and appropriate basis functions* $g_1, \ldots, g_{L_0}$.

The procedure combines smoothing techniques with ideas related to principal component analysis. Based on the hypothesis of a low-dimensional model, estimation of dimensionality relies on a goodness-of-fit criterion. Estimates of the values of basis functions at the design points are simultaneously obtained. The procedure for estimating L_0 is related to work by Lewbel (1991) and Li (1991), who in different contexts use similar ideas to estimate the rank of a matrix.

A detailed asymptotic theory is presented. It is shown that the method is a useful tool to "test" the appropriateness of a low-dimensional model and to identify its components. In particular, our procedure implies asymptotically consistent tests of specific hypothesis such as, say, $L_0 \leq 5$ or $L_0 \leq 10$. If $N, n \gg L_0$, the method will reveal the true dimension with high probability.

In Section 2 assumptions on the error terms are made precise, and normalizing conditions on basis functions and parameters are given. The most important requirement on the error term is $\text{var}(\epsilon_{ij}) = 1$, which is generalized in later sections. Section 3 contains a methodological discussion. An asymptotic theory for the resulting estimators is given in Section 4. In Section 5 the situation is considered that variances of the error terms are unknown and/or not equal to 1. The idea is then to transform the data, using estimated variances. In Section 6 a successful application of the method is presented. The data to be analyzed are family expenditures on various commodities (food, housing, fuel, etc.) in different years. The methodology introduced in this paper leads to the conclusion that the data can be appropriately modelled by a polynomial in $\log(X)$ of degree 4.

2. Basic settings.

Before starting a methodological discussion, some basic features of the model have to be made more precise.

2.1. *The error term.*

Inevitably, the desired procedure has to distinguish between deterministic components inherent in the data, that is, the basis functions, and between random fluctuations generated by noise. It is thus essential to clarify assumptions on the error terms.

ASSUMPTION 1.

(a) The ϵ_{ij}'s are independent random variables with $E\epsilon_{ij} = 0$.

(b) $\text{var}(\epsilon_{ij}) = 1$, $i = 1, \dots, n$, $j = 1, \dots, N$.

(c) There exist some $D_0, D_0^* < \infty$ such that $E\epsilon_{ij}^4 \leq D_0$ and $E\epsilon_{ij}^8 \leq D_0^*$ for all i, j.

Assumption 1(b) is, of course, very restrictive. In a few applications it will be satisfied automatically. Then one has to transform the data, using estimated variances. Details are deferred to Section 5.

2.2. *Normalization.*

Let $\underline{Y}_j := (Y_{1j}, \dots, Y_{nj})'$ and $\underline{m}_j := (m_j(X_1), \dots, m_j(X_n))'$. Similarly, set $\underline{g}_r := (g_r(X_1), \dots, g_r(X_n))'$ and $\underline{\theta}_j := (\theta_{j1}, \dots, \theta_{jL_0})'$. In the following $\underline{g}_1, \dots, \underline{g}_{L_0}$ will sometimes be called basis vectors.

When considering (1.2) more closely, it is easily seen that the components of the model are not uniquely determined. Let A denote an arbitrary regular $L_0 \times L_0$ matrix, and set $(\theta_{j1}^*, \dots, \theta_{jL_0}^*)' = \underline{\theta}_j^* := A^{-1} \cdot \underline{\theta}_j$ and $(g_1^*(X_i), \dots, g_{L_0}^*(X_i))$

$:= (g_1(X_i), \dots, g_{L_0}(X_i)) \cdot A$ for all i, j. We then obtain

$$\begin{aligned} m_j(X_i) &= \sum_{r=1}^{L_0} \theta_{jr} g_r(X_i) \\ &= \left(g_1(X_i), \dots, g_{L_0}(X_i)\right) \cdot A \cdot A^{-1} \cdot \underline{\theta}_j \\ &= \sum_{r=1}^{L_0} \theta^*_{jr} g^*_r(X_i). \end{aligned}$$

This unidentifiability can be eliminated by suitable normalizing conditions. One possible set of such normalizing conditions, imposing no restriction, are as follows:

(i) $(1/n) \sum_{i=1}^{n} g_r(X_i) g_s(X_i) = \delta_{rs}$;
(ii) $\sum_{j=1}^{N} \theta_{jr} \theta_{js} = 0$ if $r \neq s$;
(iii) $\sum_{j=1}^{N} \theta_{j1}^2 \geq \sum_{j=1}^{N} \theta_{j2}^2 \geq \cdots \geq \sum_{j=1}^{N} \theta_{jL_0}^2 > 0$.

Here, $\delta_{rs} = 1$ if $r = s$, and $\delta_{rs} = 0$ otherwise. Evidently, the normalization given by (i)–(iii) depends on n, N and on the particular design. As $n \to \infty$ it is, however, easily seen that under suitable conditions on the design (i) is asymptotically equivalent to selecting orthonormal functions (with respect to some L_2-norm).

It should be noted that (i) implies that $n^{-1/2}\underline{g}_1, \dots, n^{-1/2}\underline{g}_{L_0}$ are orthonormal vectors. Condition (iii) establishes some ordering of the basis functions. We use "g_1" to denote the function with the on-average largest influence on modelling the m_j's, "g_2" to denote the one with the on-average second-largest influence and so on. The above normalization is well suited for estimation.

Let us consider existence and uniqueness of a basis and of parameters satisfying (i)–(iii). First note that under these conditions model (1.2) implies

$$\begin{aligned} M := \frac{1}{N} \sum_{j=1}^{N} \underline{m}_j \, \underline{m}_j' &= \frac{1}{N} \sum_{j=1}^{N} \left(\sum_{r=1}^{L_0} \theta_{jr} \cdot \underline{g}_r \right) \left(\sum_{r=1}^{L_0} \theta_{jr} \cdot \underline{g}_r \right)' \\ &= \sum_{r=1}^{L_0} \frac{1}{N} \sum_{j=1}^{N} \theta_{jr}^2 \cdot \underline{g}_r \underline{g}_r'. \end{aligned} \tag{2.1}$$

It follows that $(1/N) \sum_{j=1}^{N} n\theta_{j1}^2, (1/N) \sum_{j=1}^{N} n\theta_{j2}^2, \dots$ are the largest, second-largest, ... eigenvalues of M, and that $n^{-1/2}\underline{g}_1$, $n^{-1/2}\underline{g}_2, \dots$ are the corresponding orthonormal eigenvectors.

The image space of M is span$\{\underline{m}_1, \dots, \underline{m}_N\}$, and (1.2) yields rank$(M) = L_0$ and span$\{\underline{m}_1, \dots, \underline{m}_N\}$ = span$\{\underline{g}_1, \dots, \underline{g}_{L_0}\}$. It is thus immediately seen that it is always possible to identify an appropriate basis by requiring that $n^{-1/2}\underline{g}_1$, $n^{-1/2}\underline{g}_2, \dots, n^{-1/2}\underline{g}_{L_0}$ are orthonormal eignvectors for the largest, second-largest, ..., L_0th-largest eigenvalues of M. By (2.1) this is identical to the normalization given by (i)–(iii). Furthermore, we can infer from well-known results of linear algebra that, if ">" instead of only "$\geq$" holds in (iii), $\underline{g}_1, \dots, \underline{g}_{L_0}$ and the parameters are uniquely determined up to sign changes. Clearly, one might replace $\underline{g}_r$ and θ_{jr} by $-\underline{g}_r$ and $-\theta_{jr}$ without invalidating (i)–(iii). This might be eliminated, too, by an appropriate additional normalizing condition. However, sign differences are of no importance in the present context. In the following, any comparison of$\underline{g}_r$ with a corresponding estimate will implicitly assume that $\underline{g}_r$ is equipped with an "appropriate" sign.

3. The estimation procedure. Let A be an arbitrary symmetric $k \times k$ matrix, $k \in \mathbb{N}$. In the sequel, we will use $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_k(A)$ to denote the eigenvalues of$A$, and $\underline{\gamma}_1(A), \underline{\gamma}_2(A), \dots, \underline{\gamma}_k(A)$ will denote corresponding orthonormal eigenvectors. We will require that $\gamma_{r1}(A) \geq 0$, where $\gamma_{r1}(A)$ denotes the first component of $\underline{\gamma}_r(A)$.

3.1. *A first approach.* Consider the matrix M defined by (2.1). An analysis of eigenvalues and eigenvectors of M leads to the identification of components of model (1.2). In particular, $\operatorname{rank}(M) = L_0$ and $\lambda_{L_0+1}(M) = \cdots = \lambda_n(M) = 0$. Unfortunately we do not know the true regression functions, and we cannot compute M.

A straightforward idea is to use the observations instead: Determine the $n \times n$ matrix $\widehat{M} := (1/N)\Sigma_{j=1}^N \underline{Y}_j \underline{Y}_j'$. Then analyze its eigenvalues $\lambda_1(\widehat{M}) \geq \lambda_2(\widehat{M}) \geq \cdots \geq \lambda_n(\widehat{M})$ and the corresponding scaled eigenvectors $n^{1/2}\underline{\gamma}_1(\widehat{M}), n^{1/2}\underline{\gamma}_2(\widehat{M}), \dots, n^{1/2}\underline{\gamma}_n(\widehat{M})$.

Note that $E\widehat{M} = M + I$, where I denotes the identity matrix. This motivates us to consider $\lambda_1(\widehat{M}), \lambda_2(\widehat{M}), \dots$ and $n^{1/2}\underline{\gamma}_1(\widehat{M}), n^{1/2}\underline{\gamma}_2(\widehat{M}), \dots$ as estimates of $(1/N)\Sigma_{j=1}^N n\theta_{j1}^2 + 1, (1/N)\Sigma_{j=1}^N n\theta_{j2}^2 + 1, \dots$ and $\underline{g}_1, \underline{g}_2, \dots$.

There is another, even more important motivation for using $n^{1/2}\underline{\gamma}_r(\widehat{M})$ as an estimate of$\underline{g}_r$. The question of determining functional values of appropriate basis functions at the design points is closely connected with the following problem. For given $L \leq N$ determine vectors $\underline{v}_1, \dots, \underline{v}_L \in \mathbb{R}^n$ in such a way that linear combinations of these vectors provide a "best" possible approximation to $\underline{Y}_1, \dots, \underline{Y}_N$. Selecting $n^{1/2}\underline{\gamma}_1(\widehat{M}), \dots, n^{1/2}\underline{\gamma}_L(\widehat{M})$ is optimal in a least squares sense:

$$
\begin{aligned}
\sum_{r=L+1}^{n} \lambda_r(\widehat{M}) &= \frac{1}{N} \sum_{j=1}^{n} \min_{\vartheta_{j1}, \dots, \vartheta_{jL} \in \mathbb{R}} \left\| \underline{Y}_j - \sum_{r=1}^{L} \vartheta_{jr} \cdot \underline{\gamma}_r(\widehat{M}) \right\|_2^2 \\
&= \min_{\underline{v}_1, \dots, \underline{v}_L \in \mathbb{R}^n} \frac{1}{N} \sum_{j=1}^{N} \min_{\vartheta_{j1}, \dots, \vartheta_{jL} \in \mathbb{R}} \left\| \underline{Y}_j - \sum_{r=1}^{L} \vartheta_{jr} \cdot \underline{v}_r \right\|_2^2
\end{aligned}
\tag{3.1}
$$

holds for all $L \leq N$. Relation (3.1) is rather well known; it was first established by Rao (1958).

Does this open a way to solve the problem stated in the Introduction? Since $\underline{\gamma}_r(\widehat{M})$ will "adapt" to noise, (3.1) leads us to expect that

$$\sum_{r=L_0+1}^{n} \lambda_r(\widehat{M}) \leq R_{L_0} := \frac{1}{N}\sum_{j=1}^{N} \min_{\vartheta_{j1},\dots,\vartheta_{jL_0}\in\mathbb{R}} \left\| \underline{Y}_j - \sum_{r=1}^{L_0} \vartheta_{jr}\cdot \underline{g}_r \right\|_2^2. \tag{3.2}$$

In fact, this can be rigorously proved (see Theorem 1). It follows from standard results of linear regression theory that $N \cdot R_{L_0}$ follows a χ^2 distribution with $N(n - L_0)$ degrees of freedom, if the error terms are normally distributed. This might be used to establish a procedure for estimating L_0.

Unfortunately, there are serious problems invalidating this approach to a large extent. The observations $\underline{Y}_j$ are very noisy estimates of $\underline{m}_j$. Consequently, $\underline{\gamma}_1(\widehat{M}), \underline{\gamma}_2(\widehat{M}), \dots$ will inherit a strong dependence on the error terms, unless N is extremely large. The scaled eigenvectors $n^{1/2}\underline{\gamma}_r(\widehat{M})$ will thus in general be very *bad* estimates of $\underline{g}_r$. Furthermore, we will have to expect that "$\ll$" instead of only "$\leq$" holds in (3.2).

REMARK 1. The principal idea of using eigenvectors of the $\widehat{M}$-matrix for determining components of a model of the form (1.2) is quite familiar to applied statisticians. Usually this is considered as an application of *principal component analysis* (PCA) with an additional interpretation of the $\underline{\gamma}_r(\widehat{M})$ in a modelling sense, as motivated by (3.1). References are, for example, Rao (1958), Berkey and Kent (1983), Möcks (1986) or Glaser and Ruchkin [(1976), Chapter 7]. It has to be emphasized that this is not exactly the point of view adopted in the present paper. Clearly, the main ideas leading to the above approach go back to PCA, but the basic issue of estimating (deterministic) functional components from noisy data has no interpretation in a PCA context.

3.2. *A refinement.* The above approach did not make any use of the smoothness of the m_j's. Since smoothing reduces noise, it provides a tool to refine this procedure in order to make things work.

We will consider linear smoothers. These are smoothing procedures with the property that estimates $\underline{\widehat{m}}_j$ of the $\underline{m}_j$ are obtained by multiplying a "smoother matrix" W_h with the vectors $\underline{Y}_j$ of observations, that is, $\underline{\widehat{m}}_j := W_h \cdot \underline{Y}_j$. Furthermore, we will confine ourselves to the case that W_h is a symmetric projection matrix with $\operatorname{rank}(W_h) = h$.

A detailed discussion of smoothing procedures and their respective smoother matrices can be found in Buja, Hastie and Tibshirani (1989). Quite generally, projection matrices W_h arise if smoothing is based on a least squares approximation by h prespecified functions $v_1, \dots, v_h$. Then, if the vectors $\underline{v}_1, \dots, \underline{v}_h$ of functional values at $X_1, \dots, X_n \in J \subset \mathbb{R}^d$ are independent, estimates $\underline{\widehat{m}}_j$ of $\underline{m}_j$ are obtained by

$$\underline{\widehat{m}}_j := W_h \cdot \underline{Y}_j = V_h\left(V_h'V_h\right)^{-1}V_h' \cdot \underline{Y}_j, \tag{3.3}$$

where V_h denotes the $n \times h$ matrix $(v_r(X_i))_{i,r}$. If $d = 1$, well-known and frequently applied examples of this approach are least squares approximations by polynomials of degree $h-1$, by harmonic polynomials of degree $h-1$ or by cubic B-splines based on a prespecified sequence of $h-2$ knots [cf. de Boor (1978)]. Any one of these methods is suitable for approximating smooth functions. If $L_0 \ll N, n$ and if the m_j and g_r are smooth, then

$$h > L_0,\ \underline{\tilde{m}}_j := W_h \cdot \underline{m}_j \approx \underline{m}_j,\ W_h \cdot \underline{g}_r \approx \underline{g}_r \tag{3.4}$$

will hold for some $h \ll N, n$. If, for example, the m_j's are four times continuously differentiable, it follows from well-known theorems of approximation theory that $\int_J (m_j(X) - \overline{m}_{h,j}(X))^2\, dX = O(h^{-8})$, where $\overline{m}_{h,j}$ denotes the best approximation of m_j by polynomials of degree h. For cubic B-splines, similar relations have been derived by de Boor (1978). For theoretical results characterizing a large class of projection-type smoothing methods, one might consult Cox (1988).

Under (3.4), equations (1.1) and (1.2) lead to

$$\underline{\hat{m}}_j = W_h \cdot \underline{m}_j + W_h \cdot \underline{\epsilon}_j =: \underline{\tilde{m}}_j + \underline{\tilde{\epsilon}}_j, \tag{3.5}$$

$$\underline{\tilde{m}}_j = \sum_{r=1}^{L_0} \theta_{jr}(W_h \cdot \underline{g}_r) = \sum_{r=1}^{L_0} \underline{\tilde{\theta}}_{jr} \underline{\tilde{g}}_r, \tag{3.6}$$

which holds for all j. Model (3.6) will be called the projected model. Here, we use "$\tilde{g}_r$" and "$\tilde{\theta}_{jr}$" to denote the normalized basis and parameters satisfying conditions (i)–(iii) of Section 2.2 for the projected model (usually the vectors $W_h \underline{g}_r$ will not be exactly orthogonal, and thus $\underline{\tilde{g}}_r \neq W_h \underline{g}_r$).

However, if (3.4) holds, all components of the projected model will be approximately equal to those of the true model. In particular, $\underline{\tilde{g}}_r \approx W_h \cdot \underline{g}_r \approx \underline{g}_r$. We thus might concentrate on estimating the components of the *projected* model. This means that we have to analyze the eigenvalues $(1/N)\sum_{j=1}^N n\tilde{\theta}_{jr}^2$ and eigenvectors $n^{-1/2}\underline{\tilde{g}}_r$ of the matrix $M_h := (1/N)\sum_{j=1}^N \underline{\tilde{m}}_j \underline{\tilde{m}}_j'$.

These considerations motivate the basic steps of our final approach:

1. Perform a data smoothing. For any $j = 1, \dots, N$ determine an estimate $\underline{\hat{m}}_j := W_h \cdot \underline{Y}_j$ of $\underline{m}_j$, where W_h denotes the smoother matrix associated with an appropriate smoothing procedure. W_h has to be a symmetric, $n \times n$ projection matrix with $\operatorname{rank}(W_h) = h < n$.
2. Compute the largest h eigenvalues $\hat{\lambda}_1 := \lambda_1(\widehat{M}_h) \geq \cdots \geq \hat{\lambda}_h := \lambda_h(\widehat{M}_h)$ and corresponding eigenvectors $\underline{\gamma}_1(\widehat{M}_h), \dots, \underline{\gamma}_h(\widehat{M}_h)$ of the matrix $\widehat{M}_h := (1/N) \sum_{j=1}^N \underline{\hat{m}}_j \underline{\hat{m}}_j'$.

REMARK 2. Following (3.3), let $\Gamma_h := V_h(V_h'V_h)^{-1/2}$. Clearly, $W_h = \Gamma_h\Gamma_h'$. It is now easily seen that $\lambda_r(\widehat{M}_h) = 0$, $r = h+1, \dots, n$, while the largest h eigenvalues of $\widehat{M}_h = W_h \widehat{M} W_h$ and $\widehat{\Lambda} := \Gamma_h' \widehat{M} \Gamma_h$ are equal. For any r, $\Gamma_h \cdot \underline{\gamma}_r(\widehat{\Lambda})$ is a normalized eigenvector of $\widehat{M}_h$ for $\lambda_r(\widehat{M}_h)$. It will thus be computationally simpler to determine eigenvalues and eigenvectors of $\widehat{\Lambda}$, which is only an $h \times h$ and not an $n \times n$ matrix.

Note that $E\widehat{\Lambda} = \Gamma_h' M_h \Gamma_h + I_h$, where I_h denotes the $h \times h$ identity matrix. We thus might consider $\lambda_1(\widehat{M}_h) = \lambda_1(\widehat{\Lambda})$, $\lambda_2(\widehat{M}_h) = \lambda_2(\widehat{\Lambda}), \ldots$ and $n^{1/2}\underline{\gamma}_1(\widehat{M}_h)$, $n^{1/2}\underline{\gamma}_2(\widehat{M}_h), \ldots$ as estimates of $(1/N)\Sigma_{j=1}^N n\widetilde{\theta}_{j1}^2 + 1, (1/N)\Sigma_{j=1}^N n\widetilde{\theta}_{j2}^2 + 1, \ldots$ and $\underline{\widetilde{g}}_1, \underline{\widetilde{g}}_2, \ldots$. If $h \ll n$, these estimates will no longer suffer from the drawbacks invalidating the original approach: passing over from $\underline{Y}_j$ to $\underline{\widehat{m}}_j$ is connected with a drastic reduction of noise. It holds that $n = E\|\underline{\epsilon}_j\|_2^2 \gg E\|\underline{\widetilde{\epsilon}}_j\|_2^2 = h$.

3.3. *Dimensionality.* Let us now consider the problem of dimensionality. For the projected model, relations similar to (3.1) and (3.2) can be shown to hold. However, there are some important differences: (a) The term $N \cdot \widetilde{R}_{L_0}$, defined by replacing $\underline{Y}_j$ by $\underline{\widehat{m}}_j$ and $\underline{g}_r$ by $\underline{\widetilde{g}}_r$ in (3.2), follows a χ^2 distribution with $N \cdot (h - L_0)$ degrees of freedom. As a consequence of the central limit theorem this is approximately true even for nonnormally distributed errors terms, if $h \ll n$. (b) If $h \ll N$, "$\leq$" in (3.2) might be replaced by "$\approx$."

Rigorous proofs are deferred to the next section. It should be noted that (a) and (b) rely heavily on the fact that W_h is a projection matrix. They fail to be true if this is not the case, as, for example, when using smoothing splines or kernel estimators for smoothing.

Together with the hypothesis of a low-dimensional model, the above considerations motivate the following approach to estimate the dimension of model (1.2):

3. Determine an estimate $\widehat{L}_0$ of L_0 by selecting the *smallest* L, $0 \leq L < h$, such that

$$N \cdot \sum_{r=L+1}^{h} \widehat{\lambda}_r < C_{\alpha, N \cdot (h-L)}, \tag{3.7}$$

where, for $\alpha > 0$, $C_{\alpha, N \cdot (h-L)}$ is the respective critical value of a χ^2 distribution with $N \cdot (h - L)$ degrees of freedom. If such an $L < h$ does not exist, set $\widehat{L}_0 := h$.

Here, the idea is to test whether, for $L = 0, 1, 2, 3, \ldots$, an L-dimensional model is in accordance with the data. The hypotheses of a low-dimensional model allows us to choose the smallest L which is not rejected. This is closely related to procedures proposed by Li (1991) and Lewbel (1991) in different contexts.

Of course, the above dimension estimates depend on the selection of the smoother matrix W_h. In many situations it will not be essential whether to choose, for example, polynomials or B-splines for smoothing. However, the degree of the polynomial or the number of knots will be important.

What happens if we select an inadequate h? First note that (3.5) and (3.6) hold for *any* possible W_h. However, if h is too large, we have to expect that $\Sigma_{r=L_0+1}^h \widehat{\lambda}_r \ll \widetilde{R}_{L_0}$, as outlined in Section 3.1. In this case our critical values will be too large. If, on the other hand, h is too small, we might obtain $(1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2 \ll (1/N)\sum_{j=1}^{N} n\theta_{jr}^2$ [note that necessarily $(1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2 \leq (1/N)\sum_{j=1}^{N} n\theta_{jr}^2$]. Even worse, if $h < L_0$, the matrix $M_h = (1/N)\sum_{j=1}^{N} \underline{m}_j \underline{m}_j'$ will possess less than L_0 nonzero eigenvalues. We see that any unfavorable choice of h has one single consequence: it increases the probability of selecting too few components.

These considerations show that in practice it will be most promising not just to rely on one h but to apply the method repeatedly, based on several different choices of h. For example, one might select four different smoothing parameters $h_1 < h_2 < h_3 < h_4$. Parameter h_1 should be small, while h_4 should be large enough to guarantee a negligible bias. Selection might be based on expectations about dimensionality and on assumptions about the degree of smoothness. Alternatively, an average smoothing parameter h^* obtained by cross-validation, Mallows' C_L or related methods may serve as a guideline, choosing $h_1 < h^*$, $h_2 \leq h^*$, $h_3 > h^*$ and $h_4 \gg h^*$. One then might do the following:

3a. Compute different estimates $\widehat{L}_0$ by using $h = h_1$, $h = h_2$, $h = h_3$ and $h = h_4$. Afterwards, one might do the following:

3b. Use the maximal $\widehat{L}_0$ as final estimate of L_0.

Clearly, this will increase our actual level of significance, but this effect will be tolerable if we do not rely on too many h's. Furthermore, to some extent it will be balanced by the fact that the probability of selecting too few components increases if h is too large or too small. If the subspaces spanned by $W_{h_1}, W_{h_2}, \ldots$ are nested, it is also possible to compute corrections.

REMARK 3. There are alternative ways for determining dimension estimates. For example, such estimates might be obtained by Mallows' C_L [Mallows (1973)], that is, by minimizing

$$\frac{1}{N}\sum_{j=1}^{N} \min_{\vartheta_{j1},\ldots,\vartheta_{jL} \in \mathbb{R}} \left\| \underline{Y}_j - \sum_{r=1}^{L} \vartheta_{jr} \cdot \underline{\gamma}_r(\widehat{M}_h) \right\|_2^2 + 2L$$

over L. Asymptotic properties of the resulting estimator $\widehat{L}_0^*$ can easily be derived from the results of Section 4. This method, or related ones, will be of particular interest if our primary goal is not model building, but only a "best" approximation of the regression functions in terms of (1.2).

3.4. *Basis functions.* Estimates of basis functions can be obtained from the eigenvectors of the $\widehat{M}_h$ matrix:

4. For $r = 1, \ldots, \widehat{L}_0$, estimate $\underline{g}_r$ by $\underline{\widehat{g}}_r := n^{1/2}\underline{\gamma}_r(\widehat{M}_h)$.

In order that $\underline{\tilde{g}}_r \approx \underline{g}_r$, it is required that the bias $n^{-1/2}\|\underline{g}_r - \underline{\tilde{g}}_r\|_2 \approx n^{-1/2}\|\underline{g}_r - W_h\underline{g}_r\|_2$ is small. We see that bias is more critical here than when determining $\widehat{L}_0$, where it is sufficient that $(1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2$ is "not much" smaller than

$(1/N)\Sigma_{j=1}^{N} n\theta_{jr}^2$. On the other hand, variances of the estimators decrease with both N and h/n (see Section 4). In connection with step 3a, final estimates of basis functions should thus be determined by relying on a comparably large h.

The above approach is not the only way of obtaining estimates of suitable basis functions. In a PCA context, Rice and Silverman (1991) propose a way to obtain a smooth nonparametric estimate of the eigenvector for the largest eigenvalue of a covariance matrix. Their method might be adapted to the present situation. A further alternative is as follows.

For any r, we have $\widetilde{\underline{g}}_r = \Sigma_{s=1}^{L_0} \delta_{sr} W_h \underline{g}_s$ for some $\delta_{1r}, \dots, \delta_{L_0 r} \in \mathbb{R}$. Suppose that W_h is such that (3.4) holds and that $\operatorname{rank}(\underset{\sim}{M_h}) = L_0$. Then the δ_{sr} are uniquely determined, and it is easily verified that if $\widetilde{\theta}_{jr}$ are the parameters of the projected model (3.6), then

$$\underline{m}_j(X_i) = \sum_{r=1}^{L_0} \widetilde{\theta}_{jr} \psi_r(X_i), \qquad i = 1, \dots, n,$$

holds for $j = 1, \dots, N$, where $\psi_r(X_i) := \Sigma_{s=1}^{L_0} \delta_{sr} g_s(X_i)$ for $r = 1, \dots, L_0$.

The $\psi_1, \dots, \psi_{L_0}$ thus establish another basis for model (1.2). Instead of the $\underline{g}_r$'s we might decide to estimate the ψ_r's. This can be done by the following procedure, which provides an alternative to step 4 above:

4a. For all j, r compute an estimate $\widehat{\widetilde{\theta}}_{jr}$ of $\widetilde{\theta}_{jr}$ by $\widehat{\widetilde{\theta}}_{jr} := n^{-1/2} \underline{\gamma}_r(\widehat{M}_h)' \cdot \widehat{\underline{m}}_j$.

4b. For all i, r determine $\widehat{\psi}_r^*(X_i) := (n/N\widehat{\lambda}_r)\Sigma_{j=1}^{N} \widehat{\widetilde{\theta}}_{jr} \cdot Y_{ij}$.

4c. Separately for each r, smooth $\widehat{\psi}_r^*(X_1), \dots, \widehat{\psi}_r^*(X_n)$ to obtain final estimates $\widehat{\psi}_r(X_1), \dots, \widehat{\psi}_r(X_n)$ of $\psi_r(X_1), \dots, \psi_r(X_n)$.

Based on (3.5) and (3.6), $\widehat{\widetilde{\theta}}_{j1}, \dots, \widehat{\widetilde{\theta}}_{j\hat{L}_0}$ in step 4a are the least squares estimates of $\widetilde{\theta}_{j1}, \dots, \widetilde{\theta}_{j\hat{L}_0}$ when replacing the $\widetilde{\underline{g}}_r$ by their estimates $n^{1/2}\underline{\gamma}_r(\widehat{M})$. It should be noted that the estimated parameters automatically satisfy $\Sigma_{j=1}^{N} \widehat{\widetilde{\theta}}_{jr} \widehat{\widetilde{\theta}}_{js} = 0$, $r \neq s$. Moreover, $\Sigma_{j=1}^{N} \widehat{\widetilde{\theta}}_{jr}^{2} = N \cdot \widehat{\lambda}_r / n$. It is then easily seen that, for all i, the preliminary estimates $\widehat{\psi}_1^*(X_i), \dots, \widehat{\psi}_{\hat{L}_0}^*(X_i)$ minimize

$$\sum_{j=1}^{N} \left(Y_{ij} - \sum_{r=1}^{\hat{L}_0} \widehat{\widetilde{\theta}}_{jr} \cdot v_{ir} \right)^2$$

over all $v_{ir} \in \mathbb{R}$. This motivates step 4b. The smoothing step 4c is straightforward.

In step 4a we will expect good parameter estimates for small h (compare the theoretical results in Section 4). Since we rely on the projected model, the values $n^{-1/2}\|\underline{g}_r - \widetilde{\underline{g}}_r\|_2$ are of *minor importance*. Basically, bias problems only arise in step 4c, but there we may do a different amount of smoothing for different components. This is an important difference to our original step 4. In fact, one might choose a smoothing parameter by cross-validation.

Since the parameters $\widehat{\widetilde{\theta}}_{jr}$ are constant over X_i, $i = 1, \dots, n$, it quite obviously makes sense to minimize a cross-validation function. We will not go into further details of steps 4a–4c, for this would result in overloading the present paper.

Given L_0, the idea of analyzing the $\widehat{M}_h$ matrix to obtain estimates of basis functions is closely related to the "self-modeling nonlinear regression" approach which has been proposed by Kneip and Gasser (1988) in a more general context. This follows from (3.1). Some more details can be found in Kneip (1987).

4. Theoretical results. In this section we will investigate the above approach from a theoretical point of view. In addition to finite sample results, some asymptotic theory is given. Based on (1.1), (1.2) and Assumption 1, asymptotics rely on sampling more and more observations by adding more and more curves ($N \to \infty$) and/or by adding more and more design points ($n \to \infty$). We will generally allow that L_0 and h increase with the sample size. More precisely, we implicitly assume sequences $N_0 \le N_1 \le N_2 \le \cdots$ and $n_0 \le n_1 \le \cdots$ of values for N, n, and corresponding sequences $L_{0,0} \le L_{0,1} \le \cdots$ and $h_0 \le h_1 \le \cdots$ for L_0, h. All asymptotic results refer to limits ($k \to \infty$) of tuples $(N_k, n_k, L_{0,k}, h_k)$. Hereby, for example, "$N \to \infty$" means that $\lim_k N_k \to \infty$ while, "L_0 is fixed" requires $L_{0,0} = L_{0,1} = \cdots$. Statements like "$n > h > L_0$" refer to all tuples $(n_k, h_k, L_{0,k})$; "$L_0 \ge a$" means $L_{0,k} \ge a$ for all k; and "$N/n = O(1)$" is equivalent to $N_k/n_k = O(1)$ as $k \to \infty$.

4.1. *Analysis of* $\widehat{M}_h$.

ASSUMPTION 2. *There exists a $D_1 < \infty$ such that, for all n,*

$$\max_{r,s=1,\dots,n} (W_h)_{rs} < D_1 \cdot \frac{h}{n}.$$

Assumption 2 will be satisfied for practically all reasonable choices of a smoother matrix W_h. Theorem 1 now provides a justification of (3.2) and claim (a) of Section 3.3. It provides finite sample results, except for assertion (iii), which relies on $n \to \infty$ (h, N fixed).

THEOREM 1. *In addition to the above assumptions suppose that $h > L_0$. Furthermore, assume that the linear subspace spanned by $W_h \cdot \underline{g}_1, \dots, W_h \cdot \underline{g}_{L_0}$ has dimension L_0. We then obtain*

$$\sum_{r=L_0+1}^{h} \widehat{\lambda}_r \le \frac{1}{N} \operatorname{tr}\left(\sum_{j=1}^{N} \left(I - \frac{1}{n} \sum_{r=1}^{L_0} \widetilde{\underline{g}}_r \widetilde{\underline{g}}'_r \right) W_{h \underline{\epsilon}_j \underline{\epsilon}'_j} \right) =: \widetilde{R}_{L_0}.$$

Moreover,

(i) $E\widetilde{R}_{L_0} = h - L_0$, $\operatorname{var}(\widetilde{R}_{L_0}) \leq 2(h - L_0)/N + (D_0 - 3)D_1 h(h - L_0)/(Nn)$.

(ii) *If the* ε_{ij} *are i.i.d.* $N(0,1)$ *distributed, then* $N \cdot \widetilde{R}_{L_0}$ *follows a* χ^2 *distribution with* $N \cdot (h - L_0)$ *degrees of freedom.*

(iii) *As* $n \to \infty$, $N \cdot \widetilde{R}_{L_0}$ *follows a* χ^2 *distribution with* $N \cdot (h - L_0)$ *degrees of freedom asymptotically.*

Here, I denotes the identity matrix. A proof of the theorem is contained in the Appendix.

For further work we have to impose some regularity conditions on the θ_{jr}.

ASSUMPTION 3. *There exist constants* $D_2 > 0$ *and* $D_3 > 0$ *such that, for all* n, N, L_0,

$$\left|\frac{1}{N}\sum_{j=1}^{N} n\theta_{jr}^2 - \frac{1}{N}\sum_{j=1}^{N} n\theta_{js}^2\right| \geq D_2 \cdot \frac{1}{N}\sum_{j=1}^{N} n\theta_{jr}^2 \quad \textit{for all } r, s \in \{1, \ldots, L_0\},\ r \neq s, \tag{4.1}$$

$$\frac{1}{N}\sum_{j=1}^{N} n\theta_{jL_0}^2 \geq D_3. \tag{4.2}$$

Condition (4.1) is technical in nature. The values of $(1/N)\Sigma_{j=1}^{N} n\theta_{jr}^2$ have to decrease quite rapidly as r increases. Condition (4.2) claims that $(1/N)\Sigma_{j=1}^{N} n\theta_{jL_0}^2$ does not converge to 0 as the sample size increases. It should be noted that components which are really "essential" for modelling a regression function m_j have to satisfy $n\theta_{jr}^2 > 1$. This is easy to see: for $L \leq L_0$ let $\widehat{\theta}_{j1}, \ldots, \widehat{\theta}_{jL}$ denote the least squares estimates of the parameters $\theta_{j1}, \ldots, \theta_{jL}$ when $\underline{g}_1, \ldots, \underline{g}_L$ are known. Then

$$E\left\|\underline{m}_j - \sum_{r=1}^{L} \widehat{\theta}_{jr}\underline{g}_r\right\|_2^2 \geq E\left\|\underline{m}_j - \sum_{r=1}^{L+1} \widehat{\theta}_{jr}\underline{g}_r\right\|_2^2 \quad \text{if and only if } n\theta_{jL+1}^2 \geq 1.$$

It should be noted that (4.2) is basically weaker than the conditions which are usually imposed for analyzing asymptotic properties of parametric models. If a model of the form (1.2) holds for some fixed $L_0 \in \mathbb{N}$, then one will typically assume that for each $L \in \{1, \ldots, L_0\}$ there exists a $c_L > 0$ such that

$$c_L \leq \frac{1}{N}\sum_{j=1}^{N} \frac{1}{n}\left\|\underline{m}_j - \sum_{r=0}^{L-1} \theta_{jr}\underline{g}_r\right\|_2^2 = \sum_{r=L}^{L_0} \frac{1}{N}\sum_{j=1}^{N} \theta_{jr}^2.$$

Here, $\theta_{j0} \equiv 0$ and $\underline{g}_0 \equiv 0$. Hence, in this case $n = O((1/N)\Sigma_{j=1}^{N} n\theta_{jL}^2)$ for all $L = 1, \ldots, L_0$, and the $(1/N)\Sigma_{j=1}^{N} n\theta_{jL}^2$ become large as n increases. One may note that, whenever $c_1 > 0$, at least some of these values will be of order n. In a more general context the following proposition can be derived.

PROPOSITION 1. *Let $N/n = O(1)$. Assume that there exists a density f such that, as $n \to \infty$, $(1/N)\Sigma_{i=1}^{n} v(X_i) \to \int_J v(X) f(X)\, dX$ holds for any continuous function $v\colon J \to \mathbb{R}$.*

Then, for any fixed $r \le L_0$, there exists a $d_r > 0$ such that $(1/N)\Sigma_{j=1}^{N} n\theta_{jr}^2 \ge d_r$ holds for all n, N sufficiently large.

A proof can be found in the Appendix.

Theorem 2 quantifies the difference between parameters and basis vectors of the true model and of the projected model. For $r = 1, \dots, L_0$ set $\mathrm{Bias}_{h,r} := n^{-1/2}\, \|\underline{g}_r - W_h \cdot \underline{g}_r\|_2$.

THEOREM 2. *Suppose that $\Sigma_{r=1}^{L_0} \mathrm{Bias}_{h,r} < D_2/8$ holds in addition to Assumptions 1–3. Then the following holds for any $r \in \{1, \dots, L_0\}$:*

$$\text{(i)} \qquad \left| \frac{1}{N} \sum_{j=1}^{N} n\theta_{jr}^2 - \frac{1}{N} \sum_{j=1}^{N} n\tilde{\theta}_{jr}^2 \right| \le \frac{5}{4} \cdot \frac{1}{N} \sum_{j=1}^{N} n\theta_{jr}^2 \cdot \mathrm{Bias}_{h,r}^2;$$

$$\text{(ii)} \qquad n^{-1/2} \|\underline{g}_r - \tilde{g}_r\|_2 \le \frac{4}{D_2} \mathrm{Bias}_{h,r}.$$

A proof is given in the Appendix. Assertion (i) implies that the relative difference between the average squared true and projected parameters is small, if $\mathrm{Bias}_{h,r}$ is not too large. In fact, we see that a moderate bias will be sufficient to guarantee that $(1/N)\Sigma_{j=1}^{N} n\tilde{\theta}_{jr}^2$ and $(1/N)\Sigma_{j=1}^{N} n\theta_{jr}^2$ are of the same order of magnitude. Based on our smoothness assumptions on the g_r's, one will thus expect that for a reasonable choice of the smoothing procedure Assumption 3 carries over to the following assumption.

ASSUMPTION 4. *There exist constants $D_4 > 0$ and $D_5 \ge 0$ such that, for all n, N, (4.1) and (4.2) hold when replacing $\theta_{jr}, \theta_{js}, \theta_{jL_0}$ by $\tilde{\theta}_{jr}, \tilde{\theta}_{js}, \tilde{\theta}_{jL_0}$ and D_2, D_3 by D_4, D_5.*

It should be noted that Assumption 3 and Theorem 2 imply that Assumption 4 holds if $\Sigma_{r=1}^{L_0} \mathrm{Bias}_{h,r} < D_2/8$. Moreover, a result similar to Proposition 1 can be derived.

PROPOSITION 2. *In addition to the assumptions of Proposition 1 suppose that, as $n \to \infty$, $(1/n)\|\underline{m}_j - W_h \cdot \underline{m}_j\|_2^2 \to 0$ holds for any j.*

Then, for any fixed $r \le L_0$, there exists a $\tilde{d}_r > 0$ such that $(1/N)\Sigma_{j=1}^{N} n\tilde{\theta}_{jr}^2 \ge \tilde{d}_r$ holds for all n, N sufficiently large.

A proof can be found in the Appendix. The following theorem yields results about the asymptotic behavior of eigenvalues and eigenvectors of $\widehat{M}_h$.

THEOREM 3. *In addition to Assumptions* 1–4 *suppose that* n, h, L_0 *satisfy* $n \geq h > L_0$, $h^2/N \to 0$ *as* $N \to \infty$.

Then, as $N \to \infty$, *the following hold*:

$$\text{(i)} \qquad \widehat{\lambda}_r = 1 + \frac{1}{N}\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2 + O_p\left(\left(\frac{1}{N^2}\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2\right)^{1/2}\right), \qquad r = 1, \dots, L_0;$$

$$\text{(ii)} \quad n^{-1/2}\|\underline{\widetilde{g}}_r - \underline{\widehat{g}}_r\|_2 = O_P\left(\frac{h^{1/2}}{N^{1/2}\left((1/N)\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2\right)^{1/2}}\right), \qquad r = 1, \dots, L_0;$$

$$\text{(iii)} \qquad \max_{r=1,\dots,L_0}\left|\frac{\widetilde{\lambda}_r - 1 - (1/N)\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2}{(1/N)\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2}\right| = O_P\left(hN^{-1/2}\right);$$

$$\text{(iv)} \qquad \widetilde{R}_{L_0} - \sum_{r=L_0+1}^{h} \widehat{\lambda}_r = O_P\left(\frac{h(h-L_0)}{N}\right).$$

The proof is based on general results in the perturbation theory for finite-dimensional spaces and is given in the Appendix. Note that the dimension of the matrix $\widehat{M}_h$ is allowed to increase with the sample size, which introduces a major complication.

4.2. *Dimensionality.* Let us consider the problem of estimating L_0. First suppose that a low-dimensional model of the form (1.2) holds. We can then derive the following result.

PROPOSITION 3. *Suppose that Assumptions* 1 *and* 2 *hold together with the assumptions of Proposition* 2. *Assume* $L_0 \in \mathbb{N}$ *is fixed and* $h^2(h - L_0)/N \to 0$, $h/n \to 0$ *as* $N, n \to \infty$. *Then, if* α *denotes a level of significance used to determine critical values* $C_{\alpha, N(h-L_0)}$ *in* (3.7), *we obtain*

$$P(\widehat{L}_0 \neq L_0) \to \alpha \quad \text{as } N, n \to \infty. \tag{4.3}$$

A proof can be found in the Appendix. It should be noted that Proposition 3 does not rely on Assumptions 3 and 4. Proposition 2 implies that condition (4.2) is automatically fulfilled, while (4.1) is not really required. If Assumption 4 holds, Theorem 3 also allows us to establish (4.3) in the case that L_0 increases with n, N not too fast. As $n \to \infty$ we may let $\alpha \to 0$ in such a way that $C_{\alpha, N(h-L)} - N(h - L) = o(N)$ for all $L < h$. Under the conditions of Theorem 3, $\widehat{L}_0$ is then a consistent estimator of L_0.

However, (4.3) does not generalize to really high-dimensional models ($L_0 \approx \min\{N, n\}$). A closer look at the approximations given in Theorems 1 and 3 shows that quite accurate dimension estimates can only be expected if L_0 is considerably smaller than n and $N^{1/2}$. Otherwise an $h > L_0$ has to be large compared to n, N. Then our critical values will no longer be appropriate.

This is not a disappointing result. As outlined in the Introduction, we will usually be interested in low-dimensional models only. However, our procedure yields a useful tool to check whether the data can be appropriately fitted by a low-dimensional model, say, $L_0 \leq L_0^* = 5$. If this is true, Proposition 3 applies. What happens if $L_0 > L_0^*$?

PROPOSITION 4. *Let $L_0^* \in \mathbb{N}$ be a fixed constant with $L_0 > L_0^*$. Suppose that Assumptions* 1 *and* 2 *hold together with the assumptions of Proposition* 2. *Furthermore, assume that $h^2/N \to 0$ as $n, N \to \infty$. Then*

$$P(\widehat{L}_0 > L_0^*) \to 1 \quad \textit{as } N, n \to \infty.$$

A proof can be found in the Appendix. We can conclude that our procedure basically provides an asymptotically consistent test of a hypothesis like, say, $L_0 \leq L_0^* := 5$. Even more, if this hypothesis is true, then asymptotically $\widehat{L}_0$ will reveal the true dimension with high probability. The method is quite powerful. We can infer from the theoretical results that we may detect components with $(1/N)\sum_{j=1}^N \theta_{jr}^2 = O(1/n)$. In fact, (4.3) even holds if the L_0th component does not improve the average model fit, that is, if $\sum_{j=1}^N E\|\underline{m}_j - \sum_{j=1}^{L_0} \widehat{\theta}_{jr} \underline{g}_r\|_2^2 \geq \sum_{j=1}^N E\|\underline{m}_j - \sum_{j=1}^{L_0 - 1} \widehat{\theta}_{jr} \underline{g}_r\|_2^2$, where $\widehat{\theta}_{jr}$ denote the least squares estimators of θ_{jr} (compare the discussion following Assumption 3).

Recall that in Section 3.3 we proposed to rely on steps 3a and 3b to deal with the problem of smoothing-parameter selection. The above theorems and propositions incorporate some conditions on h which should be taken into account when selecting h_1, h_2, h_3, h_4. Our theory then leads to an important conclusion: *final dimension estimates $\widehat{L}_0$ are reliable only if $\widehat{L}_0$ is very small compared to N, n.*

4.3. *Basis functions.* Let us now consider the problem of estimating basis functions. Under Assumptions 3 and 4 we can infer from Theorems 2 and 3 that

$$n^{-1}\|\underline{g}_r - \widehat{\underline{g}}_r\|_2^2 = O_P\left(\mathrm{Bias}_{h,r}^2 + \frac{h}{N\left((1/N)\sum_{j=1}^N n\widetilde{\theta}_{jr}^2\right)}\right)$$

holds for $r = 1, \dots, L_0$. Obviously, the smaller $(1/N)\sum_{j=1}^N n\widetilde{\theta}_{jr}^2$ is, the higher the variability of the estimates.

The above relation allows us to derive optimal rates of convergence in different contexts. For example, consider components with $(1/N)\sum_{j=1}^N n\widetilde{\theta}_{jr}^2 \geq c \cdot n$

for some $c < \infty$. Let us assume that $\text{Bias}_{h,r} = O(h^{-\nu})$ for some $\nu \in \mathbb{N}$. Under some regularity conditions on the design this holds for least squares approximation by polynomials, provided that $d = 1$ and that g_r is ν times continuously differentiable. Then $h = O((Nn)^{1/(2\nu+1)})$ and $(Nn)^{1/(2\nu+1)} = O(h)$ lead to

$$n^{-1}\|\underline{g}_r - \widehat{\underline{g}}_r\|_2^2 = O_P\big((Nn)^{-2\nu/(2\nu+1)}\big), \tag{4.4}$$

which is much faster than the best obtainable rates for individual estimates [cf. Stone (1982)]. Rates of convergence obviously deteriorate if $(1/N)\times \sum_{j=1}^N n\, \tilde{\theta}_{jr}^2 = o(n)$.

5. Adjusting variances. As outlined in Section 2.1, in most applications our basic assumption $\text{var}(\epsilon_{ij}) = 1$ will not be satisfied automatically.

5.1. *Homoscedastic errors.* Assume that there exist data $\mathbf{Y}_{ij}$ satisfying $\mathbf{Y}_{ij} = \mathbf{m}_j(X_i) + \epsilon_{ij}$, where the ϵ_{ij} fulfill Assumption 1 except that $\text{var}(\epsilon_{ij}) = \sigma_j^2 \neq 1$.

The theory developed in the previous sections applies to $Y_{ij} := \mathbf{Y}_{ij}/\sigma_j$, $\epsilon_{ij} = \epsilon_{ij}/\sigma_j$ and $\mathbf{m}_j(X_i)/\sigma_j =: m_j(X_i) = \sum_{r=1}^{L_0}\theta_{jr}g_r(X_i)$. According to our methodology, eigenvalues and eigenvectors of $\widehat{M}_h := (1/N)\sum_{j=1}^N W_h\underline{Y}_j\underline{Y}_j'W_h$ provide estimates of L_0 and $\underline{g}_1, \underline{g}_2, \dots$.

However, in practice we can only determine estimates $\widehat{\sigma}_j^2$ of the σ_j^2. For $d = 1$, variance estimators with very favorable asymptotic properties have been proposed, for example, by Rice (1984), Gasser, Sroka and Jennen-Steinmetz (1986) and Hall, Kay and Titterington (1990). What we then actually do when applying the procedure of Section 3.2 is to analyze

$$\widehat{M}_h^* := \frac{1}{N}\sum_{j=1}^N W_h\left(\frac{\underline{\mathbf{Y}}_j}{\widehat{\sigma}_j}\right)\left(\frac{\underline{\mathbf{Y}}_j}{\widehat{\sigma}_j}\right)' W_h = \frac{1}{N}\sum_{j=1}^N \frac{\sigma_j^2}{\widehat{\sigma}_j^2} W_h\underline{Y}_j\underline{Y}_j'W_h$$

and its eigenvalues $\widehat{\lambda}_1^* \geq \cdots \geq \widehat{\lambda}_h^*$ and eigenvectors $\underline{\gamma}_1(\widehat{M}_h^*), \dots, \underline{\gamma}_r(\widehat{M}_h^*)$. Let $\widehat{\underline{g}}_r^* := n^{1/2}\underline{\gamma}_r(\widehat{M}_h^*)$.

The question arises whether the differences between eigenvalues and eigenvectors of $\widehat{M}_h$ and $\widehat{M}_h^*$ invalidate the theoretical justification of our method. This is not the case, as is shown by Theorem 4. The theorem is based on Assumptions 1–4. Furthermore, the following additional assumption is required.

ASSUMPTION 5.

(a) *There exist constants $D_6^* < \infty, D_6 > 0$ such that $D_6^* \geq \sigma_j^2 \geq D_6$ for all N and all $j = 1, \dots, N$.*

(b) *There is a constant $D_7 < \infty$ such that $|\theta_{jr}| \leq D_7$ for all j, r.*

(c) *For all n, N the estimators $\widehat{\sigma}_j^2$ of σ_j^2 are obtained as $\widehat{\sigma}_j^2 := \max\{\Omega(\underline{\mathbf{Y}}_j), D_8\}$ $(D_6 > D_8 > 0)$, where*

$$\Omega(\underline{\mathbf{Y}}_j) := \sum_{i=1}^n\sum_{k=1}^n \omega_{ik}\mathbf{Y}_{ij}\mathbf{Y}_{kj}$$

for some $\omega_{ik} \in \mathbb{R}$. *Furthermore, there exist constants* $D_9, D_{10} < \infty$ *such that the following hold*: (1) $\sum_{k=1}^{n} |\omega_{ik}| \leq D_9 \cdot 1/n$ *and* $\sum_{k=1}^{n} |\omega_{ki}| \leq D_9 \cdot 1/n$, $i = 1, \dots, n$; (2) $|E\Omega(\underline{\mathbf{Y}}_j) - \sigma_j^2| \leq D_{10} \cdot 1/n$ *for all* n, N *and* $j = 1, \dots, N$.

Assumption 5(c) establishes conditions on the variance estimator. The structural form required for $\Omega(\cdot)$ holds for each of the estimators cited above. In the present paper these estimators are slightly modified by introducing a lower bound D_8. This has been done for technical convenience. It does not impose a real restriction for practical applications. All estimators cited satisfy condition 1 of Assumption 5(c), and condition 2 of Assumption 5(c) holds if each m_j is sufficiently smooth. For example, the results of Gasser, Sroka and Jennen-Steinmetz (1986) imply that under some regularity conditions on the design $|E\Omega(\underline{\mathbf{Y}}_j) - \sigma_j^2| = O(n^{-2})$ holds for their estimator, provided m_j is twice continuously differentiable.

The following theorem is based on the asymptotic setup used to establish Theorem 3. The only exception is that it is now required that both $N, n \to \infty$.

THEOREM 4. *In addition to Assumptions* 1–5 *suppose that* $n \geq h > L_0$, $h^2/N \to 0$ *and* $h^2/n \to 0$ *as* $N, n \to \infty$.

Then, as $N, n \to \infty$, *the following hold*:

(i) $$\hat{\lambda}_r^* = 1 + \frac{1}{N}\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2 + O_P\left(\frac{(1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2}{n} + \frac{\left((1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2\right)^{1/2}}{N^{1/2}}\right), \qquad r = 1, \dots, L_0;$$

(ii) $$n^{-1/2}\|\tilde{\underline{g}}_r - \hat{\underline{g}}_r^*\|_2 = O_P\left(\frac{h^{1/2}}{N^{1/2}\left((1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2\right)^{1/2}} + \frac{h^{1/2}}{n}\right), \qquad r = 1, \dots, L_0;$$

(iii) $$\max_{r=1,\dots,L_0}\left|\frac{\hat{\lambda}_r^* - 1 - (1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2}{(1/N)\sum_{j=1}^{N} n\tilde{\theta}_{jr}^2}\right| = O_P\left(h\left(N^{-1/2} + n^{-1}\right)\right);$$

(iv) $$\left|\sum_{r=L_0+1}^{h} \hat{\lambda}_r^* - \tilde{R}_{L_0}\right| = O_P\left(\frac{h(h - L_0)}{N} + \frac{(h - L_0)}{n + (Nn)^{1/2}}\right).$$

We see that now it is important that both N and n are large compared to h. Then we can conclude from Theorem 4 that the differences between $\hat{\lambda}_s^*$, $\hat{\underline{g}}_r^*$ and $\hat{\lambda}_s$, $\hat{\underline{g}}_r$ will not have a substantial influence on our procedure to estimate components of an appropriate model. Propositions 3 and 4 generalize to the present situation. There is, however, a small difference in the optimal rates of convergence of the estimates of the $\underline{g}_r$: instead of (4.4) we obtain $n^{-1}\|\underline{g}_r - \underline{\widehat{g}}_r^*\|_2^2 = O((Nn + n^2)^{-2\nu/(2\nu+1)})$.

5.2. *Heteroscedastic errors.* Suppose that there exist data $\mathbf{Y}_{ij}$ satisfying $\mathbf{Y}_{ij} = \mathbf{m}_j(X_i) + \varepsilon_{ij}$, where the ε_{ij} satisfy Assumption 1 except that $\operatorname{var}(\varepsilon_{ij}) = \sigma_j^2(X_i) \neq 1$. Let us assume that the functions $\sigma_j^2(\cdot)$ are smooth. To avoid a lengthy treatment, we will not consider this situation in great detail. Only some important aspects are presented.

Some additional conceptual work is necessary. Thus let us assume for the moment that we know the true variances $\sigma_j^2(X_i)$. The difficulty in dealing with heteroscedastic variances is that a model for the functions $\mathbf{m}_j(\cdot)/\sigma_j(\cdot)$ does not necessarily lead to a model for the functions $\mathbf{m}_j$. This is only true if the variances can be decomposed: there exists a function $s(\cdot)$ and N constants $\overline{\sigma}_1, \dots, \overline{\sigma}_N$ such that $\sigma_j^2(X_i) = \overline{\sigma}_j^2 \cdot s(X_i)^2$. Without restriction we may require that $(1/n)\sum_{i=1}^n s(X_i)^2 = 1$, which leads to

$$\overline{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^n \sigma_j^2(X_i), \tag{5.1}$$

$$s(X_i)^2 = \frac{1}{N}\sum_{j=1}^N \frac{\sigma_j^2(X_i)}{\overline{\sigma}_j^2}. \tag{5.2}$$

One then can determine transformed data Y_{ij} by multiplying by $1/(\overline{\sigma}_j s(X_i))$. A model $\mathbf{m}_j(X_i) = \sum_{r=1}^{L_0} \vartheta_{jr}\mathbf{g}_r(X_i)$ for the original data implies

$$m_j(X_i) := \frac{\mathbf{m}_j(X_i)}{\overline{\sigma}_j s(X_i)} = \sum_{r=1}^{L_0} \frac{\vartheta_{jr}}{\overline{\sigma}_j}\frac{\mathbf{g}_r(X_i)}{s(X_i)} =: \sum_{r=1}^{L_0} \theta_{jr} g_r(X_i) \tag{5.3}$$

Using the transformed observations, our methodology leads to estimates $\widehat{L}_0$ and $\underline{\widehat{g}}_r = (\widehat{g}_r(X_1), \dots, \widehat{g}_r(X_n))'$ of L_0 and the $\underline{g}_r$, and

$$\widehat{\mathbf{g}}_r(X_i) := s(X_i) \cdot \widehat{g}_r(X_i), \qquad i = 1, \dots, n, \tag{5.4}$$

provide estimates of the $\mathbf{g}_r(X_i)$.

Now, let us consider the general case that variances do not allow the above decomposition. But even then one might use (5.1) and (5.2) to determine constants $\overline{\sigma}_j^2$ and a function s. After transforming the data by multiplying by $1/(\overline{\sigma}_j s(X_i))$, one then might estimate the components of model (5.3). When analyzing the resulting $\widehat{M}_h$-matrix it turns out that the assertions of Theorem 3 hold unchanged. Although the transformed error terms do not satisfy Assumption 1(b), we still obtain $(1/N)\sum_{j=1}^N E\epsilon_{ij}^2 = 1$ for all i. When passing through the proofs it is easily seen that this is all that is actually needed to establish Theorem 3. We can thus expect that (5.4) yields reasonable estimates of underlying basis functions. However, $\widetilde{R}_{L_0}$ will no longer follow a $\chi^2_{N(h-L_0)}$ distribution. This evidently requires an adjustment of the critical values to be used for determining $\widehat{L}_0$. Fortunately this can quite easily be done. We still have $E\widetilde{R}_{L_0} = h - L_0$. If n is large compared to h, the following holds:

$$\operatorname{var}(\widetilde{R}_{L_0}) \approx \frac{1}{N^2}\sum_{j=1}^{N} 2\operatorname{tr}\left(\left(\left(I - \sum_{r=1}^{L_0} \underline{\gamma}_r(\widehat{M}_h)\underline{\gamma}_r(\widehat{M}_h)'\right)W_h\Sigma_j\right)^2\right),$$

where Σ_j is a diagonal matrix with

$$\operatorname{diag}(\Sigma_j) = \left(\frac{\sigma_j^2(X_1)}{\overline{\sigma}_j^2 s(X_1)^2}, \dots, \frac{\sigma_j^2(X_n)}{\overline{\sigma}_j^2 s(X_n)^2}\right).$$

Noting further that $N^{1/2}(h - L_0)^{-1/2}\widetilde{R}_{L_0}$ is asymptotically normal for large n and N, these formulas can be used to determine approximate critical values for different L and α.

How to estimate the variances $\sigma_j^2(X_i)$? Following Gasser, Sroka and Jennen-Steinmetz (1986), we might determine squared pseudoresiduals

$$r_{ij}^2 := \frac{\left(\mathbf{Y}_{ij} - C_i^{(1)}\mathbf{Y}_{i-1j} - C_i^{(2)}\mathbf{Y}_{i+1j}\right)^2}{1 + C_i^{(1)2} + C_i^{(2)2}},$$

where $C_i^{(1)} := (X_{i+1} - X_i)/(X_{i+1} - X_{i-1})$, $C_i^{(2)} := 1 - C_i^{(1)}$.

By using, for example, kernel estimators one then might smooth these squared pseudoresiduals to obtain estimates $\widehat{\sigma}_j^2(X_i)$ of $\sigma_j^2(X_i)$. Let us consider Gasser–Müller kernel estimators based on a bandwidth b and a kernel of order k [cf. Gasser and Müller (1984)]. Then under some regularity conditions on the design and on the smoothness of the m_j and $\sigma_j^2(\cdot)$ it can be shown that, for large n, bias and variance of these estimators are of the order $O(b^k)$ and $O(1/(nb))$. When considering the influence of the error in variance estimation, results very similar to those of Theorem 4 can be derived: qualitative properties of our procedure are left unchanged. It should be noted that it is important to choose a small bandwidth. In either (5.1) or (5.2) we have to determine *averages*. This introduces a further smoothing. We obtain $\operatorname{var}(\widehat{\overline{\sigma}_j^2}) = O(1/n)$ and $\operatorname{var}(\widehat{s}(X_i)) = O(1/(Nnb))$.

6. An application: family expenditure survey. In this section we apply the techniques developed above to U. K. family expenditure survey (FES) data from 1968 to 1983 [HMSO(1983)]. For each of these years the data reports the expenditures on various goods of approximately 7000 households. Separately for each year, households were selected at random from electoral registers. The data contains total expenditure and expenditures on nine commodity aggregates: housing; fuel; food; clothing; durables; transport; services; alcohol and tobacco; and "miscellaneous and other goods."

Starting with Engel (1857), a major issue in applied demand analysis has been the estimation of "cross-sectional Engel curves," that is, the conditional expectations of expenditures on a commodity aggregate given total expenditures. Most work has been done in the context of parametric models. An overview of different parametric approaches is given by Deaton (1986). All important parametric models for Engel curves are linear and, hence, of the form (1.2). Furthermore, they are low dimensional: L_0 typically varies between 2 and 3. Such models have provided reasonable fits in many applications, but none of them seems to be fully satisfactory. Nevertheless the relative success of these models might lead to the hypothesis that a low-dimensional model is sufficient for describing Engel curves. An appropriate model for the FES data will provide a first step in this direction.

For estimation it is quite convenient not to use expenditures on a commodity aggregate directly, but to rely on shares of total expenditures. This leads to "budget share Engel curves." Furthermore, to improve comparability of Engel curves over different years, it is reasonable to normalize total expenditure by dividing by mean total expenditure (separately for each year).

Let (y_{lkt}, x_{lt}) denote the resulting data. Here k indicates commodity aggregates; t, years; and l, households. Data for very rich and very poor households is sparse and not very reliable. Estimating and modelling Engel curves thus has to take place within a compact domain. Since, by normalization, the value 1 for x_{lt} corresponds to mean total expenditure in the year t, $J := [0.25, 2.5]$ seems a reasonable choice. Normalized expenditures for approximately 95% of all households fall into this range, differing a little bit from year to year.

When trying to apply the above methodology to this data, a further difficulty arises. In the present application we have a random design and the x_{lt} are not identical for different years. One now might specify a grid $0.25 =: X_0 < X_1 < X_2 < \cdots < X_n < 2.5 =: X_{n+1}$ of n points and define new "data" (Y_{ikt}, X_i), where, for given i, k, t, Y_{ikt} denotes the average over all y_{lkt} corresponding to some $x_{lt} \in [(X_{i-1} + X_i)/2, (X_i + X_{i+1})/2)$. If these intervals are small enough such that the bias is negligible, we obtain $m_{kt}(X_i) := E(y_{lkt} \mid X_i) \approx EY_{ikt}$. For the present analysis a grid of 231 points was chosen. Grid points X_i were selected in such a way that some adaptation to the design density was reached. Combining now k and t to a common variable j, we end up with data with can be written in the form (1.1). We have $n = 231$ and $N = 144$.

REMARK 4.

(a) Basically any grid $X_1, \ldots, X_n$ is appropriate which is dense enough to guarantee a small bias. Suppose we choose a comparably small number n of grid points. Then averaging takes place over a large number of y_{lkt}. This results in comparably small variances of the Y_{ikt} (as well as in "more" normality). Thus, a small n is balanced by the effect that when transforming the data we have to divide by small standard deviations (see Section 5).

(b) There is a deviation from our assumption on the independence of error terms. We have $\sum_{k=1}^{9} y_{lkt} = 1$ for all t, l, that is, expenditures on the nine com-

TABLE 1
Analysis of eigenvalues, dimensionality; $h = 7$

L	**Eigenvalues**	$\sum_{r=L}^{7} \widehat{\lambda}_r$	**Critical value ($\alpha = 0.05$)**
1	18669.60	19177.77	7.52
2	471.44	508.17	6.48
3	29.99	36.76	5.44
4	3.33	6.74	4.40
5	1.51	3.42	3.34
6	0.98	1.90	2.28
7	0.92	0.92	1.20

modity aggregates sum up to total expenditure. Any Y_{ij} is thus correlated with the observations $Y_{i\iota_s}$, $s = 1, \dots, q$, of $q = 8$ out of the 144 curves. It is now easily seen that the theoretical results given in Theorems 3 and 4 generalize to the case that as N increases there are such dependencies for certain collections of $q + 1$ out of the N curves ($q \in \mathbb{N}$ fixed). However, these dependencies may have some influence on the asymptotic distribution of $\widetilde{R}_{L_0}$. As for heteroscedastic errors one then might adjust critical values. It still holds that $E\widetilde{R}_{L_0} = h - L_0$. The asymptotic variance of $\widetilde{R}_{L_0}$ might be approximated using nonparametric estimates of covariances.

We are now ready to apply the methodology described above. Error variances are heteroscedastic and we have to use the transformations presented in Section 5.2. Smoothing was based on least squares fits of polynomials and cubic B-splines. For polynomials, generalized cross-validation brought out $h^* = 7$ as average optimal smoothing parameter.

Let us first consider the dimensionality of the model. Table 1 presents results obtained when analyzing the eigenvalues of the $\widehat{M}_h$ matrix according to step 3 of our procedure. Here, smoothing is based on cubic B-splines with five knots at 0.2, 0.7, 1.25, 1.85, 2.7; $h = 7$. Recall that $\widehat{\lambda}_r$ estimates $1 + (1/N)\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2$, if $r \leq L_0$.

The critical values given stem from χ^2 approximations. This is justified, since adjustments, as outlined above, did not lead to essential changes. Differences usually varied between 2 and 3%. The eigenvalues now lead to $\widehat{L}_0 = 5$. A closer look at the table shows that four components are highly significant, while the fifth is somewhat dubious. Further applications of the method confirm this result. It also holds for $h = 6, 8$ and for both polynomial regression and cubic B-splines. There is no evidence that more than five base functions are required. In fact, $\sum_{r=7}^{h} \widehat{\lambda}_r < (h - 6)$ was obtained throughout.

Choices of $h = 9$ or $h = 15$ lead to $\widehat{L}_0 = 4$. Table 2 reports results when smoothing is based on cubic B-splines with 49 knots ($h = 51$).

Here, noise obviously superimposes small components. We end up with $\widehat{L}_0 = 3$.

TABLE 2
Analysis of eigenvalues, dimensionality; $h = 51$

L	**Eigenvalues**	$\sum_{r=L}^{51} \widehat{\lambda}_r$	**Critical value** **($\alpha = 0.05$)**
1	18669.82	19219.51	52.39
2	471.77	549.69	51.38
3	30.58	77.92	50.36
4	3.91	47.34	49.35
5	2.30	43.43	48.34
6	2.26	41.13	47.32
7	2.17	38.87	46.31

TABLE 3
Approximation of estimated base functions

L	**Eigenvalues**	**AE($\widehat{g}_L$)**
1	18669.62	0.000004
2	471.51	0.000176
3	30.10	0.005959
4	3.34	0.021633

When combining these results we can conclude that the data supports our hypothesis that a low-dimensional model of the form (1.2) is appropriate. Furthermore, four or five base functions should be sufficient.

We now might use one of the procedures presented in Section 3 to estimate the functional values of base functions at the design points. However, a closed-form analytic expression is certainly much more desirable. Based on dimension estimation, data might again give hints:

Consider a smooth, strictly monotone transformation $x \to T(x)$. For half of the data, say, $j = 73, \dots, 144$, then determine Y_{ij}^* by averaging over all $y_{ktl}(k \equiv k(j), t \equiv t(j))$ corresponding to some x_{tl} with $T^{-1}(x_{tl}) \in [(X_{i-1} + X_i)/2, (X_i + X_{i+1})/2$. Then estimate the dimension of a model for the data (Y_{ij}^{**}, X_i), where $Y_{ij}^{**} := Y_{ij}, j = 1, \dots, 72$, and $Y_{ij}^{**} := Y_{ij}^*, j = 73, \dots, 144$. Approximately,

$$Y_{ij}^{**} = \sum_{r=1}^{L_0} \theta_{jr} g_r(X_i) + \epsilon_{ij}, \qquad j = 1, \dots, 72,$$

and

$$Y_{ij}^{**} = \sum_{r=1}^{L_0} \theta_{jr} g_r\big(T(X_i)\big) + \epsilon_{ij}, \qquad j = 73, \dots, 144.$$

hold. However, usually it will not be true that $g_r(T(\cdot)) \in \operatorname{span}\{g_1, \dots, g_{L_0}\}$, and the dimension L_0^{**} of a model for (Y_{ij}^{**}, X_i) will be larger than L_0. If $L_0^{**} = L_0$, this allows us to draw conclusions about the structure of the base functions.

Using transformations of the form $T(x) = t \cdot x$ and $T(x) = x^t$, $\widehat{L}_0^{**} = \widehat{L}_0$ was obtained for all t in a reasonably large neighborhood of 1. Other trans-

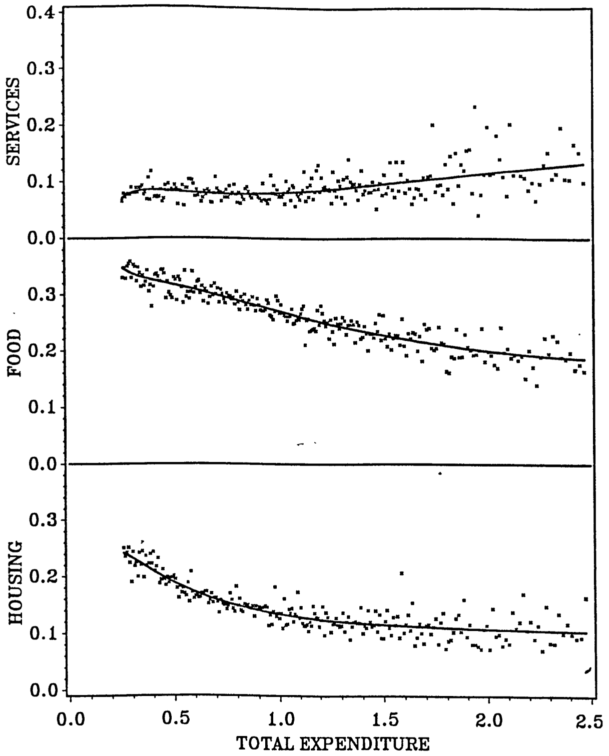

FIG. 1. *Estimated budget-share Engel curves of three commodities for the year* 1973.

formations [as, e.g., $T(x) = x + t$] lead to an increase of dimensionality. Furthermore, goodness-of-fit tests were applied to check whether base functions estimated from (Y_{ij}^*, X_i) (transformation applied to all j) are able to model the original data (Y_{ij}^*, X_i). Again results were consistent with the assumption that $g_r(T(\cdot)) \in \text{span}\{g_1, \dots, g_{L_0}\}$, $r = 1, \dots, L_0$, holds for transformations of the form $T(x) = t \cdot x$ and $T(x) = x^t$. This already suffices to fix the structure of the g_r's. It can be proved that the only functions with these properties are polynomials in $\log(X)$. Table 3 illustrates that estimated base functions can indeed be well approximated by functions of the form $\sum_{r=0}^4 \delta_r(\log(X))^r$. It shows the approximation error,

$$\text{AE}\,(\hat{g}_L) := \min_{\delta_0, \dots, \delta_4} \frac{1}{n} \sum_{i=1}^{N} \left(\hat{g}_L(X_i) - \sum_{r=0}^{4} \delta_r \big(\log(X_i) \big)^r \right)^2,$$

for the estimated base functions $\hat{g}_1, \hat{g}_2, \hat{g}_3, \hat{g}_4$ of the four highly significant components.

Estimates were obtained according to step 4 of our procedure. Smoothing was based on cubic B-splines with 7 knots at 0.2, 0.45, 0.8, 1.15, 1.5, 2.0, 2.7; $h = 9$. Recall that the theoretical results imply that the larger the value of the corresponding eigenvalue, the better the estimates of base functions.

We thus arrive at the conclusion that the budget-share Engel curves of the FES data can be described by the model $m_j(X) = \Sigma_{r=0}^{4}\theta_{jr}(\log(X))^r$. This implies the model $\Sigma_{r=0}^{4}\theta_{jr}X(\log(X))^r$ for the cross-sectional Engel curves. Different goodness-of-fit tests were applied to check this model for each of the 144 curves. Between six and eight rejections were obtained for $\alpha = 0.05$, and between one and two for $\alpha = 0.01$. This is evidently in accordance with the hypotheses that the model is appropriate. Figure 1 shows weighted least squares fits of the model for three different commodity aggregates in the year 1973.

APPENDIX

For an $n \times n$ matrix A let $\mathcal{L}(A) := \{A \cdot v \mid v \in \mathbb{R}^n\}$.

Proof of Theorem 1. Let $\underline{v}_r := n^{-1/2}\underline{\widetilde{g}}_r$ for $r = 1, \ldots, L_0$. Select orthonormal vectors $\underline{v}_{L_0+1}, \ldots, \underline{v}_h$ such that $\underline{v}_1, \ldots, \underline{v}_h$ is an orthonormal basis of $\mathcal{L}(W_h) \supset \mathcal{L}(\widehat{M}_h)$. Set $\bar{\epsilon}_{kj} := \underline{v}_k \cdot \underline{\epsilon}_j$ and $\underline{\bar{\epsilon}}_j := (\bar{\epsilon}_{ij}, \ldots, \bar{\epsilon}_{hj})'$, and let Γ_h, Γ_{L_0} and Γ_{h-L_0} denote the $n \times h, n \times L_0$ and $n \times (h - L_0)$ matrices $(\underline{v}_1, \ldots, \underline{v}_h), (\underline{v}_1, \ldots, \underline{v}_{L_0})$ and $(\underline{v}_{L_0+1}, \ldots, \underline{v}_h)$.

Clearly, $W_h = \Gamma_h\Gamma_h'$ and $\widehat{M}_h = \Gamma_h\Gamma_h'\widehat{M}_h\Gamma_h\Gamma_h'$. It is now immediately seen that the largest h eigenvalues of $\widehat{M}_h$ and of $\widehat{\Lambda} := \Gamma_h'\widehat{M}_h\Gamma_h$ are identical, or, more precisely,

$$\lambda_r(\widehat{M}_h) = \lambda_r(\widehat{\Lambda}), \qquad r = 1, \ldots, h. \tag{A.1}$$

Obviously, $\lambda_r(\widehat{M}_h) = 0$ for $r > h$. Let $\widehat{\Lambda}_{rs}$ denote the entries of $\widehat{\Lambda}$. It is immediately seen that $\widehat{\Lambda}_{rr}(1/N)\Sigma_{j=1}^{N}\epsilon_{rj}^2$ holds for any $r > L_0$.

Since trivially $\Sigma_{r=1}^{L_0}\lambda_r(\widehat{\Lambda}) \geq \Sigma_{r=1}^{L_0}\widehat{\Lambda}_{rr}$ it thus follows that

$$\begin{aligned}\sum_{r=L_0+1}^{h} \lambda_r(\widehat{M}_h) &= \sum_{r=L_0+1}^{h} \lambda_r(\widehat{\Lambda}) \\ &\leq \sum_{r=L_0+1}^{h} \widehat{\Lambda}_{rr} = \sum_{r=L_0+1}^{h} \frac{1}{N}\sum_{j=1}^{N}\bar{\epsilon}_{rj}^2,\end{aligned} \tag{A.2}$$

On the other hand, note that $(1/n)\Sigma_{r=1}^{L_0}{}'\underline{\widetilde{g}}_r\underline{\widetilde{g}}_r' = \Gamma_{L_0}\Gamma_{L_0}'$. Some easy computations lead to

$$\begin{aligned}\widetilde{R}_{L_0} &= \operatorname{tr}\left(\Gamma_{h-L_0}\Gamma_{h-L_0}'\left(\frac{1}{N}\sum_{j=1}^{N}\underline{\epsilon}_j\underline{\epsilon}_j'\right)\Gamma_{h-L_0}\Gamma_{h-L_0}'\right) \\ &= \sum_{r=L_0+1}^{h}\frac{1}{N}\sum_{j=1}^{N}\bar{\epsilon}_{rj}^2,\end{aligned} \tag{A.3}$$

which together with (A.2) proves the first assertion of the theorem.

By construction $\underline{v}_1, \dots, \underline{v}_h$ are orthonormal. It thus follows from standard results that if the ϵ_{ij} are i.i.d. $N(0,1)$ distributed, so are the $\bar{\epsilon}_{rj}$. Theorem 1(ii) is then an immediate consequence of (A.3).

More generally, Assumption 1 implies that $\underline{\bar{\epsilon}}_j$ and $\underline{\bar{\epsilon}}_k$ are independent for $j \neq k$, and

$$E\underline{\bar{\epsilon}}_j = 0, \qquad \operatorname{cov}(\underline{\bar{\epsilon}}_j) = I_h, \qquad j = 1, \dots, N, \tag{A.4}$$

where I_h denotes the $h \times h$ identity matrix. Hence, $E\widetilde{R}_{L_0} = h - L_0$. Let v_{rs} denote the elments of the vector $\underline{v}_r$. Since $\Gamma_h \Gamma_h' = W_h$, we can infer from Assumption 2 that, for any i,

$$v_{1i}^2 + \cdots + v_{hi}^2 \leq D_1 \frac{h}{n}. \tag{A.5}$$

Straightforward computations now yield

$$\begin{aligned} \operatorname{var}(\widetilde{R}_{L_0}) &= \frac{2(h-L_0)}{N} + \frac{1}{N^2} \sum_{j=1}^{N} \sum_{r=L_0+1}^{h} \sum_{i=1}^{n} v_{ri}^2 (v_{L_0+1i}^2 + \cdots + v_{hi}^2)(E\epsilon_{ij}^4 - 3) \\ &\leq \frac{2}{N}(h-L_0)D_1 + \frac{1}{N}(h-L_0)D_1 \frac{h}{n}(D_0 - 3). \end{aligned} \tag{A.6}$$

It remains to prove Theorem 1(iii). Using the Cramér–Wald device, (A.4), (A.5) and Assumption 1(c), we can deduce from standard martingale central limit theorems [cf. Pollard (1984), page 171] that, for all j, $\underline{\bar{\epsilon}}_j$ follows a $N(0, I_h)$ distribution asymptotically ($n \to \infty$). Together with (A.3) this completes the proof of the theorem. □

PROOF OF PROPOSITION 1. Let $r < L_0$, and let $N^* \in \mathbb{N}$ be such that the dimension of the space spanned by $\underline{m}_1, \dots, \underline{m}_{N^*}$ is $L_0^* \geq r$ for some $L_0^* \leq L_0$. It is then easily seen that the $N^* \times N^*$ matrix

$$Q := \left(\int_J m_j(X) m_k(X) f(X)\, dX \right)_{j,k=1,\dots,N^*}$$

possesses exactly L_0^* nonzero eigenvalues $l_1 \geq \cdots \geq l_{L_0^*} > 0$. Consider the matrices

$$Q_n := \left(\frac{1}{n} \underline{m}_j' \underline{m}_k \right)_{j,k=1,\dots,N^*} \quad \text{and} \quad M^* := \sum_{j=1}^{N^*} \frac{1}{n} \underline{m}_j \underline{m}_j'.$$

It is easily seen $\lambda_r(Q_n) = \lambda_r(M^*)$ holds for all $r \leq \min\{N^*, n\}$. By our assumption on the X_i, Q_n converges to Q as $n \to \infty$. It follows from Kato [(1966), Theorem 5.14 of Chapter 2], that

$$\lambda_r(M^*) = \lambda_r(Q_n) \to l_r, \qquad r = 1, \dots, L_0^*, \text{ as } n \to \infty.$$

Together with $N/n = O(1)$ this implies the existence of $d_r > 0$ such that $\lambda_r(M^* \cdot n/N) \geq d_r$ for all n, N sufficiently large. However,

$$\frac{1}{N}\sum_{j=1}^{N} \underline{m}_j\underline{m}_j' = M = M^* \cdot \frac{n}{N} + \frac{1}{N}\sum_{j=N^*+1}^{N} \underline{m}_j\underline{m}_j' =: M^* \cdot \frac{n}{N} + M^{**}.$$

The matrix M^{**} is positive semidefinite. We can conclude that $(1/N)\Sigma_{j=1}^{N} n\theta_{jr}^2 = \lambda_r(M) \geq \lambda_r(M^* n/N) \geq d_r$ for all n, N sufficiently large, as follows from a theorem of Anderson and Dasgupta (1963). □

PROOF OF PROPOSITION 2. The proof is analoguous to the proof of Proposition 1, when noting that $(1/n)\|\underline{m}_j - W_h\underline{m}_j\|_2^2 \to 0$ for all j implies that $\widetilde{Q}_n := ((1/n)\underline{\widetilde{m}}_j'\underline{\widetilde{m}}_k)_{j,k=1,\dots,N^*}$ converges to Q. □

Generally speaking, Theorems 2–4 make assertions about relations between eigenvalues and eigenvectors of matrices A and A^* which in some sense are not "very" different. This requires us to consider how eigenvalues and eigenvectors change when passing over from A to $A + B (B := A^* - A)$. Lemma 1 provides the basic theoretical tool to deal with this question.

We will need a matrix norm. For a real or complex $k \times k$ matrix $A, k \in \mathbb{N}$, set

$$\|A\| := \sup_{v \in \mathbb{C}^k, \|v\|_2 = 1} \|A \cdot v\|_2,$$

where $\|z\|_2 := (\Sigma_{r=1}^{k}|z|^2)^{1/2}$ for $z \in \mathbb{C}^k$. Of course, if A is real, $\|A\|$ can be determined by considering the supremum over real $v \in \mathbb{R}^k$, only.

We will use $\mathcal{EG}(A)$ to denote the set of all eigenvalues of A. For $\lambda \in \mathcal{EG}(A)$, $\mathcal{EV}(A, \lambda)$ will denote the set of all normalized eigenvectors of A for λ. We say that a vector v is "normalized" if $\|v\|_2 = 1$.

LEMMA 1. *Let A and B be real $k \times k$ matrices, $k \in \mathbb{N}$. Suppose that A is symmetric and that, for some $k_0, 1 \leq k_0 < k$, it holds that $\lambda_1(A) > \lambda_2(A) > \cdots > \lambda_{k_0}(A) > \lambda_{k_0+1}(A)$ and $\lambda_{k_0+1}(A) = \cdots = \lambda_k(A) =: \mathbf{l}$.*

For $\lambda \in \mathcal{EG}(A)$ use $U(\lambda)$ to denote the projection matrix projecting onto the eigenspace of A for λ, and let

$$\mathbf{S}(\lambda) := \sum_{\ell \in \mathcal{EG}(A)\setminus\{\lambda\}} \frac{1}{\ell - \lambda} U(\ell) \tag{A.7}$$

be the reduced resolvent of A for λ. Furthermore, for $\beta > 0$ let $\mathcal{EG}^*(A, B, \beta)$ denote the set of all $\lambda \in \mathcal{EG}(A)$ such that

$$\|\overline{\mathbf{S}}(\lambda)^{(p_1)}B\overline{\mathbf{S}}(\lambda)^{(p_2)}\cdots\overline{\mathbf{S}}(\lambda)^{(p_q)}B\overline{\mathbf{S}}(\lambda)^{(p_{q+1})}\| \leq \frac{\beta^q}{\|\mathbf{S}(\lambda)\|^{q-p}} \tag{A.8}$$

holds for all $q \in \mathbb{N}$ and $p_1, \dots, p_{q+1} \in \mathbb{N} \cup \{0\}$ with $p_1 + \cdots + p_{q+1} = p \leq q$. Here, $\overline{\mathbf{S}}(\lambda)^{(0)} = -U(\lambda)$ and $\overline{\mathbf{S}}(\lambda)^{(\delta)} = \mathbf{S}(\lambda)^\delta$, $\delta \in \mathbb{N}$.

Then if, for some $\beta \leq 1/8$, $\lambda_r(A) \in \mathcal{EG}^*(A,B,\beta)$ holds for all $r = 1,\dots,k_0$, we obtain the following:

(i) *For any $r = 1,\dots,k_0$, there is a real $\ell_r \in \mathcal{EG}(A+B)$ and a real eigenvector $u_r \in \mathcal{EV}(A+B,\ell_r)$ such that*

$$\Bigl|\ell_r - \lambda_r(A) - \operatorname{tr}\Bigl(U\bigl(\lambda_r(A)\bigr)BU\bigl(\lambda_r(A)\bigr)\Bigr)\Bigr| \leq \alpha\bigl(\lambda_r(A)\bigr)_1 \cdot \frac{1}{1-4\beta}, \tag{A.9}$$

$$\|\underline{\gamma}_r(A) - u_r\|_2 \leq \alpha\bigl(\lambda_r(A)\bigr)_2 \cdot \frac{2}{1-4\beta}, \tag{A.10}$$

where with $\lambda := \lambda_r(A)$

$$\alpha(\lambda)_1 := \sup_{q\geq 2}\ \sup_{p_1+\cdots+p_q=q-1} \frac{\bigl|\operatorname{tr}\bigl(B\overline{\mathbf{S}}(\lambda)^{(p_1)}\cdots B\overline{\mathbf{S}}(\lambda)^{(p_q)}\bigr)\bigr|}{\beta^{q-2}}, \tag{A.11}$$

$$\alpha(\lambda)_2 := \sup_{q\geq 1}\ \sup_{p_1+\cdots+p_q=q} \frac{\bigl\|\overline{\mathbf{S}}(\lambda)^{(p_1)}B\cdots\overline{\mathbf{S}}(\lambda)^{(p_q)}BU(\lambda)\bigr\|}{\beta^{q-1}}. \tag{A.12}$$

If, additionally, B is symmetric and if $\mathcal{EG}(A) = \mathcal{EG}^*(A,B,\beta)$ for some $\beta \leq 1/8$, we furthermore obtain the following:

(ii) *For any $r = 1,\dots,k_0$, it holds that $\ell_r = \lambda_r(A+B)$, and*

$$\left|\sum_{r=k_0+1}^{k} \lambda_r(A+B) - (k-k_0)\cdot\mathbf{1} - \operatorname{tr}\bigl(U(\mathbf{1})BU(\mathbf{1})\bigr)\right| \leq \frac{\alpha(\mathbf{1})_1}{1-4\beta}, \tag{A.13}$$

where $\alpha(\mathbf{1})_1$ is defined by (A.11), *with* $\lambda := \mathbf{1}$.

Proof of Lemma 1. The proof of the lemma is based on results in the perturbation theory for finite-dimensional spaces. Consider $\lambda_r(A)$, $r = 1,\dots,k_0$, and let $\eta \in \mathbb{C}$. The conditons of Lemma 1 imply that, for η near 0, $A + \eta\cdot B$ has exactly one eigenvalue $\overline{\ell}_r(\eta)$ near $\lambda_r(A)$. Moreover, $\overline{\ell}_r(\eta)$ is a holomorphic function near $\eta = 0$ [cf. a corollary of Reed and Simon (1978), pages 3–4]. Formulas (2.21) and (2.31) in Chapter II of Kato (1966) yield the corresponding Taylor series. For any $r = 1,\dots,k_0$ we then obtain, abbreviating $\lambda := \lambda_r(A)$,

$$\overline{\ell}_r(\eta) = \ell_r(\eta) := \lambda + \sum_{q=1}^{\infty} \frac{(-\eta)^q}{q} \sum_{p_1+\cdots+p_q=q-1} \operatorname{tr}\bigl(B\overline{\mathbf{S}}(\lambda)^{(p_1)}\cdots B\overline{\mathbf{S}}(\lambda)^{(p_q)}\bigr) \tag{A.14}$$

for $|\eta|$ sufficiently small. A corresponding matrix $\overline{U}(\eta)$, projecting into the eigenspace of $A + \eta B$ for $\overline{\ell}_r(\eta)$, adopts the expansion

$$\begin{aligned}\overline{U}(\eta) = U(\eta,\lambda) &:= U(\lambda) - \sum_{q=1}^{\infty}(-\eta)^q \\ &\quad\times \sum_{p_1+\cdots+p_{q+1}=q} \overline{\mathbf{S}}(\lambda)^{(p_1)}B\overline{\mathbf{S}}(\lambda)^{(p_2)}\cdots\overline{\mathbf{S}}(\lambda)^{(p_q)}B\overline{\mathbf{S}}(\lambda)^{(p_{q+1})},\end{aligned} \tag{A.15}$$

as follows from formulas (1.17) and (2.12) in Kato [(1966), Chapter 2].

The number of solutions of $p_1 + \cdots + p_q = q - 1$ is

$$(2q-2)!/\big((q-1)!\big)^2$$

(recall that $p_i \in \mathbb{N} \cup \{0\}$ for all i). It is easily verified that $(2q-2)!/\,(q\cdot((q-1)!)^2)$ and $(2q-1)!/(q!)^2$ can be bounded by 4^{q-2} and 4^q if $q \geq 2$ and $q \geq 1$. Thus, if $\beta \leq 1/8$, $\lambda \in \mathcal{EG}^*(A,B,\beta)$ implies that the series in (A.14) and (A.15) converge for all $|\eta| < 2$, and that

$$\text{(A.16)} \qquad \|U(\eta,\lambda) - U(\lambda)\| \leq \frac{4|\eta|\beta}{1-4|\eta|\beta},$$

$$\text{(A.17)} \qquad \big|\ell_r(\eta) - \lambda - \eta\,\mathrm{tr}\big(U(\lambda)BU(\lambda)\big)\big| \leq \frac{|\eta|^2\alpha(\lambda)_1}{1-4|\eta|\beta},$$

where $\alpha(\lambda)_1$ is defined (A.11). We obtain

$$\text{(A.18)} \qquad \begin{aligned} \alpha(\lambda)_1 &\leq \sup_{q\geq 2}\ \sup_{p_1+\cdots+p_{q-1}=q-1} \frac{\big\|U(\lambda)B\overline{\mathbf{S}}(\lambda)^{(p_1)}B\cdots\overline{\mathbf{S}}(\lambda)^{(p_{q-1})}BU(\lambda)\big\|}{\beta^{q-2}} \\ &\leq \frac{\beta^2}{\|\mathbf{S}(\lambda)\|}. \end{aligned}$$

To verify relation (A.18) first note that in (A.11) we have $p_i = 0$ for at least one $p_i \in \{p_1,\ldots,p_q\}$. Since $\mathrm{tr}\,(CD) = \mathrm{tr}\,(DC)$, the second supremum in (A.11) is thus equivalent to the supremum over those values of $p_1,\ldots,p_q$ with $p_q = 0$, only. By definition $\overline{\mathbf{S}}(\lambda)^{(0)} = -U(\lambda)$. The matrix $U(\lambda)$ is idempotent with $\mathrm{rank}(U(\lambda)) = 1$. It holds that $\mathrm{tr}\,(DU(\lambda)) = \mathrm{tr}\,(U(\lambda)DU(\lambda))$ and $|\mathrm{tr}\,(U(\lambda)DU(\lambda))| \leq \|U(\lambda)DU(\lambda)\|$, for any real $k \times k$ matrix D, and (A.18) is an immediate consequence.

Relations (A.16) and (A.17) imply that, for *all* $|\eta| < 2$, $\bar{\ell}_r(\eta) := \ell_r(\eta)$ is an eigenvalue of $A + \eta B$, while $U(\eta,\lambda)$ projects into the eigenspace of $A + \eta B$ for $\ell_r(\eta)$. These are consequences of Kato [(1966), Theorems 1.8 and 1.9 of Chapter 2] and of the identity theorem for holomorphic functions. In particular, $\ell_r := \ell_r(1)$ is an eigenvalue of $A + B$, and (A.17) proves assertion (A.9).

To prove assertion (A.10), first note that $u_r := U(1,\lambda)\cdot\underline{\gamma}_r(A)/\|U(1,\lambda)\cdot\underline{\gamma}_r(A)\|_2$ is a real, normalized eigenvector of $A + B$ for ℓ_r. Recall that we abbreviate $\lambda := \lambda_r(A)$. It holds that $\mathbf{S}(\lambda)\underline{\gamma}_r(A) = 0$ and, hence, $\overline{\mathbf{S}}(\lambda)^{(p)}\underline{\gamma}_r(A) = 0$, $p > 0$. Relation (A.15) thus leads to

$$\text{(A.19)} \qquad \begin{aligned} &\big\|U(1,\lambda)\cdot\underline{\gamma}_r(A) - \underline{\gamma}_r(A)\big\|_2 \\ &\quad\leq \sum_{q=1}^{\infty}\ \sum_{p_1+\cdots+p_q=q} \big\|\overline{\mathbf{S}}(\lambda)^{(p_1)}B\overline{\mathbf{S}}(\lambda)^{(p_2)}\cdots\overline{\mathbf{S}}(\lambda)^{(p_q)}BU(\lambda)\big\| \\ &\quad\leq \frac{\alpha(\lambda)_2}{1-4\beta}, \end{aligned}$$

where the last inequality in (A.19) can easily be deduced from the fact that the number of solutions of $p_1 + \cdots + p_q = q$ is $(2q-1)!/(q!(q-1)!) \leq 4^{q-1}$. Trivially,

$\alpha(\lambda)_2 \le \beta$. Using (A.19) and $\alpha(\lambda)_2 \le \beta$, relation (A.10) then follows from the fact that, for all vectors v, w with $\|w\|_2 = 1$,

$$\left\| \frac{v}{\|v\|_2} - w \right\|_2 \le \|v\|_2 \left| \frac{1}{\|v\|_2} - 1 \right| + \|v - w\|_2$$
$$\le \big| \|w\|_2 - \|v\|_2 \big| + \|v - w\|_2 \le 2\|v - w\|_2.$$

Now, consider the second part of the lemma, and suppose that B is symmetric. Theorem 1.10 in Kato [(1966), Chapter 2] implies the existence of a number of functions $\ell_{k_0+1}(\eta), \dots, \ell_k(\eta)$ (some of them might be identical) with the following properties: the $\ell_r(\cdot)$ are holomorphic functions in a region containing the real axis; for any η, $\ell_r(\eta)$ is an eigenvalue of $A + \eta B$; and $\mathbf{l} = \ell_{k_0+1}(0) = \cdots = \ell_k(0)$.

Formulas (2.21) and (2.31) of Kato (1966) provide a Taylor expansion for $\sum_{r=k_0+1}^{k} \ell_r(\eta) - (k - k_0)\mathbf{l}$ which is of the same form as those of $\ell_r(\eta) - \lambda$ in (A.14); just replace λ by $\mathbf{l}$ in the tr$(\cdot)$-terms. Accordingly, since $\mathbf{l} \in \mathcal{EG}^*(A, B, \beta)$, relation (A.17) remains true when replacing $\ell_r(\eta) - \lambda$ by $\sum_{r=k_0+1}^{k} \ell_r(\eta) - (k - k_0)\mathbf{l}$.

The term $\alpha(\mathbf{l})_1$, defined by (A.11), can then be bounded by

$$\alpha(\mathbf{l})_1 \le (k - k_0) \sup_{q \ge 2} \sup_{p_1 + \cdots + p_{q-1} = q-1} \frac{\|U(\mathbf{l}) B \overline{\mathbf{S}}(\mathbf{l})^{(p_1)} B \cdots \overline{\mathbf{S}}(\mathbf{l})^{(p_{q-1})} B U(\mathbf{l})\|}{\beta^{q-2}}$$
$$\le \frac{(k - k_0)\beta^2}{\|\mathbf{S}(\mathbf{l})\|}.$$

We thus can conclude that assertion (ii) of Lemma 1 holds, provided $\ell_r = \lambda_r(A + B)$, $r = 1, \dots, k_0$, and $\sum_{r=k_0+1}^{k} \ell_r(1) = \sum_{r=k_0+1}^{k} \lambda_r(A + B)$.

Since $\beta \le 1/8$, we can infer from (A.17) and (A.18) that, for all $r = 1, \dots, k_0$,

$$|\lambda_r(A) - \ell_r| < \frac{1}{2\|\mathbf{S}(\lambda_r(A))\|} = \frac{1}{2} \min_{r \ne s \in \{1, \dots, k_0+1\}} |\lambda_r(A) - \lambda_s(A)|.$$

Hence, $\ell_1 > \ell_2 > \cdots > \ell_{k_0} > \lambda_{k_0}(A) - (\lambda_{k_0}(A) - \mathbf{l})/2$, and to complete the proof of the lemma it only remains to show that

$$\text{(A.20)} \qquad \max_{r = k_0+1, \dots, k} |\mathbf{l} - \ell_r(1)| \le \frac{1}{2\|\mathbf{S}(\mathbf{l})\|} = \frac{\lambda_{k_0}(A) - \mathbf{l}}{2}.$$

For $\lambda = \mathbf{l}$, let $U(\eta, \lambda)$ be defined by (A.15). Since $\mathbf{l} \in \mathcal{EG}^*(A, B, \beta)$, relation (A.16) carries over to $\lambda = \mathbf{l}$, and the series in (A.15) converges for $|\eta| < 2$. We now can infer from [Kato (1966), Theorem 1.10 and formulas (2.3) and (2.12) in Chapter 2] that we have $\ell_r(1) - \mathbf{l} \in \mathcal{EG}((A + B - \mathbf{l}I) \cdot U(1, \mathbf{l}))$. Hence,

$$\text{(A.21)} \qquad \max_{r = k_0+1, \dots, k} |\mathbf{l} - \ell_r(1)| \le \|(A + B - \mathbf{l} \cdot I) \cdot U(1, \mathbf{l})\|.$$

Formulas (2.16) and (2.18) in Kato [(1966), Chapter 2] provide a series expansion of the right-hand side of (A.21):

$$\max_{r = k_0+1, \dots, k} |\mathbf{l} - \ell_r(1)|$$
$$\le \sum_{q=1}^{\infty} (-\eta)^q \sum_{p_1 + \cdots + p_{q+1} = q-1} \|\overline{\mathbf{S}}(\mathbf{l})^{(p_1)} B \overline{\mathbf{S}}(\mathbf{l})^{(p_2)} \cdots \overline{\mathbf{S}}(\mathbf{l})^{(p_q)} B \overline{\mathbf{S}}(\mathbf{l})^{(p_{q+1})}\|.$$

The number of solutions of $p_1 + \cdots + p_{q+1} = q - 1$ is

$$(2q-1)!/\big(q!(q-1)!\big) \leq 4^{q-1}.$$

Any summand on the right-hand side can be bounded by $\beta^q/\|\mathbf{S(1)}\|$. Consequently, $|\mathbf{1} - \ell_r(1)| \leq \beta/(\|\mathbf{S(1)}\|(1-4\beta))$ for all $r = k_0 + 1, \dots, k$, and $\beta \leq 1/8$ yields (A.20). This completes the proof of Lemma 1. □

The bounds established in Lemma 1 are too general to be feasible for practical computations. There are several different ways to make them more specific. Lemma 1a provides a version which is particularly suited for the proof of Theorem 2.

LEMMA 1a. *Under the conditions of Lemma* 1 *define matrices* $S(\lambda)$ *by replacing* $\ell - \lambda$ *by* $|\ell - \lambda|$ *in* (A.7). *For* $\lambda \in \mathcal{E}\mathcal{G}(A)$, *set*

$$\beta(\lambda) := \max\big\{\|B \cdot S(\lambda)\|, \|S(\lambda)\| \cdot \|B \cdot U(\lambda)\|\big\},$$

and let $\beta := \max_{r=1,\dots,k_0} \beta(\lambda_r(A))$.

Then $\lambda_r(A) \in \mathcal{E}\mathcal{G}^*(A, B, \beta)$, *for all* $r = 1, \dots, k_0$, *and*

$$\alpha\big(\lambda_r(A)\big)_1 \leq \big\|BU\big(\lambda_r(A)\big)\big\| \cdot \big\|U\big(\lambda_r(A)\big)BS\big(\lambda_r(A)\big)\big\|,$$
$$\alpha\big(\lambda_r(A)\big)_2 \leq \big\|S\big(\lambda_r(A)\big)\big\| \cdot \big\|BU\big(\lambda_r(A)\big)\big\|.$$

PROOF. For any two $k \times k$ matrices C and D it holds that

$$\|C\overline{\mathbf{S}}(\lambda)^{(p)}D\| \leq \|CS(\lambda)\|\|S(\lambda)\|^{p-1}\|D\|, \qquad p \geq 1,$$

and

$$\|CU(\lambda)D\| \leq \|CU(\lambda)\|\|U(\lambda)D\| \leq \|CU(\lambda)\|\|D\|.$$

In a straightforward way, such decompositions can be used to establish relation (A.8) for all $q, p_1, \dots, p_q$ and $\lambda \in \{\lambda_1(A), \dots, l_{k_0}(A)\}$. Note that $\|U(\lambda)\| = 1$ and $U(\lambda) \cdot U(\lambda) = U(\lambda)$. Accordingly, the asserted bounds for $\alpha(\lambda)_1$ and $\alpha(\lambda)_2$ follow from (A. 18) and (A.12). □

Lemma 1b establishes the version of Lemma 1 which is particularly suited for the proofs of Theorems 3 and 4 and Propositions 3 and 4.

LEMMA 1b. *In addition to the conditions of Lemma* 1 *suppose that* B *is symmetric. Define matrices* $S(\lambda)$ *by replacing* $\ell - \lambda$ *by* $|\ell - \lambda|$ *in* (A.7). *For* $\lambda \in \mathcal{E}\mathcal{G}$ (A) *set* $s(\lambda) := \|S(\lambda)\|$,

$$\beta(\lambda) := \max\{\|U(\lambda)BU(\lambda)\| \cdot s(\lambda), \|S(\lambda)^{1/2}BU(\lambda)\| \cdot s(\lambda)^{1/2}, \|S(\lambda)^{1/2}BS(\lambda)^{1/2}\|\}$$

and let $\beta := \max_{\lambda \in \mathcal{E}\mathcal{G}(A)} \beta(\lambda)$.

Then $\mathcal{E}\mathcal{G}^*(A, B, \beta) = \mathcal{E}\mathcal{G}(A)$, *and, for any* $\lambda \in \mathcal{E}\mathcal{G}(A)$,

$$\alpha(\lambda)_1 \leq s(\lambda)\,\big|\mathrm{tr}\big(U(\lambda)BU(\lambda)BU(\lambda)\big)\big| + \big|\mathrm{tr}\big(U(\lambda)BS(\lambda)BU(\lambda)\big)\big|. \tag{A.22}$$

Furthermore, for $r = 1,\dots,k_0$, *it holds that*

$$\alpha(\lambda_r(A))_2 \leq s(\lambda_r(A))^{1/2} \cdot \|S(\lambda_r(A))^{1/2} BU(\lambda_r(A))\|.$$

PROOF. For any two $k \times k$ matrices C and D it holds that

$$\|C\overline{\mathbf{S}}(\lambda)^{(p)}D\| \leq \|CS(\lambda)^{1/2}\|\|S(\lambda)\|^{p-1}\|S(\lambda)^{1/2}D\|, \qquad p \geq 1,$$

and

$$\|CU(\lambda)D\| \leq \|CU(\lambda)\|\|U(\lambda)D\|.$$

As is easily seen, such decompositions can be used to establish relation (A.8) for all $q, p_1,\dots,p_q$ and $\lambda \in \mathcal{E}\mathcal{G}(A)$. Accordingly, the asserted bounds for $\alpha(\lambda)_2$ follow from (A.12).

It remains to prove (A.22). Let $\lambda \in \mathcal{E}\mathcal{G}(A)$, and recall the arguments used to verify relation (A.18). We can infer that

$$\alpha(\lambda)_1 = \sup_{q \geq 2} \sup_{p_1 + \cdots + p_{q-1} = q-1} \frac{\left|\operatorname{tr}\left(C(p_1,\dots,p_{q-1};\lambda)\right)\right|}{\beta^{q-2}} \tag{A.23}$$

holds, where

$$C(p_1,\dots,p_q;\lambda) := U(\lambda)B\overline{\mathbf{S}}(\lambda)^{(p_1)}B\cdots\overline{\mathbf{S}}(\lambda)^{(p_q)}BU(\lambda).$$

Now let

$$C_1(\lambda) := U(\lambda)B\left(S(\lambda)^{1/2} + s(\lambda)^{1/2}U(\lambda)\right)$$

and note that, for $q \geq 2$, $C(p_1,\dots,p_q;\lambda) = C_1(\lambda)C_2(p_1,\dots,p_q;\lambda)C_1(\lambda)'$ for a matrix $C_2(p_1,\dots,p_q;\lambda)$ with $\|C_2(p_1,\dots,p_q;\lambda)\| \leq \beta^{q-2}$ [recall that $U(\lambda)S(\lambda) = 0$]. We thus obtain, for $q \geq 2$,

$$\begin{aligned}
&\left|\operatorname{tr}\left(C(p_1,\dots,p_{q-1};\lambda)\right)\right| \\
&\quad = \left|\operatorname{tr}\left(C_1(\lambda)C_2(p_1,\dots,p_{q-1};\lambda)C_1(\lambda)'\right)\right| \\
&\quad = \|C_2(p_1,\dots,p_q;\lambda)\| \\
&\qquad \times \left|\operatorname{tr}\left(C_1(\lambda)C_1(\lambda)'\right) - \operatorname{tr}\left(C_1(\lambda)\left(I - \frac{C_2(p_1,\dots,p_q;\lambda)}{\|C_2(p_1,\dots,p_q;\lambda)\|}\right)C_1(\lambda)'\right)\right|.
\end{aligned}$$

It is easily seen that the matrix $I - C_2(p_1,\dots,p_q;\lambda)/\|C_2(p_1,\dots,p_q;\lambda)\|$ is positive semidefinite. We can conclude that

$$\begin{aligned}
&\left|\operatorname{tr}\left(C(p_1,\dots,p_q;\lambda)\right)\right| \\
&\quad \leq \|C_2(p_1,\dots,p_q;\lambda)\| \cdot \operatorname{tr}\left(C_1(\lambda)C_1(\lambda)'\right) \\
&\quad \leq \beta^{q-2}\Big(s(\lambda)\left|\operatorname{tr}\left(U(\lambda)BU(\lambda)BU(\lambda)\right)\right| + \left|\operatorname{tr}\left(U(\lambda)BS(\lambda)BU(\lambda)\right)\right|\Big).
\end{aligned}$$

The last inequality can immediately be derived from the definitions of $C_1(\cdot)$ and $C_2(\cdot)$. Together with (A.23) this completes the proof of Lemma 1b. □

PROOF OF THEOREM 2. Definition of M implies that $\mathcal{E}\mathcal{G}(M) = \{l_1, \dots, l_{L_0}, 0\}$, where $l_r := (1/N)\sum_{j=1}^{N} n\theta_{jr}^2$. For $r = 1, \dots, L_0$ the projection matrix $U(l_r)$ projecting onto the eigenspace of M for l_r is given by

$$\begin{aligned} U(l_r) &:= \frac{1}{n}\underline{g}_r\underline{g}_{r'}, \\ S(l_r) &:= \sum_{l_r \neq \lambda \in \mathcal{E}\mathcal{G}(M)} \frac{1}{|\lambda - l_r|} U(\lambda) \\ &= \sum_{s \neq r} \frac{1}{|l_s - l_r|}\frac{1}{n}\underline{g}_s\underline{g}_s' + \frac{1}{l_r}U(0). \end{aligned}$$

Now let

$$M_h^* := (W_h - I)M = (W_h - I)\sum_{r=1}^{L_0} l_r \frac{1}{n}\underline{g}_r\underline{g}_r'.$$

In the sequel we will consider the largest L_0 eigenvalues, and corresponding eigenvectors, of $M + M_h^*$. They will prove to be identical to those of M_h. Let us thus analyze the terms relevant to establishing the assertions of Lemma 1a. Based on the assumptions of Theorem 2 we obtain the following, for all $r = 1, \dots, L_0$:

$$\|M_h^* U(l_r)\| = \|(I - W_h)l_r n^{-1/2}\underline{g}_r\|_2 = l_r \,\mathrm{Bias}_{h,r} \le \frac{l_r D_2}{8}; \tag{A.24}$$

$$\begin{aligned} \|M_h^* S(l_r)\| &\le \sum_{s \neq r} \frac{\|(I - W_h)l_s n^{-1/2}\underline{g}_s\|_2}{|l_s - l_r|} \\ &\le \sum_{s \neq r} \frac{l_s}{D_2 l_s}\,\mathrm{Bias}_{h,s} \le \frac{1}{8}; \end{aligned} \tag{A.25}$$

$$\begin{aligned} \|U(l_r)M_h^* S(l_r)\| &\le \sum_{s \neq r} \frac{l_s}{D_2 l_s}\frac{1}{n}|\underline{g}_r'(I - W_h)\underline{g}_s| \\ &\le \sum_{s \neq r} \frac{\mathrm{Bias}_{h,r}\,\mathrm{Bias}_{h,s}}{D_2} \\ &\le \frac{\mathrm{Bias}_{h,r}}{8}; \end{aligned} \tag{A.26}$$

$$\operatorname{tr}\big(U(l_r)M_h^* U(l_r)\big) = -l_r\frac{1}{n}\underline{g}_r'(I - W_h)\underline{g}_r = -l_r \,\mathrm{Bias}_{h,r}^2. \tag{A.27}$$

Necessarily $D_2 < 1$, and, hence, $\|S(l_r)\| \leq 1/(D_2 l_r)$. Therefore, relations (A.24) and (A.25) imply that, for any $r = 1, \dots, L_0$,

$$\beta(l_r) := \max\left\{\|M_h^* \cdot S(l_r)\|, \|S(l_r)\| \cdot \|M_h^* \cdot U(l_r)\|\right\} \leq \tfrac{1}{8}.$$

Lemma 1a thus yields $l_r \in \mathcal{EG}^*(M + M_h^*, 1/8)$, $r = 1, \dots, L_0$. We now can invoke Lemma 1 to analyze differences between the eigenvalues and eigenvectors of M and $M + M_h^*$. Bounds for $\alpha(l_r)_1$ and $\alpha(l_r)_2$ are to be obtained from Lemma 1a, (A.24) and (A.26). Additionally, using (A.27), it follows that for each $r = 1, \dots, L_0$ there exist a real eigenvalue $\ell_r \in \mathcal{EG}(M + M_h^*)$ and a real eigenvector $u_r \in \mathcal{EV}(M + M_h^*, \ell_r)$ such that

$$|l_r - \ell_r| \leq \frac{5}{4} l_r \operatorname{Bias}_{h,r}^2, \tag{A.28}$$

$$\left\|n^{-1/2}\underline{g}_r - u_r\right\|_2 \leq \frac{4}{D_2} \operatorname{Bias}_{h,r}. \tag{A.29}$$

By assumption, $\operatorname{Bias}_{h,r} \leq D_2/8$ and $|l_s - l_r| \geq D_2 l_r, s \neq r$. Hence, (A.28) implies $|\ell_r - l_r| < l_r D_2/2 \leq 1/2 \min_{s \neq r} |l_s - l_r|$. From this we can deduce that $\ell_1 > \ell_2 > \cdots > \ell_{L_0} > 0$.

At this point, recall that $M_h = W_h M W_h$, and note that $\lambda_r(M_h) = (1/N)\Sigma_{j=1}^N n\tilde{\theta}_{jr}^2$ and $\underline{\gamma}_r(M_h) = n^{-1/2}\underline{\tilde{g}}_r$. Obviously,

$$M + M_h^* = M_h + W_h M (I - W_h),$$

and it is easily seen that, for all $r = 1, \dots, L_0$,

$$\big(M_h + W_h M(I - W_h)\big) \cdot n^{-1/2}\underline{\tilde{g}}_r = M_h \cdot n^{-1/2}\underline{\tilde{g}}_r = \lambda_r(M_h).$$

Consequently, both $\ell_1 > \cdots > \ell_{L_0} > 0$ and $\lambda_1(M_h) \geq \cdots \geq \lambda_{L_0}(M_h)$ define collections of eigenvalues of $M_h + W_h M(I - W_h)$, while $u_1, \dots, u_{L_0}$ and $n^{-1/2}\underline{\tilde{g}}_1, \dots, n^{-1/2}\underline{\tilde{g}}_{L_0}$ are corresponding real, normalized eigenvectors. On the other hand, note that $M_h + W_h M(I - W_h) \cdot v = M_h \cdot v$ for $v \in \mathcal{L}(W_h)$, and $M_h + W_h M(I - W_h) \cdot v = W_h M(I - W_h) \cdot v \in \mathcal{L}(W_h)$ for $v \in \mathbb{R}^n \backslash \mathcal{L}(W_h)$. This shows that any real, nonzero eigenvalue of $M_h + W_h M(I - W_h)$ is necessarily an eigenvalue of M_h, too. The total number of such eigenvalues cannot exceed $\operatorname{rank}(M_h) = L_0$. Consequently, $\ell_r = \lambda_r(M_h)$ and $n^{-1/2}\underline{\tilde{g}}_r = u_r$ or $n^{-1/2}\underline{\tilde{g}}_r = -u_r$, $r = 1, \dots, L_0$. Together with (A.28) and (A.29) this completes the proof of Theorem 2. □

In the sequel, let $\tilde{\lambda}_r := (1/N)\Sigma_{j=1}^N n\tilde{\theta}_{jr}^2$, $r = 1, \dots, L_0$, and $\tilde{\lambda}_r := 1$, $r > L_0$.

LEMMA 2. *Under the conditions on L_0, h, n, N imposed in Theorem 3 let $\Lambda [\equiv \Lambda(L_0, h, n, N)]$ denote $h \times h$ diagonal matrices with diagonal entries $\Lambda_{11} > \Lambda_{22} > \cdots > \Lambda_{L_0 L_0} > \Lambda_{L_0+1 L_0+1} = \cdots = \Lambda_{hh} := 1$. Assume that there exists a $D^* > 0$ such that $|\Lambda_{rr} - \Lambda_{ss}| \geq D^* \tilde{\lambda}_r$ holds for all $r, s \in \{1, \dots, L_0 + 1\}, r \neq s$, and all L_0, h, n, N.*

For $L_0, n, N, h \in \mathbb{N}$ let $\Xi [\equiv \Xi(L_0, n, N, h)]$ denote symmetric $h \times h$ random matrices. Let ξ_{rs}, $r, s = 1, \dots, h$, denote the entries of Ξ, and assume that there *exists a sequence δ_N of real numbers with the following properties*: (1) $h \cdot \delta_N \to 0$ *as* $N \to \infty$; (2) $\sup_{r,s=1,\dots,h}(E\xi_{rs}^2/\tilde{\lambda}_r\tilde{\lambda}_s) = O(\delta_N^2)$, $N \to \infty$.

Then the following hold:

$$\text{(i)}\qquad \begin{aligned} &|\lambda_r(\Lambda + \Xi) - \Lambda_{rr}| \\ &\quad = O_P\left(\left(E\xi_{rr}^2\right)^{1/2} + \sum_{s=1}^{h} \frac{E\xi_{rs}^2}{\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}} \right), \quad r = 1,\dots,L_0; \end{aligned}$$

$$\text{(ii)}\qquad \begin{aligned} &\|\underline{\gamma}_r(\Lambda) - \underline{\gamma}_r(\Lambda + \Xi)\|_2 \\ &\quad = O_P\left(\frac{\left(\sum_{s\neq r}\left[E\xi_{rs}^2/\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}\right]\right)^{1/2}}{\tilde{\lambda}_r^{1/2}} \right), \quad r = 1,\dots,L_0; \end{aligned}$$

$$\text{(iii)}\qquad \max_{r=1,\dots,L_0} \frac{|\lambda_r(\Lambda + \Xi) - \Lambda_{rr}|}{\tilde{\lambda}_r} = O_P(h \cdot \delta_N);$$

$$\text{(iv)}\qquad \left| \sum_{r=L_0+1}^{h} \lambda_r(\Lambda + \Xi) - \sum_{r=L_0+1}^{h} (\xi_{rr} + 1) \right| = O_P\left(h(h - L_0)\delta_N^2\right).$$

PROOF. It holds that $\mathcal{EG}(\Lambda) = \{\Lambda_{11},\dots,\Lambda_{L_0L_0}, 1\}$. Furthermore, $\mathcal{EV}(\Lambda_{rr}) = \operatorname{span}\{\underline{e}_r\}$, $r = 1,\dots,L_0$, and $\mathcal{EV}(1) = \operatorname{span}\{\underline{e}_{L_0+1},\dots,\underline{e}_h\}$, where $\underline{e}_1 = (1,0,\dots,0)'$, $\dots,\underline{e}_h = (0,\dots,0,1)'$ denote the Euclidean basis of $\mathbb{R}^h$. The matrix $U(\Lambda_{rr}) := \underline{e}_r\, \underline{e}_r'$ is the projection matrix projecting into the eigenspace of Λ for Λ_{rr}, $r = 1,\dots,L_0$, while

$$U(1) := \sum_{r=L_0+1}^{h} \underline{e}_r\underline{e}_r'$$

is the projection matrix for the eigenvalue 1. For $l \in \mathcal{EG}(\Lambda)$ matrices $S(\ell)$, as required in Lemma 1b, are given by

$$S(l) = \sum_{\ell \in \mathcal{EG}(\Lambda)\setminus\{l\}} \frac{1}{|l - \ell|} U(\ell).$$

By assumption,

$$\text{(A.30)}\qquad \frac{1}{|\Lambda_{rr} - \Lambda_{ss}|} \le \frac{1}{D^* \max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}}$$

holds for all r, s with $\Lambda_{rr} \neq \Lambda_{ss}$. Hence, $s(\Lambda_{rr}) = \|S(\Lambda_{rr})\| \le 1/(D^*\tilde{\lambda}_r)$. Now, note that

$$\text{(A.31)}\qquad \begin{aligned} \operatorname{tr}\left(\underline{e}_r\underline{e}_r'\Xi\,\underline{e}_s\underline{e}_s'\right) &= \xi_{rs} \cdot \delta_{rs}, \\ \operatorname{tr}\left(\underline{e}_q\underline{e}_q'\Xi\,\underline{e}_r\underline{e}_r'\Xi\,\underline{e}_s\underline{e}_s'\right) &= \xi_{rs}^2 \cdot \delta_{qs}, \end{aligned}$$

for all r, s, q, where $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. Furthermore, $\|C\| \leq (\mathrm{tr}(C'C))^{1/2}$ for any real $h \times h$ matrix C. After some straightforward computations this allows us to derive that

$$\beta(\lambda)^2 \leq \beta^* := \sum_{r=1}^{h}\sum_{s=1}^{h} \frac{\xi_{rs}^2}{(D^{*2}\widetilde{\lambda}_r\widetilde{\lambda}_s)}$$

holds for all $\lambda \in \mathcal{E}\mathcal{G}(\Lambda)$, where $\beta(\lambda)$ is defined as in Lemma 1b. This implies

$$\begin{aligned} \beta &:= \max_{l \in \mathcal{E}G(\Lambda)} \beta(l) \leq \beta^{*\frac{1}{2}} = O_P\left(\left(\sum_{r=1}^{h}\sum_{s=1}^{h} \frac{E\xi_{rs}^2}{D^{*2}\widetilde{\lambda}_r\widetilde{\lambda}_s}\right)^{1/2}\right) \\ &= O_P(h\delta_N) = o_P(1). \end{aligned} \tag{A.32}$$

Relation (A.32) now allows us to invoke Lemma 1 in the specialized version established by Lemma 1b. Note that $\mathrm{tr}(U(\Lambda_{rr})\Xi U(\Lambda_{rr})) = \xi_{rr}$, $r \leq L_0$, and $\mathrm{tr}(U(1)\Xi U(1)) = \sum_{r=L_0+1}^{h}\xi_{rr}$. Assertions (i), (ii) and (iv) are immediate consequences of (A.9), (A.10) and (A.13), when using (A.30) and (A.31) to evaluate the bounds for $\alpha(\lambda)_1$ and $\alpha(\lambda)_2$ given in Lemma 1b. Assertion (iii) follows from (A.32), $\alpha(\Lambda_{rr})_1/\widetilde{\lambda}_r \leq \beta^* \cdot D^*$ and $\mathrm{tr}(U(\Lambda_{rr})\Xi U(\Lambda_{rr}))/\widetilde{\lambda}_r = \xi_{rr}/\widetilde{\lambda}_r \leq \beta^{*1/2} \cdot D^*$, $r \leq L_0$. □

PROOF OF THEOREM 3. Recall the arguments used to prove Theorem 1. Using the notation introduced there, it follows from (A.1) that the nonzero eigenvalues of $\widehat{M}_h$ and of $\widehat{\Lambda}$ are identical. Furthermore,

$$n^{-1/2}\underline{\widehat{g}}_r = \Gamma_h \cdot \underline{\gamma}_r(\widehat{\Lambda}), \qquad r = 1, \ldots, L_0. \tag{A.33}$$

Let Λ denote the diagonal matrix with $\Lambda_{rr} = \widetilde{\lambda}_r + 1$, $r = 1, \ldots, L_0$, and $\Lambda_{rr} := 1$, $r > L_0$. It holds that

$$n^{-1/2}\underline{\widetilde{g}}_r = \Gamma_h \cdot \underline{\gamma}_r(\Lambda) = \Gamma_h \cdot \underline{e}_r, \qquad r = 1, \ldots, L_0.$$

Consequently,

$$n^{-1/2}\|\underline{\widetilde{g}}_r - \underline{\widehat{g}}_r\|_2 = \|\Gamma_h\underline{\gamma}_r(\Lambda) - \Gamma_h\underline{\gamma}_r(\widehat{\Lambda})\|_2 = \|\underline{\gamma}_r(\Lambda) - \underline{\gamma}_r(\widehat{\Lambda})\|_2.$$

Hence, in order to prove the theorem it suffices to show that the asserted stochastic bounds apply to $|\Lambda_{rr} - \lambda_r(\widehat{\Lambda})|$, $|\sum_{r=L_0+1}^{h}\lambda_r(\widehat{\Lambda}) - \widetilde{R}_{L_0}|$ and $\|\underline{\gamma}_r(\Lambda) - \underline{\gamma}_r(\widehat{\Lambda})\|_2$ (recall that $\Lambda_{rr} = \widetilde{\lambda}_r + 1$). Let

$$\begin{aligned} \Xi := \widehat{\Lambda} - \Lambda &= \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=1}^{L_0} n^{1/2}\widetilde{\theta}_{jr}\,\underline{e}_r\right)\left(\sum_{r=1}^{h}\bar{\epsilon}_{rj}\underline{e}_r\right)' \\ &\quad + \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=1}^{h}\bar{\epsilon}_{rj}\underline{e}_r\right)\left(\sum_{r=1}^{L_0} n^{1/2}\widetilde{\theta}_{jr}\underline{e}_r\right)' \\ &\quad + \frac{1}{N}\sum_{j=1}^{N}\sum_{r=1}^{h}\sum_{s=1}^{h}\left(\bar{\epsilon}_{rj}\bar{\epsilon}_{sj} - \delta_{rs}\right)\underline{e}_r\underline{e}_s' \\ &:= \Xi^{(1)} + \Xi^{(2)} + \Xi^{(3)}, \end{aligned} \tag{A.34}$$

where $\delta_{rs} = 1$ for $r = s$, and $\delta_{rs} = 0$ otherwise. The random variables $\bar{\epsilon}_{rj}$ and $\bar{\epsilon}_{sk}$ are independent for $j \neq k$, satisfying $E\bar{\epsilon}_{rj} = 0$, $E\bar{\epsilon}_{rj}^2 = 1$ and $E\bar{\epsilon}_{rj}\bar{\epsilon}_{sj} = 0$, $r \neq s$. Similar to (A.6), some easy computations show that there exists a constant $D_{11}, \infty > D_{11} \geq 0$, such that $E\bar{\epsilon}_{rj}^4 \leq D_{11}$. For $r, s \in \{1, \dots, h\}$ this leads to the following:

$$\text{(A.35)} \qquad E\xi_{rs}^{(1)^2} = \frac{\tilde{\lambda}_r}{N}, \qquad r \leq L_0; \qquad E\xi_{rs}^{(1)^2} = 0, \qquad r \geq L_0;$$

$$\text{(A.36)} \qquad E\xi_{rs}^{(2)^2} = \frac{\tilde{\lambda}_s}{N}, \qquad s \leq L_0; \qquad E\xi_{rs}^{(2)^2} = 0, \qquad s \geq L_0;$$

$$\text{(A.37)} \qquad E\xi_{rs}^{(3)^2} \leq \frac{D_{11}}{N},$$

for all r, s. These results might be combined to obtain

$$\text{(A.38)} \qquad \sup_{q, s = 1, \dots, h} \frac{E\xi_{qs}^2}{\tilde{\lambda}_q \tilde{\lambda}_s} = O\left(\frac{1}{N}\right) \quad \text{and} \quad \sup_{s = 1 \dots, h} \frac{E\xi_{rs}^2}{\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}} = O\left(\frac{1}{N}\right),$$

$r = 1, \dots, h$. Let $D^* := \min\{D_4, D_5\}$. Assumption 4 implies that $|\Lambda_{rr} - \Lambda_{ss}| \geq D^* \max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}$ holds for all $r, s \in \{1, \dots, L_0 + 1\}$, $r \neq s$. The assertions of the theorem now are consequences of Lemma 2, when additionally noting that

$$\tilde{R}_{L_0} = \frac{1}{N} \sum_{j=1}^{N} \sum_{r = L_0 + 1}^{h} \bar{\epsilon}_{rj}^2 = \sum_{r = L_0 + 1}^{h} (\xi_{rr} + 1). \qquad \square$$

PROOF OF PROPOSITION 3. Since we do not rely on Assumption 4, we have to consider the eigenvalues $\hat{\lambda}_r$ first. Recall that $\hat{\lambda}_r := \lambda_r(\widehat{M}_h) = \lambda_r(\hat{\Lambda})$, $r = 1, \dots, h$. Let Λ^* denote the diagonal matrix with diagonal entries $\Lambda_{rr}^* = 1 + 0.9^{r-1} \cdot \tilde{\lambda}_r$, $r = 1, \dots, L_0$, and $\Lambda_{rr}^* := 1$, $r > L_0$. Using the notation introduced in the proof of Theorem 3, we obtain $\hat{\Lambda} = \Lambda^* + \Xi + (\Lambda - \Lambda^*)$. Since $\Lambda - \Lambda^*$ is positive semidefinite, it follows from a theorem of Anderson and Dasgupta (1963) that $\lambda_r(\hat{\Lambda}) \geq \lambda_r(\Lambda^* + \Xi)$. Proposition 2 implies that there exists a $d > 0$ such that $\tilde{\lambda}_{L_0} > d$. Hence, $|\Lambda_{rr}^* - \Lambda_{ss}^*| \geq \min\{0.1 \cdot 0.9^{L_0 - 2}, 0.9^{L_0 - 1} \cdot d\} \cdot \tilde{\lambda}_r$ for all $r, s = 1, \dots, L_0 + 1$, $r \neq s$. Since, furthermore, (A.38) generalizes to the present situation, we might invoke Lemma 2 to analyze the eigenvalues of $\Lambda^* + \Xi$ Lemma 2(iii) and $N^{1/2}h^{-1/2} \to \infty$ now lead to

$$\min_{r = 1, \dots, L_0} N^{1/2} h^{-1/2} (\hat{\lambda}_r - 1) \geq \min_{r = 1, \dots, L_0} N^{1/2} h^{-1/2} \big(\lambda_r(\Lambda^* + \Xi) - 1\big) \to \infty$$

in probability. At the same time there obviously exists a constant $C_\alpha^{**} < \infty$ such that

$$N^{-1/2}(h - L)^{-1/2} \big(C_{\alpha, N(h-L)} - N(h - L)\big) \leq C_\alpha^{**}$$

for all N and $h > L$. It follows that

$$\begin{aligned}&P\left(\min_{L=1,\dots,L_0}\left(N\sum_{r=L}^{h}\widehat{\lambda}_r - C_{\alpha,N(h-L)}\right) < 0\right)\\ &\quad\le P\left(\min_{L=1,\dots,L_0}\left(N^{1/2}(h-L)^{-1/2}\sum_{r=L}^{h}(\widehat{\lambda}_r - 1) - C_{\alpha}^{**}\right) < 0\right) \to 0.\end{aligned} \tag{A.39}$$

Relation (A.39) shows that $P(\widehat{L}_0 < L_0) \to 0$, and it only remains to prove that $P(\widehat{L}_0 > L_0) \to \alpha$. We can infer from Lemma 2(iv) and from our assumptions on h, N that

$$\begin{aligned}&N^{1/2}(h-L_0)^{-1/2}\sum_{r=L_0+1}^{h}\widehat{\lambda}_r\\ &\quad\ge N^{1/2}(h-L_0)^{-1/2}\sum_{r=L_0+1}^{h}\lambda_r(\Lambda^* + \Xi)\\ &\quad= N^{1/2}(h-L_0)^{-1/2}\widetilde{R}_{L_0} + o_P(1).\end{aligned} \tag{A.40}$$

Now, however, consider the diagonal matrix Λ^{**} with $\Lambda_{rr}^{**} = 1 + 1.1^{L_0 - r}\cdot\widetilde{\lambda}_r$, $r = 1,\dots,L_0$ and $\Lambda_{rr}^{**} := 1, r > L_0$. We have $\widehat{\Lambda} = \Lambda^{**} + \Xi + (\Lambda - \Lambda^{**})$ and $\lambda_r(\widehat{\Lambda}) \le \lambda_r(\Lambda^{**} + \Xi)$, since $\Lambda - \Lambda^{**}$ is negative semidefinte. Arguments similar to those used above show that we might invoke Lemma 2(iv) to obtain

$$\begin{aligned}N^{1/2}(h-L_0)^{-1/2}\sum_{r=L_0+1}^{h}\widehat{\lambda}_r &\le N^{1/2}(h-L_0)^{-1/2}\sum_{r=L_0+1}^{h}\lambda_r(\Lambda^{**} + \Xi)\\ &= N^{1/2}(h-L_0)^{-1/2}\widetilde{R}_{L_0} + o_P(1).\end{aligned} \tag{A.41}$$

At this point recall the notation introduced in the proof of Theorem 1. It can be inferred from Assumption 1 that there is a constant $D_0^{**} < \infty$ such that $E\bar{\epsilon}_{rj}^8 \le D_0^{**}$ for all r, j. By using standard martingale central limit theorems we thus can deduce from Theorem 1(i) that

$$\begin{aligned}&N^{1/2}(h-L_0)^{-1/2}\widetilde{R}_{L_0} = N^{-1/2}(h-L_0)^{-1/2}\sum_{j=1}^{N}\sum_{r=L_0+1}^{n}\bar{\epsilon}_{rj}^2\\ &\quad\text{is } AN\bigl(N^{1/2}(h-L_0)^{1/2}, 2\bigr).\end{aligned} \tag{A.42}$$

On the other hand, we can obtain from standard results that, for any $\alpha > 0$,

$$\left|N^{-1/2}(h-L_0)^{-1/2}C_{\alpha,N(h-L_0)} - C_{\alpha}^*\right| \to 0$$

as $N \to \infty$, where C_{α}^* denotes the corresponding critical value of a $N(N^{1/2}(h - L_0)^{1/2}, 2)$ distribution. When combining this with (A.40)–(A.42), we can conclude that

$$
\begin{aligned}
&P\left(N \sum_{r=L_0+1}^{h} \widehat{\lambda}_r > C_{\alpha, N(h-L_0)}\right) \\
&\quad = P\left(N^{1/2}(h-L_0)^{-1/2} \sum_{r=L_0+1}^{h} \widehat{\lambda}_r > N^{-1/2}(h-L_0)^{-1/2} C_{\alpha, N(h-L_0)}\right) \to \alpha,
\end{aligned}
$$

which proves that $P(\widehat{L}_0 > L_0) \to \alpha$. □

PROOF OF PROPOSITION 4. Let $L_0^{(h)} \leq L_0$ denote the dimension of the projected model, and recall that $\sum_{j=1}^{N} \widetilde{\theta}_{jr}\widetilde{\theta}_{js} = 0$ if $r \neq s$. The assumptions of the proposition imply that $L_0^{(h)} > L_0^*$ for all n sufficiently large. Some easy computations then show that

$$
\begin{aligned}
\widehat{M}_h = {} & \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=1}^{L_0^*+1} n^{1/2}\widetilde{\theta}_{jr}\underline{\widetilde{g}}_r + W_h\underline{\epsilon}_j\right)\left(\sum_{r=1}^{L_0^*+1} n^{1/2}\widetilde{\theta}_{jr}\underline{\widetilde{g}}_r + W_h\underline{\epsilon}_j\right)' \\
& + \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=L_0^*+2}^{L_0^{(h)}} n^{1/2}\widetilde{\theta}_{jr}\underline{\widetilde{g}}_r\underline{\epsilon}_j' W_h + \sum_{r=L_0^*+2}^{L_0^{(h)}} n^{1/2}\widetilde{\theta}_{jr} W_h\underline{\epsilon}_j\underline{\widetilde{g}}_r'\right) \\
& + \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{r=L_0^*+2}^{L_0^{(h)}} n^{1/2}\widetilde{\theta}_{jr}\underline{\widetilde{g}}_r\right)\left(\sum_{r=L_0^*+2}^{L_0^{(h)}} n^{1/2}\widetilde{\theta}_{jr}\underline{\widetilde{g}}_r\right)' \\
:= {} & \widehat{M}_h^{(1)} + \widehat{M}_h^{(2)} + \widehat{M}_h^{(3)}.
\end{aligned}
$$

Set $\widehat{M}_h^{(2)} = 0$ and $\widehat{M}_h^{(3)} = 0$ if $L_0^*+2 > L_0^{(h)}$. Let Λ^* denote the diagonal matrix with diagonal entries $\Lambda_{rr}^* = 1 + 0.9^{r-1} \cdot \widetilde{\lambda}_r, r = 1, \ldots, L_0^* + 1$, and $\Lambda_{rr}^* := 1, r > L_0^* + 1$. Replacing L_0 by $L_0^* + 1$, define matrices $\Lambda, \widehat{\Lambda}$ and Ξ analogous to those in the proof of Theorem 3. We have $\lambda_r(\widehat{M}_h^{(1)}) = \lambda_r(\widehat{\Lambda})$, for all $r = 1, \ldots, L_0^* + 1$, and $\widehat{\Lambda} = \Lambda^* + \Xi + (\Lambda - \Lambda^*)$. Arguments similar to those used in the proof of Proposition 3 now show that Lemma 2 (iii) leads to

$$
\begin{aligned}
\text{(A.43)} \qquad \min_{r=1,\ldots,L_0^*+1}\left(\lambda_r(\widehat{M}_h^{(1)}) - 1\right) &\geq \min_{r=1,\ldots,L_0^*+1}\left(\lambda_r(\Lambda^* + \Xi) - 1\right) \\
&= 0.9^{L_0^*}\widetilde{\lambda}_{L_0^*+1}\left(1 + o_P(1)\right).
\end{aligned}
$$

Since the normalization implies

$$
\frac{1}{N}\sum_{j=1}^{N} n\widetilde{\theta}_{jr}^2 \leq \frac{1}{N}\sum_{j=1}^{N} n\widetilde{\theta}_{jL_0^*+1}^2 = \widetilde{\lambda}_{L_0^*+1} \quad \text{for } r > L_0^* + 1,
$$

it is easily verified that

$$\text{(A.44)}\qquad \begin{aligned} \max_{r=1,\dots,n} \left|\lambda_r(\widehat{M}_h^{(2)})\right| &\le \left(\operatorname{tr}(\widehat{M}_h^{(2)\prime}\widehat{M}_h^{(2)})\right)^{1/2} \\ &= O_P\left(\frac{h}{N^{1/2}}\widehat{\lambda}_{L_0^*+1}^{1/2}\right) = o_P(\widetilde{\lambda}_{L_0^*+1}^{1/2}) \end{aligned}$$

Note that necessarily $L_0^{(h)} \le h$. Clearly, $\widehat{M}_h^{(3)}$ is positive semidefinite. It follows that $\lambda_r(\widehat{M}_h) \ge \lambda_r(\widehat{M}_h^{(1)}) - \max_{r=1,\dots,n} |\lambda_r(\widehat{M}_h^{(2)})|$. Furthermore, Proposition 2 implies the existence of a $d > 0$ such that $\widetilde{\lambda}_{L_0^*+1} \ge d$ for all h, n, N sufficiently large. This can be combined with (A.43) and (A.44) to obtain that

$$\min_{r=1,\dots,L_0^*+1} N^{1/2}h^{-1/2}(\widehat{\lambda}_r - 1) \to \infty$$

in probability. As an immediate consequence, relation (A.39) generalizes to the present situation; just replace L_0 by $L_0^* + 1$. Hence, $P(\widehat{L}_0 \le L_0^*) \to 0$, which completes the proof of the proposition. □

PROOF OF THEOREM 4. Let $\widehat{\Lambda}^* := \Gamma_h'\widehat{M}_h^*\Gamma_h$. Using the same arguments as in the proof of Theorem 3, we obtain that in order to prove Theorem 4 it suffices to show that $|\Lambda_{rr} - \lambda_r(\widehat{\Lambda}^*)|$, $|\sum_{r=L_0+1}^h \lambda_r(\widehat{\Lambda}^*) - \widetilde{R}_{L_0}|$ and $\|\underline{\gamma}_r(\Lambda) - \underline{\gamma}_r(\widehat{\Lambda}^*)\|_2$ adopt the asserted bounds.

As above, this will be derived by making use of Lemma 2. However, now $\widehat{\Lambda}^* - \Lambda$ contains additional terms. It holds that

$$\Xi := \widehat{\Lambda}^* - \Lambda = \Xi^{(1)} + \Xi^{(2)} + \Xi^{(3)} + \Xi^{(4)},$$

where $\Xi^{(1)}, \Xi^{(2)}$ and $\Xi^{(3)}$ are defined by (A.34), and where

$$\Xi^{(4)} := \frac{1}{N}\sum_{j=1}^N \left(\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - \frac{\sigma_j^2}{\sigma_j^2}\right)\left(\sum_{r=1}^h (n^{1/2}\widetilde{\theta}_{jr} + \bar{\epsilon}_{rj})\underline{e}_r\right)\left(\sum_{r=1}^h (n^{1/2}\widetilde{\theta}_{jr} + \bar{\epsilon}_{rj})\underline{e}_r\right)'.$$

Additionally, we have to analyze the asymptotic behavior of $\xi_{rs}^{(4)}$ for all r, s. Therefore, first note that $\Omega(\underline{\mathbf{Y}}_j) - E\Omega(\underline{\mathbf{Y}}_j) = \sum_{i,k}\omega_{ik}\epsilon_{ij}\epsilon_{ik} - E\sum_{i,k}\omega_{ik}\epsilon_{ij}\epsilon_{ik}$. Based on our assumptions on the $\epsilon_{ij} = \epsilon_{ij}/\sigma_j$, an application of an inequality of Whittle (1960) shows that there exists a constant $D_{12} < \infty$ such that, for any $\alpha \in \{1, 2\}$ and all j, N,

$$\text{(A.45)}\qquad E\big(\Omega(\underline{\mathbf{Y}}_j) - E\Omega(\underline{\mathbf{Y}}_j)\big)^{2\alpha} \le D_{12}\left(\sum_{i,k}\omega_{ik}^2\right)^{\alpha} \le D_{12}D_9^{2\alpha}n^{-\alpha}.$$

Together with condition 2 of Assumption 5(c) this yields $E(\Omega(\underline{\mathbf{Y}}_j) - \sigma_j^2)^{2\alpha} \le D_{13}n^{-\alpha}$, $D_{13} < \infty$, which obviously carries over to $E(\widehat{\sigma}_j^2 - \sigma_j^2)^{2\alpha} \le D_{14}n^{-\alpha}$,

$D_{14} < \infty$. It follows that there is a constant $D_{15} < \infty$ such that, for $\alpha \in \{1, 2\}$ and all j, N,

$$E\left(\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - \frac{\sigma_j^2}{\sigma_j^2}\right)^{2\alpha} = E\left(\frac{\sigma_j^2(\sigma_j^2 - \widehat{\sigma}_j^2)}{\sigma_j^2\widehat{\sigma}_j^2}\right)^{2\alpha} \le E\left(\frac{\sigma_j^2(\sigma_j^2 - \widehat{\sigma}_j^2)}{D_6 D_8}\right)^{2\alpha} \le D_{15} n^{-\alpha}. \tag{A.46}$$

Set $\widetilde{\theta}_{jr} := 0$ for $r = L_0 + 1, \dots, h$. Based on the independence of the error terms for different j, (A.46) leads to

$$\begin{aligned} \operatorname{var}(\xi_{rs}^{(4)}) &\le \frac{1}{N^2} \sum_{j=1}^{N} E\left(\left(\frac{\sigma_j^2}{\widehat{\sigma}_j^2} - \frac{\sigma_j^2}{\sigma_j^2}\right)(n^{1/2}\widetilde{\theta}_{jr} + \overline{\epsilon}_{rj})(n^{1/2}\widetilde{\theta}_{js} + \overline{\epsilon}_{sj})\right)^2 \\ &\le \frac{1}{N^2} \sum_{j=1}^{N} (D_{15} n^{-2})^{1/2} \left(E(n^{1/2}\widetilde{\theta}_{jr} + \overline{\epsilon}_{rj})^4 (n^{1/2}\widetilde{\theta}_{js} + \overline{\epsilon}_{sj})^4\right)^{1/2} \\ &\le D_{16} \frac{\widetilde{\lambda}_r^{1/2}\widetilde{\lambda}_s^{1/2}}{N}, \end{aligned} \tag{A.47}$$

for some $D_{16} < \infty$. This is easily obtained by making use of the Cauchy–Schwarz inequality, and by noting that, by assumption 5(b), $(1/N)\Sigma n^2 \widetilde{\theta}_{jr}^2 \widetilde{\theta}_{js}^2 \le n \cdot D_7^2 \widetilde{\lambda}_r^{1/2} \widetilde{\lambda}_s^{1/2}$. If both $r > L_0$ and $s > L_0$, then $\widetilde{\theta}_{jr} = 0$ and $\widetilde{\theta}_{js} = 0$, and we can derive a sharper bound for the variance:

$$\begin{aligned} \operatorname{var}(\xi_{rs}^{(4)}) &\le \frac{1}{N^2} \sum_{j=1}^{N} (D_{15} n^{-2})^{1/2} (E\,\overline{\epsilon}_{rj}^4 \overline{\epsilon}_{sj}^4)^{1/2} \\ &\le \frac{D_{16}^*}{Nn}, \qquad r, s > L_0, \end{aligned} \tag{A.48}$$

for some $D_{16}^* < \infty$.

It remains to consider $E\xi_{rs}^{(4)}$. Since, by relation (A.45) and by condition 2 of Assumption 5(c), $P(\Omega(\underline{\mathbf{Y}}_j) < D_8) = o(n^{-1})$ uniformly for all j, we can immediately infer from condition 1 of Assumption 5(c) that there is a $D_{17} < \infty$ such that $E(\sigma_j^2 - \widehat{\sigma}_j^2)\overline{\epsilon}_{rj} \le D_{17} n^{-1}$, for all j, N, r. On acount of a Taylor expansion of $1/\widehat{\sigma}_j^2$, this allows us to derive the existence of a constant $D_{18} < \infty$ such that, for all r, s,

$$\begin{aligned} \left|E\xi_{rs}^{(4)}\right| &\le \left|\frac{1}{N} \sum_{j=1}^{N} E \frac{\sigma_j^2(\sigma_j^2 - \widehat{\sigma}_j^2)}{\sigma_j^4} (n^{1/2}\widetilde{\theta}_{jr} + \overline{\epsilon}_{rj})(n^{1/2}\widetilde{\theta}_{js} + \overline{\epsilon}_{sj})\right| \\ &\quad + \frac{1}{N} \sum_{j=1}^{N} \left(E \frac{\sigma_j^2(\sigma_j^2 - \widehat{\sigma}_j^2)^4}{D_8^4 \sigma_j^2}\right)^{1/2} \left(E(n^{1/2}\widetilde{\theta}_{jr} + \overline{\epsilon}_{rj})^2 (n^{1/2}\widetilde{\theta}_{js} + \overline{\epsilon}_{sj})^2\right)^{1/2} \\ &\le D_{18} \frac{\widetilde{\lambda}_r^{1/2}\widetilde{\lambda}_s^{1/2}}{n}. \end{aligned} \tag{A.49}$$

We now might combine (A.35)–(A.37) and (A.47)–(A.49) to obtain

$$|E\xi_{rs}^{(4)}|^2 \le D_{19}\left(\frac{\lambda_r}{N} + \frac{\tilde{\lambda}_s}{N} + \frac{\tilde{\lambda}_r\tilde{\lambda}_s}{n^2}\right),$$

for some $0 < D_{19} < \infty$, and hence

$$\sup_{q,s=1,\dots,h} \frac{E\xi_{qs}^2}{\tilde{\lambda}_q\tilde{\lambda}_s} = O\left(\frac{1}{N} + \frac{1}{n^2}\right) \quad \text{and} \quad \sup_{s=1,\dots,h} \frac{E\xi_{rs}^2}{\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}} = O\left(\frac{1}{N} + \frac{\tilde{\lambda}_r}{n^2}\right),$$

$r = 1, \dots, h$. The assertions of the theorem now are consequences of Lemma 2, when additionally noting that, by (A.48) and (A.49),

$$\sum_{r=L_0+1}^{h} (\xi_{rr} + 1) = \tilde{R}_{L_0} + O_P\left(\sum_{r=L_0+1}^{h} \xi_{rr}^{(4)}\right) = \tilde{R}_{L_0} + O_P\left(\frac{h - L_0}{n + (Nn)^{1/2}}\right).$$

This completes the proof of the theorem. □

REFERENCES

Anderson, T. W. and Dasgupta, S. (1963). Some inequalities on characteristic roots of matrices. *Biometrika* **50** 522–524.

Berkey, C. S. and Kent, R. L. (1983). Longitudinal principal components and nonlinear regression models of early childhood growth. *Annals of Human Biology* **10** 522–536.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–509.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

Deaton, A. (1986). Demand analysis. In *Handbook of Econometrics* (Z. Grilliches and M. D. Intriligator, eds.) **3** 1767–1839. North-Holland, New York.

De Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

Engel, E. (1857). Die Produktions- und Consumtionsverhältnisse des Königreiches Sachsen. [Reprinted in *Bull. Inst. Internat. Statist.* **9** 1–54 (1895).]

Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

Glaser, E. M. and Ruchkin, D. S. (1976). *Principles of Neurobiological Signal Analysis.* Academic, New York.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika* **77** 521–529.

HMSO (1983). Family Expenditure Survey, annual base tapes (1968–1983), Dept. Employment, Statistics Division, Her Majesty's Stationary Office, London. (The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.)

Kato, T. (1966). *Perturbation Theory for Linear Operators.* Springer, New York.

Kneip, A. (1987). Selbstmodellierende nichtlineare Regression. Ph.D. dissertation, Dept. Mathematics, Univ. Heidelberg.

Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *Ann. Statist.* **16** 82–112.

Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica* **59** 711–730.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–327.

MALLOWS, C. L. (1973). Some comments on C_P. *Technometrics* **15** 661–675.

MÖCKS, J. (1986). The effect of latency variations in principal component analysis of event-related potentials. *Psychophysiology* **23** 480–484.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

RAO, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics* **14** 1–17.

REED, M. and SIMON, B. (1978). *Analysis of Operators* **4**. Academic, London.

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1231.

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

WHITTLE, P. (1960). Bounds on the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

INSTITUT DE STATISTIQUE
UNIVERSITÉ CATHOLIQUE DE LOUVAIN
34, VOIE DU ROMAN PAYS
1348 LOUVAIN-LA-NEUVE
BELGIUM