cemmap
centre for microdata methods and practice

Adaptive nonparametric instrumental variables estimation: empirical choice of the regularization parameter

Joel Horowitz

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

ADAPTIVE NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION: EMPIRICAL CHOICE OF THE REGULARIZATION PARAMETER

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208
USA

September 2012

ABSTRACT

In nonparametric instrumental variables estimation, the mapping that identifies the function of interest, g say, is discontinuous and must be regularized (that is, modified) to make consistent estimation possible. The amount of modification is controlled by a regularization parameter. The optimal value of this parameter depends on unknown population characteristics and cannot be calculated in applications. Theoretically justified methods for choosing the regularization parameter empirically in applications are not yet available. This paper presents such a method for use in series estimation, where the regularization parameter is the number of terms in a series approximation to g. The method does not require knowledge of the smoothness of g or of other unknown functions. It adapts to their unknown smoothness. The estimator of g based on the empirically selected regularization parameter converges in probability at a rate that is at least as fast as the asymptotically optimal rate multiplied by $(\log n)^{1/2}$, where n is the sample size. The asymptotic integrated mean-square error (AIMSE) of the estimator is within a specified factor of the optimal AIMSE.

Key words: Ill-posed inverse problem, regularization, sieve estimation, series estimation, nonparametric estimation

JEL Classification: C13, C14, C21

I thank Xiaohong Chen, Joachim Freyberger, Sokbae Lee, and Vladimir Spokoiny for helpful discussions. This research was supported in part by NSF grant SES-0817552.

ADAPTIVE NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION: EMPIRICAL CHOICE OF THE REGULARIZATION PARAMETER

1. INTRODUCTION

This paper is about estimating the unknown function g in the model

(1.1) $$Y = g(X) + U;\ E(U \mid W = w) = 0$$

for almost every w or, equivalently,

(1.2) $$E[Y - g(X) \mid W = w] = 0$$

for almost every w. In this model, g is a function that satisfies regularity conditions but is otherwise unknown, Y is a scalar dependent variable, X is a continuously distributed explanatory variable that may be correlated with U (that is, X may be endogenous), W is a continuously distributed instrument for X, and U is an unobserved random variable. The data are an independent random sample of (Y, X, W). The paper presents a theoretically justified, empirical method for choosing the regularization parameter that is needed for estimation of g.

Existing nonparametric estimators of g in (1.1)-(1.2) can be divided into two main classes: sieve (or series) estimators and kernel estimators. Sieve estimators have been developed by Ai and Chen (2003), Newey and Powell (2003); Blundell, Chen, and Kristensen (2007); and Horowitz (2012). Kernel estimators have been developed by Hall and Horowitz (2005) and Darolles, Fan, Florens, and Renault (2011). Florens and Simoni (2010) describe a quasi-Bayesian estimator based on kernels. Hall and Horowitz (2005) and Chen and Reiss (2011) found the optimal rate of convergence of an estimator of g. Horowitz (2007) gave conditions for asymptotic normality of the estimator of Hall and Horowitz (2005). Horowitz and Lee (2012) showed how to use the sieve estimator of Horowitz (2012) to construct uniform confidence bands for g. Newey, Powell, and Vella (1999) present a control function approach to estimating g in a model that is different from (1.1)-(1.2) but allows endogeneity of X and achieves identification through an instrument. The control function model is non-nested with (1.1)-(1.2) and is not discussed further in this paper. Chernozhukov, Imbens, and Newey (2007); Horowitz and Lee (2007); and Gagliardini, and Scaillet (2012) have developed methods for estimating a quantile-regression version of model (1.1)-(1.2). Chen and Pouzo (2008, 2009) developed a method for estimating a large class of nonparametric and semiparametric conditional moment models with possibly non-smooth moments. This class includes the quantile-regression version of (1.1)-(1.2).

As is explained further in Section 2 of this paper, the relation that identifies g in (1.1)-(1.2) creates an ill-posed inverse problem. That is, the mapping from the population distribution

of (Y, X, W) to g is discontinuous. Consequently, g cannot be estimated consistently by replacing unknown population quantities in the identifying relation with consistent estimators. To achieve a consistent estimator it is necessary to regularize (or modify) the mapping that identifies g. The amount of modification is controlled by a parameter called the regularization parameter. The optimal value of the regularization parameter depends on unknown population characteristics and, therefore, cannot be calculated in applications. Although there have been proposals of informal rules-of-thumb for choosing the regularization parameter in applications, theoretically justified empirical methods are not yet available.

This paper presents an empirical method for choosing the regularization parameter in sieve or series estimation, where the regularization parameter is the number of terms in the series approximation to g. The method consists of optimizing a sample analog of a weighted version of the integrated mean-square error of a series estimator of g. The method does not require *a priori* knowledge of the smoothness of g or of other unknown functions. It adapts to their unknown smoothness. The estimator of g based on the empirically selected regularization parameter also adapts to unknown smoothness. It converges in probability at a rate that is at least as fast as the asymptotically optimal rate multiplied by $(\log n)^{1/2}$, where n is the sample size. Moreover, its asymptotic integrated mean-square error (AIMSE) is within a specified factor of the optimal AIMSE. The paper does not address question of whether the factor of $(\log n)^{1/2}$ can be removed or is an unavoidable price that must be paid for adaptation. This question is left for future research.

Section 2 provides background on the estimation problem and the series estimator that is used with the adaptive estimation procedure. This section also reviews the relevant mathematics and statistics literature. The problems treated in that literature are simpler than (1.1)-(1.2). Section 3 describes the proposed method for selecting the regularization parameter. Section 4 presents the results of Monte Carlo experiments that explore the finite-sample performance of the method. Section 5 presents an empirical example, and Section 6 presents concluding comments. All proofs are in the appendix.

2. BACKGROUND

This section explains the estimation problem and the need for regularization, outlines the sieve estimator that is used with the adaptive estimation procedure, and reviews the statistics literature on selecting the regularization parameter.

2.1 The Estimation Problem and the Need for Regularization

Let X and W be continuously distributed random variables. Assume that the supports of X and W are $[0,1]$. This assumption does not entail a loss of generality, because it can be satisfied by, if necessary, carrying out monotone increasing transformations of X and W. Let f_{XW} and f_W, respectively, denote the probability density functions of (X,W) and W. Define

$$m(w) = E(Y \mid W = w) f_W(w).$$

Let $L_2[0,1]$ be the space of real-valued, square-integrable functions on $[0,1]$. Define the operator A from $L_2[0,1] \to L_2[0,1]$ by

$$(Ah)(w) = \int_{[0,1]} h(x) f_{XW}(x,w) dx,$$

where h is any function in $L_2[0,1]$. Then g in (1.1)-(1.2) satisfies $Ag = m$.

Assume that A is one-to-one, which is necessary for identification of g. Then $g = A^{-1}m$. If f_{XW}^2 is integrable on $[0,1]^2$, then zero is a limit point (and the only limit point) of the singular values of A. Consequently, the singular values of A^{-1} are unbounded, and A^{-1} is a discontinuous operator. This is the ill-posed inverse problem. Because of this problem, g could not be estimated consistently by replacing m in $g = A^{-1}m$ with a consistent estimator, even if A were known. To estimate g consistently, it is necessary to regularize (or modify) A so as to remove the discontinuity of A^{-1}. A variety of regularization methods have been developed. See, for example, Engl, Hanke, and Neubauer (1996); Kress (1999); and Carrasco, Florens, and Renault (2007), among many others. The regularization method used in this paper is series truncation, which is a modification of the Petrov-Galerkin method that is well-known in the theory of integral equations. See, for example, Kress (1999, pp. 240-245). It amounts to approximating A with a finite-dimensional matrix. The singular values of this matrix are bounded away from zero, so the inverse of the approximating matrix is a continuous operator. The details of the method are described further in Section 2.2.

2.2 Sieve Estimation and Regularization by Series Truncation

The adaptive estimation procedure uses a two-stage estimator that is a modified version of Horowitz's (2012) sieve estimator of g. The estimator is defined in terms of series expansions of g, m, and A. Let $\{\psi_j : j = 1,2,...\}$ be a complete, orthonormal basis for $L_2[0,1]$. The expansions are

$$g(x) = \sum_{j=1}^{\infty} b_j \psi_j(x),$$

$$m(w) = \sum_{k=1}^{\infty} m_k \psi_k(w),$$

and

$$f_{XW}(x,w) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} c_{jk} \psi_j(x) \psi_k(w),$$

where

$$b_j = \int_{[0,1]} g(x) \psi_j(x) dx,$$

$$m_k = \int_{[0,1]} m(w) \psi_k(w) dw,$$

and

$$c_{jk} = \int_{[0,1]^2} f_{XW}(x,w) \psi_j(x) \psi_k(w) dx dw.$$

To estimate g, we need estimators of m_k, c_{jk}, m, and f_{XW}. Denote the data by $\{Y_i, X_i, W_i : i = 1,...,n\}$, where n is the sample size. The estimators of m_k and c_{jk}, respectively, are $\hat{m}_k = n^{-1} \sum_{i=1}^{n} Y_i \psi_k(W_i)$ and $\hat{c}_{jk} = n^{-1} \sum_{i=1}^{n} \psi_j(X_i) \psi_k(W_i)$. The estimators of m and f_{XW}, respectively, are $\hat{m}(w) = \sum_{k=1}^{J_n} \hat{m}_k \psi_k(w)$ and $\hat{f}_{XW}(x,w) = \sum_{j=1}^{J_n} \sum_{k=1}^{J_n} \hat{c}_{jk} \psi_j(x) \psi_k(w)$, where J_n is a series truncation point that, for now, is assumed to be non-stochastic. It is assumed that as $n \to \infty$, $J_n \to \infty$ at a rate that is specified in Section 3.1. Section 3.3 describes an empirical method for choosing J_n. Define the operator $\hat{A}$ that estimates A by

$$(\hat{A}h)(w) = \int_{[0,1]} h(x) \hat{f}_{XW}(x,w) dx.$$

The first-stage estimator of g is defined as

(2.2) $$\tilde{g} = \hat{A}^{-1} \hat{m}.$$

To obtain the second-stage estimator that is used with this paper's adaptive estimation procedure, let $\langle \cdot, \cdot \rangle$ denote the inner product in $L_2[0,1]$. For $j = 1,...,J_n$, define $\tilde{b}_j = \langle \tilde{g}, \psi_j \rangle$. The $\tilde{b}_j$'s are the generalized Fourier coefficients of $\tilde{g}$ with the basis functions $\{\psi_j\}$. Let $J \le J_n$ be a positive integer. The second-stage estimator of g is

(2.3) $$\hat{g}_J = \sum_{j=1}^{J} \tilde{b}_j \psi_j.$$

If J is chosen to minimize the AIMSE of $\hat{g}_J$, then $\left\|\hat{g}_J - g\right\|$ converges in probability to 0 at the optimal rate of Chen and Reiss (2011). See Proposition A.1 of the appendix. However, this choice of J is not available in applications because it depends on unknown population parameters. The adaptive estimation procedure consists of choosing J in (2.3) to minimize a sample analog of a weighted version of the AIMSE of $\hat{g}_J$. This procedure is explained in Section 3.1. Let $\hat{J}$ denote the resulting value of J. The adaptive estimator of g is $\hat{g}_{\hat{J}}$. Under the regularity conditions of Section 3.2, $\left\|\hat{g}_{\hat{J}} - g\right\|$ converges in probability to 0 at a rate that is at least as fast as the optimal rate times $(\log n)^{1/2}$. Moreover, the AIMSE of $\hat{g}_{\hat{J}}$ is within a factor of $2+(4/3)\log n$ of the AIMSE of the infeasible estimator that minimizes the AIMSE of $\hat{g}_J$. Achieving these results does not require knowledge of the smoothness of g or the rate of convergence of the singular values of A.

An alternative to the estimator (2.3) consists of using the estimator (2.2) with $\hat{J}$ in place of J_n. However, replacing J_n with $\hat{J}$ causes the lengths of the series in $\hat{A}$ and $\hat{m}$ to be variables of the optimization problem. The methods of proof of this paper do not apply in this case. The asymptotic properties of $\tilde{g}$ with $\hat{J}$ in place of J_n are unknown.

2.3 Review of Related Mathematics and Statistics Literature

Ill-posed inverse problems in models that are similar to but simpler than (1.1)-(1.2) have long been studied in mathematics and statistics. Two important characteristics of (1.1)-(1.2) are that (1) the operator A is unknown and must be estimated from the data and (2) the distribution of $V \equiv Y - E[g(X)|W]$, is unknown. The mathematics and statistics literatures contain no methods for choosing the regularization parameter in (1.1)-(1.2) under these conditions.

A variety of ways to choose regularization parameters are known in mathematics and numerical analysis. Engl, Hanke, and Neubauer (1996), Mathé and Pereverzev (2003), Bauer and Hohage (2005), Wahba (1977), and Lukas (1993, 1998) describe many. Most of these methods assume that A is known and that the "data" are deterministic or that $Var(Y \mid X = x)$ is known and independent of x. Such methods are not suitable for econometric applications.

Spokoiny and Vial (2011) describe a method for choosing the regularization parameter in an estimation problem in which A is known and V is normally distributed. The resulting estimator of g converges at a rate that is within a factor of $(\log n)^p$ of the optimal rate for a

suitable $p > 0$. Loubes and Ludeña (2008) also consider a setting in which A is known. Efromovich and Koltchinskii (2001), Cavalier and Hengartner (2005), Hoffmann and Reiss (2008), and Marteau (2006, 2009) consider settings in which A is known up to a random perturbation and, possibly, a truncation error but is not estimated from the data. Johannes and Schwarz (2010) treat estimation of g when the eigenfunctions of A^*A are known to be trigonometric functions, where A^* is the adjoint of A. Loubes and Marteau (2009) treat estimation of g when the eigenfunctions of A^*A are known but are not necessarily trigonometric functions and the eigenvalues must be estimated from data. Among the settings in the mathematics and statistics literature, this is the closest to the one considered here. Loubes and Marteau obtain non-asymptotic oracle inequalities for the risk of their estimator and show that, for a suitable $p > 1$, their estimator's rate of convergence is within a factor of $(\log n)^p$ of the asymptotically optimal rate. However, the eigenfunctions of A^*A are not known in econometric applications. Section 3 describes a method for selecting J empirically when neither the eigenvalues nor eigenfunctions of A^*A are known. In contrast to Loubes and Marteau (2009), the results of this paper are asymptotic. However, Monte Carlo experiments that are reported in Section 4 indicate that the adaptive procedure works well with samples of practical size. Parts of the proofs in this paper are similar to parts of the proofs of Loubes and Marteau (2009).

3. MAIN RESULTS

This section begins with an informal description of the method for choosing J. Section 3.2 presents the formal results.

3.1 Description of the Method for Choosing J

Define $E_A(\cdot)$ as the mean of the leading term of the asymptotic expansion of the random variable $(\cdot)$. Specifically, if $Z_n = \tilde{Z}_n + r_n$, where Z_n, $\tilde{Z}_n$, and r_n are random variables, $E(\tilde{Z}_n)$ exists, and $r_n = o_p(\tilde{Z}_n)$ as $n \to \infty$, then $E_A(Z_n) = E(\tilde{Z}_n)$. Define the asymptotically optimal J as the value that minimizes the asymptotic integrated mean-square error (AIMSE) of $\hat{g}_J$ as an estimator of g. The AIMSE is $E_A\|\hat{g}_J - g\|^2$. Denote the asymptotically optimal value of J by J_{opt}. It is shown in Proposition A.1 of the appendix that under the regularity conditions of Section 3.2, $E_A\left\|\hat{g}_{J_{opt}} - g\right\|^2$ converges to zero at the fastest possible rate (Chen and Reiss

2011). However, $E_A\|\hat{g}_J - g\|^2$ depends on unknown population parameters, so it cannot be minimized in applications. We replace the unknown parameters with sample analogs, thereby obtaining a feasible estimator of a weighted version of $E_A\|\hat{g}_J - g\|^2$. Let $\hat{J}$ denote the value of J that minimizes the feasible estimator. Note that $\hat{J}$ is a random variable. The adaptive estimator of g is $\hat{g}_{\hat{J}}$. The AIMSE of the adaptive estimator is $E_A\|\hat{g}_{\hat{J}} - g\|^2$. Under the regularity conditions of Section 3.2,

(3.1) $$E_A\|\hat{g}_{\hat{J}} - g\|^2 \le [2 + (4/3)\log(n)]E_A\|\hat{g}_{J_{opt}} - g\|^2.$$

Thus, the rate of convergence in probability of $\|\hat{g}_{\hat{J}} - g\|$ is within a factor of $(\log n)^{1/2}$ of the optimal rate. Moreover, the AIMSE of $\hat{g}_{\hat{J}}$ is within a factor of $2 + (4/3)\log n$ of the AIMSE of the infeasible optimal estimator $\hat{g}_{J_{opt}}$.

The following notation is used in addition to that already defined. For any positive integer J, let A_J be the operator on $L_2[0,1]$ that is defined by

$$(A_J h)(w) = \int_{[0,1]} h(x)a_J(x,w)dx,$$

where

$$a_J(x,w) = \sum_{j=1}^{J}\sum_{k=1}^{J} c_{jk}\psi_j(x)\psi_k(w).$$

Let J_n be the series truncation point defined in Section 2.2. For any $x \in [0,1]$, define

$$\delta_n(x,Y,X,W) = [Y - g_{J_n}(X)]\sum_{k=1}^{J_n}\psi_k(W)(A_{J_n}^{-1}\psi_k)(x),$$

$$S_n(x) = n^{-1}\sum_{i=1}^{n}\delta_n(x,Y_i,X_i,W_i),$$

and

$$g_J = \sum_{j=1}^{J} b_j\psi_j.$$

The following proposition is proved in the appendix.

Proposition 1: Let assumptions 1-6 of Section 3.2 hold. Then, as $n \to \infty$,

$$\hat{g}_J - g_J = \sum_{j=1}^{J}\langle S_n, \psi_j\rangle\psi_j + r_n,$$

where

$$\|r_n\| = o_p\left(\left\|\sum_{j=1}^{J}\langle S_n, \psi_j\rangle \psi_j\right\|\right)$$

for all $J \le J_n$. ■

Now for any $J \le J_n$,

$$E_A\|\hat{g}_J - g\|^2 = E_A\|\hat{g}_J - g_J\|^2 + \|g_J - g\|^2$$

$$= E_A\|\hat{g}_J - g_J\|^2 + \|g\|^2 - \|g_J\|^2.$$

Therefore, it follows from Proposition 1 that

$$E_A\|\hat{g}_J - g\|^2 = E_A\sum_{j=1}^{J}\langle S_n, \psi_j\rangle^2 + \|g\|^2 - \|g_J\|^2.$$

Define

$$T_n(J) = E_A\sum_{j=1}^{J}\langle S_n, \psi_j\rangle^2 - \|g_J\|^2.$$

Assume that $J_n \ge J_{opt}$. Then because $\|g\|^2$ does not depend on J,

$$J_{opt} = \arg\min_{J>0} T_n(J).$$

We now put $T_n(J)$ into an equivalent form that is more convenient for the analysis that follows. Observe that

$$(A_{J_n}^{-1}\psi_k)(x) = \sum_{j=1}^{J_n} c^{jk}\psi_j(x),$$

where c^{jk} is the (j,k) element of the inverse of the $J_n \times J_n$ matrix $[c_{jk}]$. This inverse exists under the assumptions of Section 3.2. Therefore,

$$\sum_{k=1}^{J_n}\psi_k(W)(A_{J_n}^{-1}\psi_k)(x) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} c^{jk}\psi_k(W)\psi_j(x)$$

$$= \sum_{j=1}^{J_n}\psi_j(x)[(A_{J_n}^{-1})^*\psi_j](W),$$

where $*$ denotes the adjoint operator. It follows that

$$\delta_n(x, Y, X, W) = [Y - g_{J_n}(X)]\sum_{j=1}^{J_n}\psi_j(x)[(A_{J_n}^{-1})^*\psi_j](W)$$

and

$$\left\langle \delta_n(\cdot, Y, X, W), \psi_j \right\rangle = [Y - g_{J_n}(X)][(A_{J_n}^{-1})^* \psi_j](W).$$

Therefore,

$$T_n(J) = E_A \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^* \psi_j](W_i) \right\}^2 - \|g_J\|^2.$$

It follows from lemma 3 of the appendix and the assumptions of Section 3.2 that

$$E_A \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^* \psi_j](W_i) \right\}^2$$

$$= n^{-1} E_A \sum_{j=1}^{J} [Y - g_{J_n}(X)]^2 \{[(A_{J_n}^{-1})^* \psi_j](W)\}^2.$$

Therefore,

$$T_n(J) = n^{-1} E_A \sum_{j=1}^{J} [Y - g_{J_n}(X)]^2 \{[(A_{J_n}^{-1})^* \psi_j](W)\}^2 - \|g_J\|^2.$$

This is the desired form of $T_n(J)$.

$T_n(J)$ depends on the unknown parameters g_{J_n} and A_{J_n} and on the operator E_A. Therefore, $T_n(J)$ must be replaced by an estimator for use in applications. One possibility is to replace g_{J_n}, A_{J_n}, g_J, and E_A with $\tilde{g}$, $\hat{A}$, $\hat{g}_J$, and the empirical expectation, respectively. This gives the estimator

$$\breve{T}_n(J) \equiv n^{-2} \sum_{i=1}^{n} \left\{ [Y_i - \tilde{g}(X_i)]^2 \sum_{j=1}^{J} \{(\hat{A}^{-1})^* \psi_j](W_i)\}^2 \right\} - \|\hat{g}_J\|^2.$$

However, $\breve{T}_n$ is unsatisfactory for two reasons. First, it does not account for the effect on $E_A \|\hat{g}_{\hat{J}} - g\|^2$ of the randomness of $\hat{J}$. This randomness is the source of the factor of $\log n$ on the right-hand side of (3.1). Second, some algebra shows that

(3.2) $$\|\hat{g}_J\|^2 - \|g_J\|^2 = \|\hat{g}_J - g_J\|^2 + 2\langle g_J, \hat{g}_J - g_J \rangle.$$

The right-hand side of (3.2) is asymptotically non-negligible, so the estimator of T_n must compensate for its effect.

It is shown in the appendix that these problems can be overcome by using the estimator

$$(3.3) \qquad \hat{T}_n(J) \equiv (2/3)(\log n)n^{-2}\sum_{i=1}^{n}\left\{[Y_i - \tilde{g}(X_i)]^2\sum_{j=1}^{J}\{(\hat{A}^{-1})^*\psi_j](W_i)\}^2\right\} - \left\|\hat{g}_J\right\|^2.$$

We obtain $\hat{J}$ by solving the problem

$$(3.4) \qquad \underset{J:1\le J\le J_n}{\text{minimize}}: \hat{T}_n(J),$$

where J_n is the truncation parameter used to obtain $\tilde{g}$, and J_n satisfies assumption 6 of Section 3.2. Section 3.2 gives conditions under which $\hat{g}_{\hat{J}}$ satisfies inequality (3.1) The problem of choosing J_n in applications is discussed in Section 3.3.

3.2 Formal Results

This section begins with the assumptions under which $\hat{g}_{\hat{J}}$ is shown to satisfy (3.1). A theorem that states the result formally follows the presentation of the assumptions.

Let A^* denote the adjoint operator of A. Define $U = Y - g(X)$. For each positive integer J and any positive, increasing sequence $\{\nu_j : j = 1,2,...\}$, define the set of functions

$$\mathcal{H}_{J\nu} = \left\{h \in L_2[0,1]: \left\|h - \sum_{j=1}^{J}\left\langle h,\psi_j\right\rangle\psi_j\right\| \le \nu_J^{-1}\right\}.$$

For each positive integer J, define the set of functions

$$\mathcal{K}_J = \left\{h = \sum_{j=1}^{J}h_j\psi_j : \sum_{j=1}^{J}h_j^2 = 1\right\}.$$

and the scalar parameter

$$\rho_J = \sup_{\nu\in\mathcal{K}_J}\left[\left\|(A^*A)^{1/2}\nu\right\|\right]^{-1}.$$

The parameter ρ_J^{-2} is the inverse of the smallest eigenvalue of the $J\times J$ matrix whose (j,k) element is $\sum_{\ell=1}^{\infty}c_{j\ell}c_{k\ell}$. In addition, ρ_J is a generalization of the sieve measure of ill-posedness defined by Blundell, Chen, and Kristensen (2007)[1].

The assumptions are as follows.

---

[1] Blundell, Chen, and Kristensen (2007) define the sieve measure of ill-posedness as $\sup_{k\in\mathcal{H}_J:\|h\|=1}\left[\left\|(A^*A)^{1/2}h\right\|\right]^{-1}$, where $\mathcal{H}_J \subset \mathcal{K}_J$ is a Sobolev space. The sieve measure of ill-posedness and ρ_J are the same if the eigenvectors of $\sum_{\ell=1}^{\infty}c_{j\ell}c_{k\ell}$ ($j,k = 1,...,J$) are in $\mathcal{H}_J$.

Assumption 1: (i) The supports of X and W are $[0,1]$. (ii) (X,W) has a bounded probability density function f_{XW} with respect to Lebesgue measure. The probability density function of W, f_W, is non-zero almost everywhere on $[0,1]$.

Assumption 2: (i) There is a finite constant C_Y such that $E(Y^2 | W = w) \leq C_Y$ for each $w \in [0,1]$. (ii) There are finite constants $C_U > 0$, $c_{U1} > 0$, and $c_{U2} > 0$ such that $E(|U|^j | W = w) \leq C_U^{j-2} j! E(U^2 | W = w)$ and $c_{U1} \leq E(U^2 | W = w) \leq c_{U2}$ for every integer $j \geq 2$ and $w \in [0,1]$.

Assumption 3: (i) (1.1) has a solution $g \in L_2[0,1]$. (ii) The estimators $\tilde{g}$ and $\hat{g}_J$ are as defined in (2.2)-(2.3).

Assumption 4: The operator A is one-to-one.

Assumption 5: (i) The basis functions $\{\psi_j\}$ are orthonormal, complete on $L_2[0,1]$. (ii) There is a non-decreasing, positive sequence $\{\nu_j : j = 1,2,...\}$ such that $j^{-s}\nu_j$ is bounded away from 0 for all j and some $s > 3$, and $g \in \mathcal{H}_{J\nu}$. If ν_j increases exponentially fast as j increases, then $j^{-s}\nu_j \to \infty$ for any finite s. (iii) There are constants C_ψ and τ with $0 < C_\psi < \infty$ and $0 \leq \tau < (s-3)/2$ such that $\sup_{0 \leq x \leq 1} |\psi_j(x)| \leq C_\psi j^\tau$ for each $j = 1,2,...$. (iv) There are constants $\alpha > 1/2$, $\varepsilon > 0$, $C < \infty$ and D with $\sum_{j=1}^{\infty} j^{2\alpha} b_j^2 < D < \infty$ such that for all $\delta \in (1/2, \alpha)$,

$$\sup_{h \in \mathcal{H}_{J\delta D}} \frac{\|A_J - A)h\|}{\|h\|} \leq CJ^{-1-\varepsilon}\rho_J^{-1} \text{ and } \frac{\|(A_J - A)(g_J - g)\|}{\|g_J - g\|} \leq C\rho_J^{-1},$$

where $\mathcal{H}_{J\delta D} = \left\{h = \sum_{j=1}^{J} h_j \psi_j : \sum_{j=1}^{J} j^{2\delta} h_j^2 \leq D\right\}$.

Assumption 6: As $n \to \infty$, (i) $\rho_{J_n}(J_n^3/n)^{1/2} \to 0$, (ii) $\rho_{J_n}(J_n^4/n)^{1/2} \to \infty$, and (iii) $J_n^{1+4\tau} / \rho_{J_n}^2 \to 0$.

Assumptions 1 and 2 are smoothness and boundedness conditions. Assumption 3 defines the model and estimators of g. Assumption 4 is required for identification of g. Assumption 5 specifies properties of the basis functions $\{\psi_j\}$. Assumption 5(ii) specifies the accuracy of a truncated series approximation to g and is similar to assumption 2.1 of Chen and Reiss (2011). Assumption 5(ii) is satisfied with $\nu_j \propto j^s$ if g has s square-integrable derivatives and the basis functions belong to a class that includes trigonometric functions, Legendre polynomials that are

shifted and scaled to be orthonormal on $[0,1]$, and B-splines that have been orthonormalized by, say, the Gram-Schmidt procedure. Assumption 5(iii) is satisfied by trigonometric functions ($\tau = 0$), shifted and scaled Legendre polynomials ($\tau = 1/2$), and orthonormalized cubic B-splines ($\tau = 3/2$) if s is large enough. Legendre polynomials require $s > 4$, and cubic B-splines require $s > 6$. Assumption 5(iv) requires the basis functions to be such that a truncated series approximation to A is sufficiently accurate. Assumption 5(iv) restricts the magnitudes of the off-diagonal elements of the infinite dimensional matrix whose (j,k) component is c_{jk} and is analogous to the diagonality restrictions of Hall and Horowitz (2005). Assumption 6 restricts the rate of increase of J_n as $n \to \infty$ and further restricts the size of τ. Assumption 5 implies that the function m satisfies

$$\left\| m - \sum_{j=1}^{J} m_j \psi_j \right\| \leq C_1 J^{-1-\varepsilon} \rho_J^{-1}$$

for some constants $C_1 < \infty$ and $\varepsilon > 0$. See lemma 4 in the appendix for a proof.

For sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, define $a_n \asymp b_n$ if a_n / b_n is bounded away from 0 and ∞ as $n \to \infty$. The problem of estimating g is said to be mildly ill-posed if $\rho_J \asymp J^r$ for some finite $r > 0$ and severely ill posed if $\rho_J \asymp e^{\beta J}$ for some finite $\beta > 0$. Suppose that $\nu_j \asymp j^s$ and $s < \infty$. Then in the mildly ill-posed case, $J_{opt} \propto n^{1/(2r+2s+1)}$ and $\left\| \hat{g}_{J_{opt}} - g \right\| = O_p[n^{-s/(2r+2s+1)}]$ (Blundell, Chen, and Kristensen 2007; Chen and Reiss 2011). In the severely ill-posed case, $J_{opt} = O(\log n)$ and $\left\| \hat{g}_{J_{opt}} - g \right\| = O_p[(\log n)^{-s}]$. Rates approaching the parametric rate $O_p(n^{-1/2})$ are possible if $s = \infty$ but depend on the details of ρ_J and the ν_j's. The results of this section hold in the mildly and severely ill-posed cases and for finite and infinite values of s.

We now have the following theorem.

Theorem 3.1: Let assumptions 1-6 hold. Then $\hat{g}_{\hat{J}}$ satisfies inequality (3.1). ■

3.3 Choosing J_n in Applications

J_n in problem (3.4) depends on ρ_J and, therefore, is not known in applications. This section describes a way to choose J_n empirically. It is shown that inequality (3.1) holds with the empirically selected J_n. The method for choosing J_n has two parts. The first part consists of

specifying a value for J_n that satisfies assumption 6. This value depends on the unknown quantity ρ_J. The second step consists of replacing ρ_J with an estimator.[2]

To take the first step, define J_{n0} by

(3.6) $$J_{n0} = \arg \min_{J=1,2,...} \{\rho_J^2 J^{3.5}/n : \rho_J^2 J^{3.5}/n - 1 \geq 0\} .$$

Because $\rho_J^2 J^{3.5}/n$ is an increasing function of J, J_{n0} is the smallest integer for which $\rho_J^2 J^{3.5}/n \geq 1$. For example, if $\rho_J = J^{\beta}$ for some $\beta > 0$, then J_{n0} is the integer that is closest to and at least as large as $n^{1/(3.5+2\beta)}$. If $\rho_J = e^{\beta J}$ for some $\beta > 0$, then $J_{n0} = O(\log n)$.

J_{n0} satisfies assumption 6 if τ is not too large but is not feasible because it depends on ρ_J. We obtain a feasible estimator of J_{n0} by replacing ρ_J^2 in (3.6) with an estimator. To specify the estimator, let $\hat{A}_J$ ($J = 1, 2, ...$) be the operator on $L_2[0,1]$ whose kernel is

$$\hat{a}_J(x,w) = \sum_{j=1}^{J}\sum_{k=1}^{J} \hat{c}_{jk}\psi_j(x)\psi_k(w); \quad x, w \in [0,1].$$

The estimator of ρ_J^2 is denoted by $\hat{\rho}_J^2$ and is defined by

$$\hat{\rho}_J^{-2} = \inf_{h \in \mathcal{K}_J} \left\| \hat{A}_J^* \hat{A}_J h \right\| .$$

Thus, $\hat{\rho}_J^{-2}$ is the smallest eigenvalue of $\hat{A}_J^* \hat{A}_J$. The estimator of J_{n0} is

$$\hat{J}_{n0} = \arg \min_{J=1,2,...} \{\hat{\rho}_J^2 J^{3.5}/n : \hat{\rho}_J^2 J^{3.5}/n - 1 \geq 0\} .$$

The main result of this paper is given by the following theorem.

Theorem 3.2: Let assumptions 1-6 hold. Assume that either $\rho_J \asymp J^{\beta}$ (mildly ill-posed case) or $\rho_J \asymp e^{\beta J}$ (severely ill-posed case) for some finite $\beta > 0$. Then (i) $P(\hat{J}_{n0} = J_{n0}) \to 1$ as $n \to \infty$. (ii) Let $\hat{\hat{g}}_{\hat{J}}$ be the estimator of g that is obtained by replacing J_n with $\hat{J}_{n0}$ in (3.4). Then

$$E_A \left\| \hat{\hat{g}}_{\hat{J}} - g \right\|^2 \leq [2 + (4/3)\log(n)] E_A \left\| \hat{g}_{J_{opt}} - g \right\|^2 . \quad \blacksquare$$

---

[2] Blundell, Chen, and Kristensen (2007) proposed estimating the rate of increase of ρ_J as J increases by regressing an estimator of $\log(\rho_J)$ on $\log J$ or J for the mildly and severely ill-posed cases, respectively. Blundell, Chen, and Kristensen (2007) did not explain how to use this result to select a specific value of J for use in estimation of g.

Thus the estimator $\hat{\hat{g}}_{\hat{J}}$ that is based on the estimator $\hat{J}_{n0}$ satisfies the same inequality as the estimator $\hat{g}_{\hat{J}}$ that is based on a non-stochastic but infeasible J_n.[3]

4. MONTE CARLO EXPERIMENTS

This section describes the results of a Monte Carlo study of the finite-sample performance of $\hat{g}_{\hat{J}}$. There are 1000 Monte Carlo replications in each experiment. The basis functions $\{\psi_j\}$ are Legendre polynomials that are centered and scaled to be orthonormal on $[0,1]$. J_n is chosen using the empirical method that is described in Section 3.3.

The Monte Carlo experiments use two different designs. There are 5 experiments with Design 1 and 2 experiments with Design 2. In Design 1, the sample size is 1000. Realizations of (X,W) were generated from the model

$$(4.1) \qquad f_{XW}(x,w) = 1 + 2\sum_{j=1}^{\infty} c_j \cos(j\pi x)\cos(j\pi w),$$

where $c_j = 0.7\,j^{-1}$ in experiment 1, $c_j = 0.6\,j^{-2}$ in experiment 2, $c_j = 0.52\,j^{-4}$ in experiment 3, $c_j = 1.3\exp(-0.5j)$ in experiment 4, and $c_j = 2\exp(-1.5j)$ in experiment 5. In all experiments, the marginal distributions of X and W are $U[0,1]$, and the conditional distributions are unimodal with an arch-like shape. The estimation problem is mildly ill-posed in experiments 1-3 and severely ill-posed in experiments 4-5.

The function g is

$$(4.2) \qquad g(x) = b_0 + \sqrt{2}\sum_{j=1}^{\infty} b_j \cos(j\pi x),$$

where $b_0 = 0.5$ and $b_j = j^{-4}$ for $j \geq 1$. This function is plotted in Figure 1. The series in (4.1) and (4.2) were truncated at $j = 100$ for computational purposes. Realizations of Y were generated from

$$Y = E[g(x)|W] + V,$$

where $V \sim N(0, 0.01)$.

---

[3] It is possible that the efficiency of $\hat{g}$ can be improved by re-estimating its Fourier coefficients using a series length of $\hat{J}$ instead of truncating the series of length $\hat{J}_{n0}$ that is $\tilde{g}$. We do not investigate this possibility here. The resulting estimator, like $\hat{\hat{g}}_{\hat{J}}$, would have an AIMSE that is within a factor of $\log n$ of the optimal AIMSE.

Monte Carlo Design 2 mimics estimation of an Engel curve from the data used in the empirical example of Section 5. The data consist of household-level observations from the British Family Expenditure Survey (FES), which is a popular data source for studying consumer behavior. See Blundell, Pashardes, and Weber (1993), for example. We use a subsample of 1516 married couples with one or two children and an employed head of household. In these data, X denotes the logarithm of total income, and W denotes the logarithm of annual income from wages and salaries of the head of household. Blundell, Chen, and Kristensen (2007) and Blundell and Horowitz (2007) discuss the validity of W as an instrument for X. In the Monte Carlo experiment, the dependent variable Y is generated as described below.

The experiment mimics repeated sampling from the population that generated the FES data. The data $\{X_i, W_i : i = 1,...,1516\}$ were transformed to be in $[0,1]^2$ by using the transformations

$$X_i \to (X_i - \min_{1\le j\le 1516} X_j)/(\max_{1\le j\le 1516} X_j - \min_{1\le j\le 1516} X_j)$$

and

$$W_i \to (W_i - \min_{1\le j\le 1516} W_j)/(\max_{1\le j\le 1516} W_j - \min_{1\le j\le 1516} W_j).$$

The transformed data were kernel smoothed using the kernel $K(v) = (15/16)(1-v^2)^2 I(|v| \le 1)$ to produce a density $f_{FES}(x, w; \sigma)$, where σ is the bandwidth parameter. The bandwidths are $\sigma = 0.05$ and 0.10, respectively, in experiments 6 and 7. This range contains the cross-validation estimate, $\sigma = 0.07$, of the optimal bandwidth for estimating the density of the FES data. Numerical evaluation of ρ_J showed that $\rho_J \propto e^{\beta J}$ with $\beta = 3.1$ when $\sigma = 0.05$ and $\beta = 3.4$ when $\sigma = 0.10$. Thus, the estimation problem is severely ill-posed in the Design 2 experiments. The experiments use $g(x) = \Phi[(x - 0.5)/0.24]$, which mimics the share of household expenditures on some good or service. The dependent variable Y is

(4.3) $\quad Y = E[g(X)|W] + V$,

where (X, W) is randomly distributed with the density $f_{FES}(x, w; \sigma)$ and $V \sim N(0, 0.01)$ independently of (X, W).

Each experiment consisted repeating the following procedure 1000 times. A sample of $n = 1516$ observations of (X, Z) was generated by independent random sampling from the distribution whose density is $f_{FES}(x, w; \sigma)$. Then 1516 corresponding observations of Y were generated from (4.3). Finally, g was estimated using the methods described in this paper.

The results of the experiments are displayed in Table 1, which shows the empirical means of $\left\|\hat{g}_{J_{opt}} - g\right\|^2$ and $\left\|\hat{\hat{g}}_{\hat{J}} - g\right\|^2$, the ratio

$$R \equiv \frac{Empirical\ mean\ of\ \left\|\hat{\hat{g}}_{\hat{J}} - g\right\|^2}{Empirical\ mean\ of\ \left\|\hat{g}_{J_{opt}} - g\right\|^2},$$

and $B \equiv 2 + (4/3)\log n$, which is the theoretical asymptotic upper bound on R from inequality (3.1). The results show that the differences between the empirical means of $\left\|\hat{g}_{J_{opt}} - g\right\|^2$ and $\left\|\hat{\hat{g}}_{\hat{J}} - g\right\|^2$ are small and that the ratio R is well within the theoretical bound B in all of the experiments. In experiments 3 and 7, $R = 1$ because $\hat{J} = J_{opt}$ in all Monte Carlo replications.

5. AN EMPIRICAL EXAMPLE

This section presents the estimate of an Engel curve for food that is obtained by applying the methods of Section 3 to the FES data described in Section 4. As in Section 4, the basis functions are Legendre polynomials that are shifted and scaled to be orthonormal on $[0,1]$. J_n is chosen using the empirical method of Section 3.3, and $\hat{J}$ is chosen by solving problem (3.4) after replacing J_n with $\hat{J}_{n0}$. Computation of $\hat{\rho}_J$ showed that $\hat{\rho}_J \propto e^{3.3J}$, so the estimation problem is severely ill posed.

The estimated Engel curve is shown in Figure 2. It is nearly linear. This may be surprising, but a test of the hypothesis that the true Engel curve is linear against a nonparametric alternative (Horowitz 2006) does not reject the hypothesis of linearity ($p > 0.1$). Similarly, in parametric instrumental variables estimation under the assumption that g is a polynomial function of X, it is not possible to reject the hypothesis that the coefficients of terms of degree higher that one are zero. The result of the specification test of Horowitz (2006) implies that a 90% confidence region centered on the true Engel curve would contain a linear function.[4]

[4] The hypotheses that Engel curves for services and other goods are linear also cannot be rejected. The inability to reject linearity of several Engel curves is likely due to the severe ill-posedness of the estimation problem, which prevents estimation of the curves with sufficient accuracy to discriminate between linearity and nonlinearity.

6. CONCLUSIONS

This paper has presented a theoretically justified, empirical method for choosing the regularization parameter in nonparametric instrumental variables estimation. The method does not require *a priori* knowledge of smoothness or other unknown population parameters. The method and the resulting estimator of the unknown function g adapt to the unknown smoothness of g and the density of (X,W). The results of Monte Carlo experiments indicate that the method performs well with samples of practical size.

It is likely that the ideas in this paper can be applied to the multivariate model $Y = g(X,Z)+U$, $E(U|W,Z)=0$, where Z is a continuously distributed, exogenous explanatory variable or vector and W is an instrument for the endogenous variable X. This model is more difficult than (1.1)-(1.2), because it requires selecting at least two regularization parameters, one for X and one or more for the components of Z. The multivariate model will be addressed in future research.

APPENDIX

This appendix presents proofs of Proposition 1 and A.1, Theorems 3.1 and 3.2, and several lemmas that are used in the proofs of the propositions and theorems. Assumptions 1-6 hold throughout. Define $\mathcal{J}_n = \{J : 1 \le J \le J_n\}$. For $J \in \mathcal{J}_n$, define

$$S_n(J) \equiv n^{-1} E[Y - g_{J_n}(X)]^2 \sum_{j=1}^{J} \{(A_{J_n}^{-1})^* \psi_j](W)\}^2 ,$$

$$\tilde{S}_n(J) \equiv n^{-2} \sum_{i=1}^{n} \left\{ [Y_i - g_{J_n}(X_i)]^2 \sum_{j=1}^{J} \{(A_{J_n}^{-1})^* \psi_j](W_i)\}^2 \right\},$$

and

$$\hat{S}_n(J) = n^{-2} \sum_{i=1}^{n} \left\{ [Y_i - \tilde{g}(X_i)]^2 \sum_{j=1}^{J} \{(\hat{A}^{-1})^* \psi_j](W_i)\}^2 \right\}.$$

Define $U_i = Y_i - g(X_i)$ $(i = 1,...,n)$.

We begin with five lemmas that are used in the proof of Proposition 1. Then Propositions 1 and A.1 are proved. Four additional lemmas that are used in the proof of Theorem 3.1 are presented after the proofs of the propositions. Finally, Theorems 3.1 and 3.2 are proved.

Lemma 1: Let $J \le J_n$. Then

(A.1) $$\sup_{h\in\mathcal{K}_J}\left\|A_{J_n}^{-1}h\right\| = \rho_J$$

and

(A.2) $$\sup_{h\in\mathcal{K}_J}\left\|(A_{J_n}^{-1})^* h\right\| = \rho_J \,.$$

Proof: Only (A.1) is proved. The proof of (A.2) is similar. Let I denote the identity operator in $\mathcal{K}_{J_n}$. Note that $\mathcal{K}_J \subset \mathcal{K}_{J_n}$. For $h \in \mathcal{K}_J$,

$$(A^{-1} - A_{J_n}^{-1})h = -[I + A_{J_n}^{-1}(A - A_{J_n})]^{-1} A_{J_n}^{-1}(A - A_{J_n})A_{J_n}^{-1}h \,.$$

But $A_{J_n}^{-1}(A - A_{J_n})A_{J_n}^{-1}h = 0$ for $h \in \mathcal{K}_J$. Therefore,

(A.3) $$\left\|A_{J_n}^{-1}h\right\| = \left\|A^{-1}h\right\| .$$

Now,

$$\rho_J = \sup_{h\in\mathcal{K}_J}\left\|A^{-1}h\right\| ,$$

so it follows from (A.3) that

$$\rho_J = \sup_{h\in\mathcal{K}_J}\left\|A_{J_n}^{-1}h\right\| .$$

Q.E.D.

Lemma 2: The following hold as $n \to \infty$:

(A.4) $$\sup_{h\in\mathcal{H}_{J\delta D},\, \|h\|=1}\left\|(\hat{A} - A_{J_n})h\right\| = J_n^{1/2} O_p(n^{-1/2})$$

and

(A.5) $$\sup_{h\in\mathcal{H}_{J\delta D},\, \|h\|=1}\left\|(\hat{A} - A_{J_n})^* h\right\| = J_n^{1/2} O_p(n^{-1/2}) \,.$$

Proof: Only (A.4) is proved. The proof of (A.5) is similar. Let $h_j = \langle \psi_j, h \rangle$. For any $h \in \mathcal{H}_{J\delta D}$,

(A.6) $$\left\|(\hat{A}-A_{J_n})h\right\|^2 = \sum_{k=1}^{J_n}\left[\sum_{j=1}^{J_n} h_j(\hat{c}_{jk}-c_{jk})\right]^2$$

$$= \sum_{k=1}^{J_n}\left[\sum_{j=1}^{J_n} (j^{\delta}h_j)\left(\frac{\hat{c}_{jk}-c_{jk}}{j^{\delta}}\right)\right]^2 \quad ,$$

$$\leq \sum_{k=1}^{J_n}\left(\sum_{j=1}^{J_n} j^{2\delta}h_j^2\right)\sum_{j}^{J_n}\frac{(\hat{c}_{jk}-c_{jk})^2}{j^{2\delta}},$$

where the last line follows from the Cauchy-Schwarz inequality. But $\sum_{j=1}^{\infty} j^{2\delta}h_j^2 < D$. Therefore,

$$\left\|(\hat{A}-A_{J_n})h\right\|^2 \leq D\sum_{k=1}^{J_n}\sum_{j=1}^{J_n}\frac{(\hat{c}_{jk}-c_{jk})^2}{j^{2\delta}}$$

for any $h \in \mathcal{H}_{J\delta D}$. In addition, $E(\hat{c}_{jk}-c_{jk})^2 \leq Cn^{-1}$ for some constant $C<\infty$, every (j,k) and every sequence of Fourier coefficients $\{c_{jk}\}$. Therefore $\hat{c}_{jk}-c_{jk} = O_p(n^{-1/2})$ for every (j,k) and sequence $\{c_{jk}\}$, where . $O_p(n^{-1/2})$ does not depend on the sequence. It follows that,

$$\sup_{h\in\mathcal{H}_{J\delta D}}\left\|(\hat{A}-A_{J_n})h\right\|^2 \leq O_p(n^{-1})\sum_{k=1}^{J_n}\sum_{j=1}^{J_n} j^{-2\delta}$$

$$= J_n O_p(n^{-1})\sum_{j=1}^{J_n}\frac{1}{j^{2\delta}}$$

$$= J_n O_p(n^{-1}).$$

Q.E.D.

Lemma 3: As $n\to\infty$,

$$\sum_{j=1}^{J}\left\langle S_n,\psi_j\right\rangle^2 = \sum_{j=1}^{J}\left\{n^{-1}\sum_{i=1}^{n}U_i[(A_{J_n}^{-1})^*\psi_j](W_i)\right\}^2 + r_n,$$

where $|r_n| = b_n\rho_J^2/n$, $b_n = o_p(1)$, and b_n does not depend on J. Moreover, there are finite constants M_1 and M_2 such that

$$M_1\rho_J^2/n \le E\sum_{j=1}^{J}\left\{n^{-1}\sum_{i=1}^{n}U_i[(A_{J_n}^{-1})^*\psi_j](W_i)\right\}^2 \le M_2(\rho_J^2 J/n)$$

for every $J \in \mathcal{J}_n$.

Proof: We have

$$S_n(x) = n^{-1}\sum_{i-1}^{n}\delta_n(x, Y_i, X_i, W_i)$$

$$= \sum_{j=1}^{J_n}\psi_j(x)\left\{n^{-1}\sum_{i=1}^{n}[Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^*\psi_j](W_i)\right\}.$$

Therefore,

$$\sum_{j=1}^{J}\left\langle S_n, \psi_j\right\rangle^2 = \sum_{j=1}^{J}\left\{n^{-1}\sum_{i=1}^{n}[Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^*\psi_j](W_i)\right\}^2.$$

But

$$n^{-1}\sum_{i=1}^{n}[Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^*\psi_j](W_i) = R_{nj1} + R_{nj2},$$

where

$$R_{nj1} = n^{-1}\sum_{i=1}^{n}U_i[(A_{J_n}^{-1})^*\psi_j](W_i)$$

and

$$R_{nj2} = -n^{-1}\sum_{i=1}^{n}[g_{J_n}(X_i) - g(X_i)][(A_{J_n}^{-1})^*\psi_j](W_i).$$

$E(R_{nj1}) = 0$, and

$$\sum_{j=1}^{J}Var(R_{nj1}) = n^{-1}\sum_{j=1}^{J}E\sigma_U^2(W)\{[(A_J^{-1})^*\psi_j](W)]^2\},$$

where $\sigma_U^2(w) = E(U^2 \mid W = w)$. But σ_U^2 and f_W are bounded from above Therefore, use of lemma 1 yields

$$\sum_{j=1}^{J}Var(R_{nj1}) \le M_{\ 2}n^{-1}J\rho_J^2$$

for some finite constant M_2. Similarly, σ_U^2 and f_W are bounded. away from 0, so

$$\sum_{j=1}^{J} Var(R_{nj1}) \geq M_1 n^{-1} \rho_J^2 .$$

for finite constant $M_1 > 0$. Therefore,

(A.6) $$M_1 n^{-1} \rho_J^2 \leq \sum_{j=1}^{J} Var(R_{nj1}) \leq M_2 n^{-1} J \rho_J^2$$

for every $j \leq J$. In addition,

$$E(R_{nj2}) = -\int_{[0,1]^2} f_{XW}(x,w)[g_{J_n}(x) - g(x)][(A_{J_n}^{-1})^* \psi_j](w)dxdw$$

$$= -\left\langle A(g_{J_n} - g), (A_{J_n}^{-1})^* \psi_j \right\rangle$$

$$= -\left\langle (A - A_{J_n})(g_{J_n} - g), (A_{J_n}^{-1})^* \psi_j \right\rangle .$$

Therefore, the Cauchy-Schwarz inequality gives

$$\left| E(R_{nj2}) \right| \leq \left\| (A - A_{J_n})(g_{J_n} - g) \right\| \left\| (A_{J_n}^{-1})^* \psi_j \right\| .$$

But $\left\| (A - A_{J_n})(g_n - g) \right\| = O(\rho_{J_n}^{-1} J_n^{-s})$ for some $s > 3$ by assumption 5, and $\left\| (A_{J_n}^{-1})^* \psi_j \right\| \leq \rho_J$ by lemma 1. Therefore, $|E(R_{nj2})| = O(\rho_J \rho_{J_n}^{-1} J_n^{-s}) = o(\rho_J / n^{1/2})$ for every $J \leq J_n$. Also $Var(R_{nj2}) \leq J_n^{-2s} \rho_J^2 / n$. It follows from Markov's inequality and assumption 6 that for some positive sequence $\{b_n\}$ with $b_n \to 0$ as $n \to \infty$,

(A.7) $$\sum_{j=1}^{J} R_{nj2}^2 \leq b_n \rho_J^2 / n$$

for every $j \leq J$ with probability arbitrarily close to 1. The lemma follows by combining (A.6) and (A.7). Q.E.D.

Lemma 4: There are constants $C_1 < \infty$ and $\varepsilon > 0$ such that

$$\left\| m - \sum_{j=1}^{J} m_j \psi_j \right\| \leq C_1 J^{-1-\varepsilon} \rho_J^{-1} .$$

Proof: Define $m_J = \sum_{k=1}^{J} m_k \psi_k$. Then

$$m - m_J = \sum_{k=J+1}^{\infty} \sum_{j=1}^{\infty} b_j c_{jk} \psi_k .$$

But

$$(A - A_J)g = \sum_{k=J+1}^{\infty}\sum_{j=1}^{\infty} b_j c_{jk}\psi_k + \sum_{k=1}^{J}\sum_{j=J+1}^{\infty} b_j c_{jk}\psi_k$$

$$= (m - m_J) + \sum_{k=1}^{J}\sum_{j=J+1}^{\infty} b_j c_{jk}\psi_k .$$

Therefore, it follows from the triangle inequality that

$$\|m - m_J\| \leq \|(A - A_J)g\| + \left\| \sum_{k=1}^{J}\sum_{j=J+1}^{\infty} b_j c_{jk}\psi_k \right\| .$$

But

$$\left\| \sum_{k=1}^{J}\sum_{j=J+1}^{\infty} b_j c_{jk}\psi_k \right\|^2 = \sum_{k=1}^{J}\left(\sum_{j=J=1}^{\infty} b_j c_{jk} \right)^2$$

$$\leq \sum_{k=1}^{\infty}\left(\sum_{j=J=1}^{\infty} b_j c_{jk} \right)^2 = \|A(g_n - g)\|^2 .$$

Therefore,

$$\|m - m_J\| \leq \|(A - A_J)g\| + \|A(g_n - g)\|$$

$$= \|(A - A_J)g_n\| + 2\|(A - A_J)(g_n - g)\|.$$

The lemma now follows from assumption 5. Q.E.D.

Lemma 5: Let $\varepsilon > 0$ be as in assumption 5(iii). Then $J_n^{1+\varepsilon/2}\|\tilde{g} - g\| = o_p(1)$.

Proof: Let δ be such that $1/2 < \delta < 1/2 + \varepsilon/2$ and assumption 5(iii) holds. Define $\hat{h} = \arg\min_{h \in \mathcal{H}_{J_n \delta D}} \|\hat{A}h - \hat{m}\|$.

We show that $J_n^{1+\varepsilon/2}\|\hat{h} - g\| = o_p(1)$. We further show that this implies that with probability approaching 1 as $n \to \infty$, the constraint $h \in \mathcal{H}_{J_n \delta D}$ is not binding. Therefore, $\hat{h} = \tilde{g}$ with probability approaching 1. It follows that $J_n^{1+\varepsilon/2}\|\tilde{g} - g\| = o_p(1)$.

By assumption 5 and the triangle inequality

(A.10) $$\|\hat{h} - g\| \leq \|\hat{h} - g_n\| + \|g_n - g\|$$

$$= \|\hat{h} - g_n\| + O(\nu_J^{-1}).$$

Now

$$\left\|\hat{h}-g_n\right\| \le \rho_{J_n}\left\|A(\hat{h}-g_n)\right\|$$

$$= \rho_{J_n}\left\|(\hat{A}\hat{h}-\hat{m})-(\hat{A}-A)\hat{h}+(\hat{m}-m)-A(g_n-g)-(Ag-m)\right\|$$

$$= \rho_{J_n}\left\|(\hat{A}\hat{h}-\hat{m})-(\hat{A}-A)\hat{h}+(\hat{m}-m)-A(g_n-g)\right\|$$

$$\le \rho_{J_n}\left\|(\hat{A}\hat{h}-\hat{m})\right\|+\rho_{J_n}\left\|(\hat{A}-A)\hat{h}\right\|+\rho_{J_n}\left\|\hat{m}-m\right\|+\rho_{J_n}\left\|A(g_n-g)\right\|.$$

Now

$$\left\|(\hat{A}-A)\hat{h}\right\| \le \left\|(\hat{A}-A_{J_n})\hat{h}\right\|+\left\|(A_{J_n}-A)\hat{h}\right\|$$

$$= O_p(J_n/n)^{1/2}+O(J_n^{-1-\varepsilon}\rho_{J_n}^{-1})$$

by lemma 2 and assumption 5. In addition, standard arguments combined with lemma 4 show that $\left\|\hat{m}-m\right\| = O_p[(J_n/n)^{1/2}]+O(J_n^{-1-\varepsilon}\rho_{J_1}^{-1})$. Therefore,

$$\left\|\hat{h}-g_n\right\| \le \rho_{J_n}\left\|\hat{A}\hat{h}-\hat{m}\right\|+\rho_{J_n}\left\|A(g_n-g)\right\|+O_p[\rho_{J_n}(J_n/n)^{1/2}]+O(J_n^{-1-\varepsilon}).$$

Now assumption 5 implies that

$$\left\|A(g_n-g)\right\| = \left\|(A-A_{J_n})(g_n-g)\right\| = O(\nu_{J_n}^{-1}\rho_{J_n}^{-1}).$$

Therefore,

$$\left\|\hat{h}-g_n\right\| \le \rho_{J_n}\left\|\hat{A}\hat{h}-\hat{m}\right\|+O_p[\rho_{J_n}(J_n/n)^{1/2}]+O(J_n^{-1-\varepsilon}).$$

Now

$$\left\|\hat{A}\hat{h}-\hat{m}\right\| \le \left\|\hat{A}g-\hat{m}\right\|$$

$$= \left\|(\hat{A}-A)g+(Ag-m)-(\hat{m}-m)\right\|$$

$$\le \left\|(\hat{A}-A)g\right\|+\left\|\hat{m}-m\right\|$$

$$= O_p[(J_n/n)^{1/2}]+O(J_n^{-1-\varepsilon}\rho_{J_n}^{-1}).$$

Therefore,

$$\left\|\hat{h}-g_n\right\| = O_p[\rho_{J_n}(J_n/n)^{1/2}]+O(J_n^{-1-\varepsilon})$$

and

$$\left\|\hat{h} - g\right\| = O_p[\rho_{J_n}(J_n / n)^{1/2}] + O(J_n^{-1-\varepsilon}) + O(\nu_{J_n}^{-1}).$$

It follows by combining this result with assumption 5 and 6 that

(A.9) $$J_n^{1+\varepsilon/2}\left\|\hat{h} - g\right\| = o_p(1).$$

We now show that the constraint $\hat{h} \in \mathcal{H}_{J_n \delta D}$ does not bind. Let $\hat{h}_j = \left\langle \hat{h}, \psi_j \right\rangle$ ($j = 1, ..., J_n$). Then

$$\hat{h} = \sum_{j=1}^{J_n} \hat{h}_j \psi_j .$$

We show that as $n \to \infty$,

(A.10) $$P\left(\sum_{j=1}^{J_n} j^{2\delta} \hat{h}_j^2 < D\right) \to 1.$$

To do this, observe that

$$\sum_{j=1}^{J_n} j^{2\delta} \hat{h}_j^2 = \sum_{j=1}^{J_n} j^{2\delta}[b_j + (\hat{h}_j - b_j)]^2 .$$

It follows from Minkowski's inequality and assumption 5 that

$$\left\{\sum_{j=1}^{J_n} j^{2\delta}[b_j + (\hat{h}_j - b_j)]^2\right\}^{1/2} \leq \left(\sum_{j=1}^{J_n} j^{2\delta} b_j^2\right)^{1/2} + \left[\sum_{j=1}^{J_n} j^{2\delta}(\hat{h}_j - b_j)^2\right]^{1/2} .$$

But

$$\left(\sum_{j=1}^{J_n} j^{2\delta} b_j^2\right)^{1/2} + \left[\sum_{j=1}^{J_n} j^{2\delta}(\hat{h}_j - b_j)^2\right]^{1/2} < D + J_n^{\delta}\left\|\hat{h} - g\right\|$$

$$= D + J_n^{-1-\varepsilon/2+\delta} o_p(1)$$

$$= D + o_p(1).$$

It follows that as $n \to \infty$,

(A.11) $$P(\tilde{g} = \hat{h}) \to 1.$$

Q.E.D.

Proof of Proposition 1: We have

$$A_{J_n}\tilde{g} + (\hat{A} - A_{J_n})\tilde{g} = \hat{m}.$$

Therefore

$$\tilde{g} = A_{J_n}^{-1}\hat{m} - A_{J_n}^{-1}(\hat{A} - A_{J_n})\tilde{g}$$

$$= A_{J_n}^{-1}\hat{m} - A_{J_n}^{-1}(\hat{A} - A_{J_n})g_{J_n} - A_{J_n}^{-1}(\hat{A} - A_{J_n})(\tilde{g} - g_{J_n}).$$

It follows that

$$\tilde{g} - g_{J_n} = A_{J_n}^{-1}\hat{m} - A_{J_n}^{-1}\hat{A}g_{J_n} - R_n,$$

where $R_n = A_{J_n}^{-1}(\hat{A} - A_{J_n})(\tilde{g} - g_{J_n})$. Some algebra shows that $A_{J_n}^{-1}\hat{m} - A_{J_n}^{-1}\hat{A}g_{J_n} = S_n$. Therefore,

$$\hat{g}_J - g_J = \sum_{j=1}^{J}\left\langle S_n, \psi_j \right\rangle \psi_j - \sum_{j=1}^{J}\left\langle R_n, \psi_j \right\rangle \psi_j ,$$

and

$$r_n = -\sum_{j=1}^{J}\left\langle R_n, \psi_j \right\rangle \psi_j$$

Now

$$\left\langle \psi_j, R_n \right\rangle = \left\langle \psi_j, A_{J_n}^{-1}(\hat{A} - A_{J_n})(\tilde{g} - g_{J_n}) \right\rangle$$

$$= \left\langle (A_{J_n}^{-1})^* \psi_j, (\hat{A} - A_{J_n})(\tilde{g} - g_{J_n}) \right\rangle.$$

The Cauchy-Schwarz inequality gives

$$\left\| \left\langle \psi_j, R_n \right\rangle \right\| \le \left\| (A_{J_n}^{-1})^* \psi_j \right\| \left\| (\hat{A} - A_{J_n})(\tilde{g} - g_{J_n}) \right\|.$$

Therefore,

$$\left\| \left\langle \psi_j, R_n \right\rangle \right\| = \rho_J O_p[(J_n / n)^{1/2}] \left\| \tilde{g} - g_{J_n} \right\|$$

by lemmas 1 and 2 and (A.11). It follows that

$$\|r_n\| = \left(\sum_{j=1}^{J} \left\langle R_n, \psi_j \right\rangle^2 \right)^{1/2}$$

$$= J^{1/2} \rho_J O_p(n^{-1/2}) J_n^{1/2} \left\| \tilde{g} - g_{J_n} \right\|$$

for every $J \in \mathcal{J}_n$. But $J^{1/2} J_n^{1/2} \left\| \tilde{g} - g_{J_n} \right\| = o_p(1)$ for every $J \in \mathcal{J}_n$ by lemma 5. Therefore,

$$\|r_n\| = \rho_J O_p(n^{-1/2}) o_p(1).$$

The proposition follows by combining this result with lemma 3. Q.E.D.

Proposition A.1: For any $J \le J_n$, $\left\| \hat{g}_J - g \right\|^2 = O_p(\rho_J^2 J / n + \nu_J^{-2})$.

Proof: It follows from Proposition 1 that for or any $J \le J_n$,

(A.12) $E_A\|\hat{g}_J - g_J\|^2 = E\sum_{j=1}^{J}\langle S_n, \psi_j\rangle^2$.

Applying Lemma 3 to the right-hand side of (A.12) yields

$$E_A\|\hat{g}_J - g_J\|^2 = E\sum_{j=1}^{J}\langle S_n, \psi_j\rangle^2 \le M_2\rho_J^2 J/n$$

In addition,

$$E_A\|\hat{g}_J - g\|^2 = E_A\|\hat{g}_J - g_J\|^2 + \|g_J - g\|^2.$$

By assumption 4, $\|g_J - g\|^2 \le \nu_J^{-2}$. Therefore,

$$E_A\|\hat{g}_J - g\|^2 = O(\rho_J^2 J/n + \nu_J^{-2}).$$

Choosing J to optimize this rate gives Chen's and Reiss's (2011) minimax optimal rate for functions in $\mathcal{H}_{J\nu}$. The conclusion of the lemma follows from Markov's inequality. Q.E.D.

Lemma 6: Given any $\varepsilon > 0$,

$$\left|\frac{\tilde{S}_n(J) - S_n(J)}{S_n(J)}\right| \le \varepsilon$$

for each $J \in \mathcal{J}_n$ with probability approaching 1 as $n \to \infty$.

Proof: Define

$$\tilde{S}_{n1}(J) = n^{-2}\sum_{i=1}^{n}U_i^2\sum_{j=1}^{J}\{[(A_{J_n}^{-1})^*\psi_j](W_i)\}^2,$$

$$\tilde{S}_{n2}(J) = -2n^{-2}\sum_{i=1}^{n}U_i[g_{J_n}(X_i) - g(X_i)]\sum_{j=1}^{J}\{[(A_{J_n}^{-1})^*\psi_j](W_i)\}^2,$$

and

$$\tilde{S}_{n3}(J) = n^{-2}\sum_{i=1}^{n}[g_{J_n}(X_i) - g(X_i)]^2\sum_{j=1}^{J}\{[(A_{J_n}^{-1})^*\psi_j](W_i)\}^2.$$

Then

$$\tilde{S}_n(J) = \tilde{S}_{n1}(J) + \tilde{S}_{n2}(J) + \tilde{S}_{n3}(J).$$

Consider, first, convergence of $\tilde{S}_{n1}(J)/S_n(J)$. By Lemma 1,

$$\left\|(A_{J_n}^{-1})^*\psi_j\right\|^2 \le \rho_J^2$$

for every $j \le J$, $J \in \mathcal{J}_n$. This result together with assumption 5(iii) implies that $\{[(A_{J_n}^{-1})^* \psi_j](w)\}^2 \le c_1 J_n^{1+2\tau} \rho_J^2$ for some constant $c_1 < \infty$ and every $j \le J$ and $J \in \mathcal{J}_n$. Define

$$K_j(w) = \{[(A_{J_n}^{-1})^* \psi_j](w)\}^2 .$$

Then

$$\tilde{S}_{n1}(J) = n^{-2} \sum_{i=1}^{n} U_i^2 \sum_{j=1}^{J} K_j(W_i)$$

for every $J \in \mathcal{J}_n$. Moreover,

$$\bar{K}_{nJ}(w) \equiv \left(\frac{\rho_J^2}{n} \right)^{-1} n^{-1} \sum_{j=1}^{J} K_j(w) \le c_1 J_n^{1+2\tau} .$$

Now let $a_n = n^d$ for some constant $d > 0$ such $n^{1-2d} / J_n^{4+4\tau} \to \infty$ as $n \to \infty$. Such a d exists under assumption 6. Let B_n denote the event $\max_{1 \le i \le n} U_i^2 \le a_n$. Let $\bar{B}_n$ denote the complement of B_n. It follows from Markov's inequality that $P(\bar{B}_n) \to 0$ as $n \to \infty$. We have

$$\left(\frac{\rho_J^2}{n} \right)^{-1} \tilde{S}_{n1}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(B_n) + n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(\bar{B}_n),$$

where I is the indicator function. Define

$$\tilde{S}_{n1a}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(B_n)$$

and

$$\tilde{S}_{n1b}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(\bar{B}_n) .$$

For any $\varepsilon > 0$,

$$P\left[\max_{J \in \mathcal{J}_n} \left(\frac{\rho_J^2}{n} \right)^{-1} |\tilde{S}_{n1}(J) - E\tilde{S}_{n1}(J)| > 2\varepsilon \right] \le P\left[\max_{J \in \mathcal{J}_n} |\tilde{S}_{n1a}(J) - E\tilde{S}_{n1a}(J)| > \varepsilon \right]$$

$$+ P\left[\max_{J \in \mathcal{J}_n} |\tilde{S}_{n1b}(J) - E\tilde{S}_{n1b}(J)| > \varepsilon \right].$$

Now

(A.13) $$P\left[\max_{J \in \mathcal{J}_n} |\tilde{S}_{n1b}(J) - E\tilde{S}_{n1b}(J)| > \varepsilon \right] \to 0$$

as $n \to \infty$ because $P(\bar{B}_n) \to 0$. Now consider $\tilde{S}_{n1a}$. By Hoeffding's inequality

$$P[|\tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a}(J) \mid B_n)| > \varepsilon \mid B_n] \le 2\exp[-2\varepsilon^2 n / (c_1^2 J_n^{4+4\tau} a_n^2)]$$

$$\le 2\exp[-2\varepsilon^2 n^{1-2d} / (c_1^2 J_n^{4+4\tau})]$$

for every $J \in \mathcal{J}_n$. Therefore,

$$P\left[\max_{J \in \mathcal{J}_n} |\tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a}(J)| > \varepsilon \mid B_n\right] \le 2J_n \exp[-2\varepsilon^2 n^{1-2d} / (c_1^2 J_n^{4+4\tau})].$$

In the mildly ill-posed case, $J_n = o\left[n^{1/(2r+2)}\right]$ with $r \ge 2$. In the severely ill-posed case, $J_n = o(\log n)$. Therefore,

(A.14) $$P\left[\max_{J \in \mathcal{J}_n} |\tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a}(J)| > \varepsilon \mid B_n\right] \to 0$$

as $n \to \infty$. Because $P(B_n) \to 1$, it follows from (A.13) and (A.14) that

$$P\left[\max_{J \in \mathcal{J}_n} \left(\frac{\rho_J^2}{n}\right)^{-1} |\tilde{S}_{n1}(J) - E(\tilde{S}_{n1})| > \varepsilon\right] \to 0.$$

But $E\tilde{S}_{n1}(J) = S_n(J)[1 + o(1)]$ and $S_n(J) \ge C\rho_J^2 / n$ for some finite constant C uniformly over $J \in \mathcal{J}_n$. Therefore,

(A.15) $$\frac{\tilde{S}_{n1}(J) - S_n(J)}{S_n(J)} \le \varepsilon$$

with probability approaching 1 as $n \to \infty$ for any $\varepsilon > 0$ and every $J \in \mathcal{J}_n$.

Now consider $\tilde{S}_{n2}(J) / S_n(J)$. Assumption 5 implies that $|g_{J_n}(x) - g(x)| = o(J_n^{-2-2\tau})$ as $n \to \infty$ uniformly over $x \in [0,1]$ for some constant $c_2 < \infty$. Therefore,

$$|\tilde{S}_{n2}(J)| \le o(J_n^{-2-2\tau}) n^{-2} \sum_{i=1}^{n} |U_i| \sum_{j=1}^{J} \{[(A_{J_n}^{-1})^* \psi_j](W_i)\}^2 .$$

Now $\{[(A_{J_n}^{-1})^* \psi_j](w)\}^2 \le c_3 J_n^{1+2\tau} \rho_J^2$ for every $j \le J$ and some constant $c_3 < \infty$. Therefore,

$$|\tilde{S}_{n2}(J)| \le o(1) \rho_J^2 n^{-2} \sum_{i=1}^{n} |U_i|$$

uniformly over $J \in \mathcal{J}$. It follows from the strong law of large numbers that

$$|\tilde{S}_{n2}(J)| \le o(1) \rho_J^2 n^{-1} [E(|U|) + o_p(1)]$$

for every $J \in \mathcal{J}_n$. Therefore,

(A.16) $$\frac{|\tilde{S}_{n2}(J)|}{S_n(J)} = c_4 o(1)[E(|U|) + o_p(1)] = o_p(1)$$

for every $J \in \mathcal{J}_n$. A similar argument gives

$$|\tilde{S}_{n3}(J)| = o(1)\rho_J^2 n^{-1}$$

for some constant $C < \infty$ for every $J \in \mathcal{J}_n$, so

(A.17) $$\frac{|\tilde{S}_{n3}(J)|}{S_n(J)} = o(1)$$

for every $J \in \mathcal{J}_n$. The lemma follows by combining (A.15)-(A.17). Q.E.D.

Lemma 7: Given any $\varepsilon > 0$,

$$\left| \frac{\hat{S}_n(J) - \tilde{S}_n(J)}{S_n(J)} \right| \leq \varepsilon$$

for each $J \in \mathcal{J}_n$ with probability approaching 1 as $n \to \infty$.

Proof: Define

$$K_j(w) = \{[(A_{J_n}^{-1})^* \psi_j](w)\}^2,$$

$$\hat{K}_j(w) = \{[(\hat{A}^{-1})^* \psi_j](w)\}^2,$$

$$\Delta g(x) = \tilde{g}(x) - g_{J_n}(x),$$

$$\Delta K_j(w) = \hat{K}_j(w) - K_j(w),$$

$$\Delta S_{n1}(J) = -2n^{-2} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)] \Delta g(X_i) \sum_{j=1}^{J} K_j(W_i),$$

$$\Delta S_{n2}(J) = n^{-2} \sum_{i=1}^{n} [\Delta g(X_i)]^2 \sum_{j=1}^{J} K_j(W_i),$$

$$\Delta S_{n3}(J) = n^{-2} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)]^2 \sum_{j=1}^{J} \Delta K_j(W_i),$$

$$\Delta S_{n4}(J) = -2n^{-2} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)] \Delta g(X_i) \sum_{j=1}^{J} \Delta K_j(W_i),$$

and

$$\Delta S_{n5}(J) = n^{-2} \sum_{i=1}^{n} [\Delta g(X_i)]^2 \sum_{j=1}^{J} \Delta K_j(W_i).$$

Then

$$\hat{S}_n(J) - \tilde{S}_n(J) = \sum_{k=1}^{5} \Delta S_{nk}(J).$$

Because $S_n(J) \geq C\rho_J^2 / n$ for some constant $C < \infty$, it suffices to prove that

$$\max_{J \in \mathcal{J}} (\rho_J^2 / n)^{-1} |\hat{S}_n(J) - \tilde{S}_n(J)| \leq \varepsilon$$

with probability approaching 1 as $n \to \infty$.

Consider ΔS_{n1}. We have $\left\| \tilde{g} - g_{J_n} \right\| = O_p[\rho_{J_n}(J_n / n)^{1/2} + \nu_{J_n}^{-1}] = O_p[\rho_{J_n}(J_n / n)^{1/2}]$ under assumption 6. In addition, by $\{[(A_{J_n}^{-1})^* \psi_j](w)\}^2 \leq c_2 J_n^{1+2\tau} \rho_J^2$ for some constant $c_2 < \infty$ and each $J \in \mathcal{J}_n$ and $w \in [0,1]$. Moreover, $E[K_j(W)] \leq c_3 \rho_J^2$ for some constant $c_3 < \infty$. Therefore,

$$\left|\Delta S_{n1}(J)\right| = O_p[\rho_{J_n}(J_n / n)^{1/2}] n^{-2} \sum_{i=1}^{n} |Y_i - g_{J_n}(X_i)| \sum_{j=1}^{J} K_j(W_i)$$

$$= O_p[\rho_{J_n}(J_n / n)^{1/2}] n^{-2} \sum_{i=1}^{n} |U_i - [g_{J_n}(X_i) - g(X_i)]| \sum_{j=1}^{J} K_j(W_i)$$

$$= O_p[\rho_{J_n}(J_n / n)^{1/2}] \left\{ n^{-2} \sum_{i=1}^{n} |U_i| \sum_{j=1}^{J} K_j(W_i) + O_p(\nu_{J_n}^{-1}) n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{J} K_j(W_i) \right\}.$$

Using $E[K_j(W)] \leq c_3 \rho_J^2$, it follows from Markov's inequality that

$$\left|\Delta S_{n1}(J)\right| = O_p[\rho_{J_n}(J_n / n)^{1/2}](\rho_J^2 J / n)[1 + o_p(1)]$$

for every $J \in \mathcal{J}_n$, where $o_p(1)$ does not depend on J. But $\rho_{J_n}(J_n / n)^{1/2} = o(1)$ by assumption 6. Therefore,

$$(\rho_J^2 / n)^{-1} \left|\Delta S_{n1}(J)\right| = O_p[\rho_{J_n}(J_n / n)^{1/2}] J[1 + o_p(1)]$$

$$= O_p[\rho_{J_n}(J_n^3 / n)^{1/2}]$$

for every $J \in \mathcal{J}_n$. It follows that for any $\varepsilon > 0$

(A.18) $(\rho_J^2 / n)^{-1} |\Delta S_{n1}(J)| < \varepsilon$

with probability approaching 1 as $n \to \infty$. A similar argument shows that

(A.19) $(\rho_J^2/n)^{-1}\left|\Delta S_{n2}(J)\right| < \varepsilon$

with probability approaching 1 for every $J \in \mathcal{J}_n$.

Now consider $\Delta S_{n3}(J)$. We have

$$\Delta S_{n3}(J) = n^{-2}\sum_{i=1}^{n}\{U_i - [g_{J_n}(X_i) - g(X_i)]\}^2 \sum_{j=1}^{J}\Delta K_j(W_i)$$

$$= n^{-2}\sum_{i=1}^{n}U_i^2\sum_{j=1}^{J}\Delta K_j(W_i) - 2n^{-2}\sum_{i=1}^{n}U_i[g_{J_n}(X_i) - g(X_i)]\sum_{j=1}^{J}\Delta K_j(W_i)$$

$$+ n^{-2}\sum_{i=1}^{n}[g_{J_n}(X_i) - g(X_i)]^2\sum_{j=1}^{J}\Delta K_j(W_i)$$

$$\equiv \Delta S_{n3a}(J) + \Delta S_{n3b}(J) + \Delta S_{n3c}(J).$$

Some algebra shows that

$$\Delta K_j = 2[(A_{J_n}^{-1})^*\psi_j]\{[(\hat{A}^{-1})^* - (A_{J_n}^{-1})^*]\psi_j\} + \{[(\hat{A}^{-1})^* - (A_{J_n}^{-1})^*]\psi_j\}^2.$$

Moreover,

$$(\hat{A}^{-1})^* - (A_{J_n}^{-1})^* = -[I + (A_{J_n}^{-1})^*\Delta A^*]^{-1}(A_{J_n}^{-1})^*(\Delta A^*)(A_{J_n}^{-1})^*,$$

where I is the identity operator in $\mathcal{K}_{J_n}$ and $\Delta A^* = \hat{A}^* - A_{J_n}^*$. Now

$$\left\|\Delta A^*\right\| = \left\|\hat{A} - A_{J_n}\right\|$$

$$= O_p[(J_n/n)^{1/2}]$$

for every $J \in \mathcal{J}_n$ by lemma 2. Therefore, it follows from Lemma 1 and Markov's inequality that

$$\left\|[(\hat{A}^{-1})^* - (A_{J_n}^{-1})^*]\psi_j\right\| = \rho_J^2 O_p[(J_n/n)^{1/2}],$$

$$\left\|\Delta K_j\right\| = \rho_J^3 O_p[(J_n/n)^{1/2}],$$

and

$$\Delta S_{n3a}(J) = \rho_J^3 J O_p(J_n^{1/2} n^{-3/2})$$

for every $J \in \mathcal{J}_n$. It follows that $(\rho_J^2/n)^{-1}\Delta S_{n3a}(J) = O_p[\rho_J(J_n^3/n)^{1/2}] = o_p(1)$, where $o_p(1)$ does not depend on J and the last equality follows from assumption 6. Similar arguments apply to ΔS_{n3b}, ΔS_{n3c}, ΔS_{n4}, and ΔS_{n5}. The lemma follows by combining these results with (A.18) and (A.19). Q.E.D.

Lemma 6: The following inequality holds for every $J \in \mathcal{J}_n$ as $n \to \infty$.

$$\left\| \hat{g}_J - g_J \right\|^2 \leq (4/3)(\log n) S_n(J)[1 + o_p(1)],$$

where $o_p(1)$ does not depend on J.

Proof: Let s_{nJ} denote the leading term of the asymptotic expansion of $\left\| \hat{g}_J - g_J \right\|^2$. By Proposition 1 and lemma 3,

$$\text{(A.20)} \quad s_{nJ} = \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} U_i [(A_{J_n}^{-1})^* \psi_j](W_i) \right\}^2 .$$

Define

$$\sigma_j^2 = E\{U_i [(A_{J_n}^{-1})^* \psi_j](W_i)\}^2,$$

$$V_{ij} = U_i [(A_{J_n}^{-1})^* \psi_j](W_i),$$

$$R_{nj} = n^{-1} \sum_{i=1}^{n} V_{ij},$$

and

$$\xi_{nj} = [(4/3)\sigma_j^2 n^{-1} \log n]^{1/2}.$$

Then

$$s_{nJ} = \sum_{j=1}^{J} R_{nj}^2 .$$

By Bernstein's inequality

$$P(|R_{nj}| > \xi_{nj}) \leq 2 \exp\left(- \frac{n \xi_{nj}^2}{4\sigma_j^2 + 2cJ^{(1+2\tau)/2} \xi_{nj}} \right)$$

for some finite constant $c > 0$ that does not depend on j. But σ_j^2 is bounded away from 0 for all j. Therefore,

$$P(|R_{nj}| > \xi_{nj}) \leq 2 \exp\left[- \frac{(4/3)\log n}{4 + \varepsilon/3} \right] = 2n^{-4/(12+\varepsilon)}$$

for any $\varepsilon > 0$, all sufficiently large n, and all $j \leq J$. Therefore,

$$P\left(\bigcap_{j=1}^{J_n} |R_{nj}| > \xi_{nj}\right) \leq \sum_{j=1}^{J_n} P(|R_{nj}| > \xi_{nj})$$

$$\leq 2J_n n^{-4/(12+\varepsilon)} \to 0$$

as $n \to \infty$ if ε is sufficiently small, where the last relation follows from assumption 6 and the observation that ρ_{J_n} increases at least as fast as J_n^{β} for some $\beta > 0$. Combining this result with (A.20) gives

$$s_{nJ} \leq \sum_{j=1}^{J} \xi_{nj}^2$$

$$= (4/3)n^{-1}(\log n)\sum_{j=1}^{J} \sigma_j^2$$

$$= (4/3)(\log n)S_n(J)[1 + o_p(1)]$$

for every $J \in \mathcal{J}_n$, where $o_p(1)$ does not depend on J. Q.E.D.

Lemma 9: The following inequality holds.

$$\left\|\hat{g}_{\hat{J}}\right\|^2 - \left\|g_{\hat{J}}\right\|^2 \leq 3\left\|\hat{g}_{\hat{J}} - g_{\hat{J}}\right\|^2 + 0.5\left\|g_{\hat{J}} - g\right\|^2 + 0.5\left\|g_{J_{opt}} - g\right\|^2$$

$$+ 2\left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle.$$

Proof: The proof of this lemma is similar to the proof of lemma 3.4(ii) of Loubes and Marteau (2009). We have

$$\left\|\hat{g}_{\hat{J}}\right\|^2 = \left\|(\hat{g}_{\hat{J}} - g_{\hat{J}}) + g_{\hat{J}}\right\|^2$$

$$= \left\|\hat{g}_{\hat{J}} - g_{\hat{J}}\right\|^2 + 2\left\langle g_{\hat{J}}, \hat{g}_{\hat{J}} - g_{\hat{J}}\right\rangle + \left\|g_{\hat{J}}\right\|^2.$$

Therefore,

$$\left\|\hat{g}_{\hat{J}}\right\|^2 - \left\|g_{\hat{J}}\right\|^2 = \left\|\hat{g}_{\hat{J}} - g_{\hat{J}}\right\|^2 + 2\left\langle g_{\hat{J}}, \hat{g}_{\hat{J}} - g_{\hat{J}}\right\rangle.$$

Define $\Sigma_{J_{opt}:\hat{J}} = \Sigma_{j=J_{opt}+1}^{\hat{J}}$, if $\hat{J} > J_{opt}$, $-\Sigma_{j=\hat{J}+1}^{J_{opt}}$ if $\hat{J} < J_{opt}$, and 0 if $\hat{J} = J_{opt}$. Then,

$$2\left\langle g_{\hat{J}}, \hat{g}_{\hat{J}} - g_{\hat{J}} \right\rangle = 2\sum_{j=1}^{\hat{J}} b_j(\tilde{b}_j - b_j)$$

$$= 2\sum_{j=1}^{J_{opt}} b_j(\tilde{b}_j - b_j) + 2\sum_{J_{opt}:\hat{J}} b_j(\tilde{b}_j - b_j)$$

$$= 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle + 2\sum_{J_{opt}:\hat{J}} b_j(\tilde{b}_j - b_j)$$

and

(A.21) $$\left\|\hat{g}_{\hat{J}}\right\|^2 - \left\|g_{\hat{J}}\right\|^2 = \left\|\hat{g}_{\hat{J}} - g_{\hat{J}}\right\|^2 + 2\sum_{J_{opt}:\hat{J}} b_j(\tilde{b}_j - b_j) + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle.$$

Define

$$R_n = 2\sum_{J_{opt}:\hat{J}} b_j(\tilde{b}_j - b_j).$$

Then

$$|R_n| \le 2\sum_{j=1}^{\infty} |I(j \le \hat{J}) - I(j \le J_{opt})|\, |b_j(\tilde{b}_j - b_j)|,$$

where $I(\cdot)$ is the indicator function. But

$$|I(j \le \hat{J}) - I(j \le J_{opt})| = [I(j \le \hat{J}) + I(j \le J_{opt})]\, |I(j \le \hat{J}) - I(j \le J_{opt})|$$

$$\le I(j \le \hat{J})I(j > J_{opt}) + I(j \le J_{opt})I(j > \hat{J}).$$

Therefore,

$$|R_n| \le 2\sum_{j=1}^{\infty} I(j \le J_{opt})I(j > \hat{J})\, |b_j(\tilde{b}_j - b_j)| + 2\sum_{j=1}^{\infty} I(j \le \hat{J})I(j > J_{opt})\, |b_j(\tilde{b}_j - b_j)|.$$

By the Cauchy-Schwarz inequality,

$$|R_n| \le 2\left(\sum_{j=\hat{J}}^{\infty} b_j^2\right)^{1/2}\left[\sum_{j=1}^{J_{opt}} (\tilde{b}_j - b_j)^2\right]^{1/2} + 2\left(\sum_{j=J_{opt}}^{\infty} b_j^2\right)^{1/2}\left[\sum_{j=1}^{\hat{J}} (\tilde{b}_j - b_j)^2\right]^{1/2}.$$

In addition, $2ab \le a^2/2 + 2b^2$ for any real numbers a and b. Therefore,

$$|R_n| \le 0.5\sum_{j=J_{opt}}^{\infty} b_j^2 + 0.5\sum_{j=\hat{J}}^{\infty} b_j^2 + 2\sum_{j=1}^{J_{opt}} (\tilde{b}_j - b_j)^2 + 2\sum_{j=1}^{\hat{J}} (\tilde{b}_j - b_j)^2$$

(A.22) $$= 0.5\left\|g_{J_{opt}} - g\right\|^2 + 0.5\left\|g_{\hat{J}} - g\right\|^2 + 2\left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + 2\left\|\hat{g}_{\hat{J}} - g_{\hat{J}}\right\|^2.$$

The lemma follows by substituting (A.22) into (A.21). Q.E.D.

Proof of Theorem 3.1: Define $a_n = (2/3)\log(n)$ and

$$\hat{Q}_n(J) = \hat{T}_n(J) + \|g\|^2$$

$$= a_n \hat{S}_n(J) + \|g\|^2 - \|\hat{g}_J\|^2.$$

Then $\hat{J}$ minimizes $\hat{Q}_n(J)$ over $J \in \mathcal{J}_n$. By lemmas 6 and 7,

$$\hat{Q}_n(J) = a_n S_n(J)[1 + o_p(1)] + \|g\|^2 - \|\hat{g}_J\|^2$$

for all $J \in \mathcal{J}_n$, where $o_p(1)$ does not depend on J, so

$$\hat{Q}_n(\hat{J}) = a_n S_n(\hat{J})[1 + o_p(1)] + \|g\|^2 - \|\hat{g}_{\hat{J}}\|^2.$$

It follows that

$$a_n S_n(\hat{J})[1 + o_p(1)] + \|g\|^2 - \|g_{\hat{J}}\|^2 = \hat{Q}_n(\hat{J}) + \|\hat{g}_{\hat{J}}\|^2 - \|g_{\hat{J}}\|^2.$$

An application of lemma 9 gives

$$a_n S_n(\hat{J})[1 + o_p(1)] + \|g\|^2 - \|g_{\hat{J}}\|^2 \le \hat{Q}_n(\hat{J}) + 3\|\hat{g}_{\hat{J}} - g_{\hat{J}}\|^2 + 0.5(\|g\|^2 - \|g_{\hat{J}}\|^2)$$

$$+ .5\|g_{J_{opt}} - g\|^2 + 2\|\hat{g}_{J_{opt}} - g_{J_{opt}}\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle.$$

In addition, it follows from Proposition 1 that $a_n S_n(\hat{J})$ dominates $\|\hat{g}_{\hat{J}} - g_{\hat{J}}\|^2$ as $n \to \infty$. Therefore,

$$a_n S_n(\hat{J})[1 + o_p(1)] + 0.5(\|g\|^2 - \|g_{\hat{J}}\|^2) - 3\|\hat{g}_{\hat{J}} - g_{\hat{J}}\|^2$$

$$\le \hat{Q}_n(\hat{J}) + .5\|g_{J_{opt}} - g\|^2 + 2\|\hat{g}_{J_{opt}} - g_{J_{opt}}\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle.$$

By lemma 8,

$$a_n[(4/3)\log n]^{-1}\|\hat{g}_{\hat{J}} - g_{\hat{J}}\|^2 \le a_n S_n(\hat{J})[1 + o_p(1)],$$

where $o_p(1)$ does not depend on $\hat{J}$. Therefore,

$$0.5\left\|\hat{g}_{\hat{J}} - g\right\|^2 [1 + o_p(1)] \le \hat{Q}_n(\hat{J}) + .5\left\|g_{J_{opt}} - g\right\|^2$$

$$+ 2\left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle,$$

where $o_p(1)$ does not depend on $\hat{J}$. But $\hat{Q}_n(\hat{J}) \le \hat{Q}_n(J_{opt})$, so

$$0.5\left\|\hat{g}_{\hat{J}} - g\right\|^2 [1 + o_p(1)] \le \hat{Q}_n(J_{opt}) + .5\left\|g_{J_{opt}} - g\right\|^2$$

$$+ 2\left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle.$$

In addition,

$$\hat{Q}_n(J_{opt}) = a_n \hat{S}_n(J_{opt}) + \|g\|^2 - \left\|\hat{g}_{J_{opt}}\right\|^2.$$

Therefore, by lemmas 6 and 7,

$$\hat{Q}_n(J_{opt}) = a_n S_n(J_{opt})[1 + o_p(1)] + \|g\|^2 - \left\|\hat{g}_{J_{opt}}\right\|^2.$$

But

$$\left\|\hat{g}_{J_{opt}}\right\|^2 = \left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + \left\|g_{J_{opt}}\right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle,$$

so

$$\hat{Q}_n(J_{opt})$$

$$= a_n S_n(J_{opt})[1 + o_p(1)] + \|g\|^2 - \left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 - \left\|g_{J_{opt}}\right\|^2 - 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}}\right\rangle$$

and

$$0.5\left\|\hat{g}_{\hat{J}} - g\right\|^2 [1 + o_p(1)] \le a_n S_n(J_{opt})[1 + o_p(1)] + \left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 + 1.5\left\|g_{J_{opt}} - g\right\|^2.$$

Now, $E_A\left\|\hat{g}_{J_{opt}} - g_{J_{opt}}\right\|^2 = S_n(J_{opt})$. Therefore, for any $n > 1$,

$$0.5E_A\left\|\hat{g}_{\hat{J}} - g\right\|^2 \le (a_n + 1)S_n(J_{opt}) + 1.5\left\|g_{J_{opt}} - g\right\|^2$$

$$\le (a_n + 1)E_A\left\|\hat{g}_{J_{opt}} - g\right\|^2,$$

and

$$E_A\left\|\hat{g}_{\hat{J}} - g\right\|^2 \le 2(a_n + 1)E_A\left\|\hat{g}_{J_{opt}} - g\right\|^2. \quad \text{Q.E.D.}$$

Proof of Theorem 3.2: We first prove part (ii) of the theorem. J_{n0} satisfies the assumptions of Theorem 3.1, so the conclusion of Theorem 3.1 holds with J_{n0} in place of J_n. Therefore, part (ii) of the theorem follows from the fact that, by part (i), $\hat{J}_{n0} = J_{n0}$ with probability approaching 1 as $n \to \infty$.

We now prove part (i) of the theorem. For $J = 1,2,...$ Define $L_n(J) = \rho_J^2 J^{3.5}/n$ and $\hat{L}_n(J) = \hat{\rho}_J^2 J^{3.5}/n$. Let $C > 2$ be a finite constant. Define $\Theta_n = \{J = 1,2,...: |J - J_{n0}| \le C\}$. We first show that

$$\text{(A.23)} \quad \max_{J \in \Theta_n} \left| \frac{\hat{L}_n(J) - L_n(J)}{L_n(J)} \right| = \rho_{J_{n0}} J_{n0} O_p(n^{-1/2}) = o_p(1),$$

as $n \to \infty$. To do this, observe that

$$\hat{L}_n(J)/L_n(J) = \hat{\rho}_J^{-2}/\rho_J^{-2},$$

$$\rho_J^{-1} = \inf_{\nu \in \mathcal{K}_J} \frac{\|A\nu\|}{\|\nu\|},$$

and

$$\hat{\rho}_J^{-1} = \inf_{\nu \in \mathcal{K}_J} \frac{\|\hat{A}\nu\|}{\|\nu\|}.$$

For any $\nu \in \mathcal{K}_J$,

$$\frac{\|\hat{A}\nu\|}{\|\nu\|} = \frac{\|A\nu + (\hat{A} - A)\nu\|}{\|\nu\|}$$

$$\le \frac{\|A\nu\|}{\|\nu\|} + \frac{\|(\hat{A} - A)\nu\|}{\|\nu\|}$$

$$\le \frac{\|A\nu\|}{\|\nu\|} + \sup_{\nu \in \mathcal{K}_J} \frac{\|(\hat{A} - A)\nu\|}{\|\nu\|}.$$

Similarly

$$\frac{\left\|\hat{A}\nu\right\|}{\left\|\nu\right\|} \geq \frac{\left\|A\nu\right\|}{\left\|\nu\right\|} - \frac{\left\|(\hat{A}-A)\nu\right\|}{\left\|\nu\right\|}$$

$$\geq \frac{\left\|A\nu\right\|}{\left\|\nu\right\|} - \sup_{\nu\in\mathcal{K}_J} \frac{\left\|(\hat{A}-A)\nu\right\|}{\left\|\nu\right\|}.$$

A slight modification of the proof of lemma 2 shows that

$$\sup_{\nu\in\mathcal{K}_J} \frac{\left\|(\hat{A}-A)\nu\right\|}{\left\|\nu\right\|} = JO_p(n^{-1/2}).$$

Therefore, for $r_n = JO_p(n^{-1/2})$,

$$\inf_{\nu\in\mathcal{K}_J} \frac{\left\|\hat{A}\nu\right\|}{\left\|\nu\right\|} \leq \inf_{\nu\in\mathcal{K}_J} \frac{\left\|A\nu\right\|}{\left\|\nu\right\|} + r_n.$$

Similarly

$$\inf_{\nu\in\mathcal{H}_{J_S}} \frac{\left\|\hat{A}\nu\right\|}{\left\|\nu\right\|} \geq \inf_{\nu\in\mathcal{K}_J} \frac{\left\|A\nu\right\|}{\left\|\nu\right\|} - r_n.$$

It follows that

$$\hat{\rho}_J^{-1} \leq \rho_J^{-1} + r_n$$

and

$$\hat{\rho}_J^{-1} \geq \rho_J^{-1} - r_n.$$

Therefore, $|\hat{\rho}_J^{-1}/\rho_J^{-1} - 1| \leq \rho_J r_n$ and

(A.24) $$\left|\frac{\hat{L}_n(J) - L_n(J)}{L_n(J)}\right| = \rho_J JO_p(n^{-1/2}).$$

(A.23) follows from (A.24) and the observation that $\rho_J \asymp \rho_{J_{n0}}$ for $J \in \Theta_n$.

Now define

$$\tilde{J}_{n0} = \arg \min_{J=1,2,...;\, J\in\Theta_n} \{\hat{\rho}_J^2 J^{3.5}/n :\ \rho_J^2 J^{3.5}/n - 1 \geq 0\}$$

We show that $\lim_{n\to\infty} P(\tilde{J}_{n0} = J_{n0}) = 1$. It follows from (A.23) that

$$\hat{L}_n(J_{n0}) < L_n(J_{n0})\left[1 + \rho_{J_{n0}} J_{n0} O_p(n^{-1/2})\right]$$

and

$$L_n(\tilde{J}_{n0}) < \hat{L}_n(\tilde{J}_{n0})\left[1 + \rho_{J_{n0}} J_{n0} O_p(n^{-1/2})\right]$$

In addition $\hat{L}_n(\tilde{J}_{n0}) \le \hat{L}_n(J_{n0})$. Therefore,

(A.25) $L_n(\tilde{J}_{n0}) < L_n(J_{n0})\left[1+\rho_{J_{n0}} J_{n0} O_p(n^{-1/2})\right]$.

Now let $\mathcal{N} = \Theta_n - J_{n0}$. Then

$$J_{n0}+1 = \min_{J\in\mathcal{N}}\{L_n(J) : L_n(J)-1>0\}.$$

But $L_n(J_{n0}+1) > L_n(J_{n0})$. Therefore, it follows from (A.25) that $\tilde{J}_{n0} \notin \mathcal{N} - J_{n0}$, so $\tilde{J}_{n0} = J_{n0}$.

If $\hat{J}_{n0} < J_{n0} - C$, then $\hat{L}_n(\hat{J}_{n0}) - 1 < \hat{L}_n(J_{n0} - C) - 1$. Therefore, by (A.23)

$$0 < \hat{L}_n(J_{n0} - C) - 1 < L_n(J_{n0} - C)[1+\rho_{J_{n0}} J_{n0} O_p(n^{-1/2})] - 1,$$

which is impossible because J_{n0} minimizes $L_n(J)$ subject to $L_n(J) - 1 \ge 0$. Therefore, $\hat{J}_{n0} < J_{n0} - C$ cannot happen when n is large. In addition, it follows from (A.23) that $\hat{L}_n(J_{n0}+1) - 1 > 0$ with probability approaching 1 as $n \to \infty$. Therefore, with probability approaching 1 as $n \to \infty$, $\hat{J}_{n0} \le J_{n0} + 1$ and $\hat{J}_{n0} > J_{n0} + C$ cannot happen. Q.E.D.

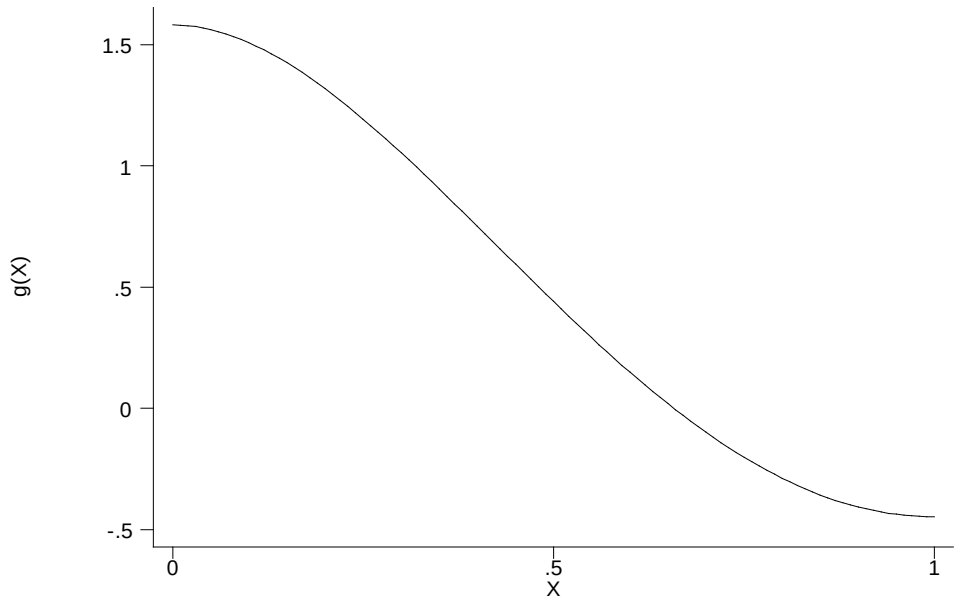

Figure 1: Graph of $g(x)$

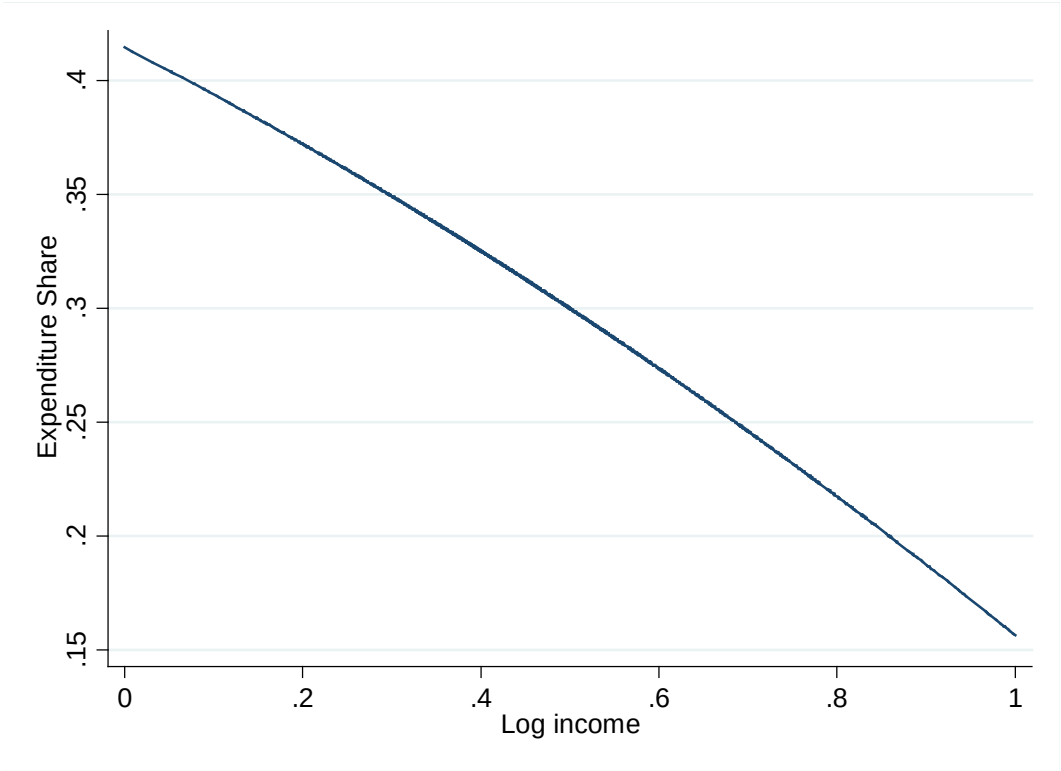

Figure 2: Estimated Engel Curve for Food

TABLE 1: RESULTS OF MONTE CARLO EXPERIMENTS

Exp't No.	Design	Empirical mean of $\left\|\hat{g}_{J_{opt}} - g\right\|^2$	Empirical mean of $\left\|\hat{\hat{g}}_{\hat{J}} - g\right\|^2$	Ratio of empirical means, R	Theoretical asymptotic upper bound on R
1	1	0.0957	0.100	1.045	11.2
2	1	0.0983	0.138	1.400	11.2
3	1	0.100	0.100	1.000	11.2
4	1	0.0940	0.0978	1.040	11.2
5	1	0.103	0.203	1.977	11.2
6	2	0.0020	0.0039	1.999	11.8
7	2	0.0029	0.0029	1.000	11.8

REFERENCES

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, 71, 1795-1843.

Bauer, F. and T. Hohage (2005). A Lepskij-type stopping rule for regularized Newton methods, *Inverse Problems*, 21, 1975-1991.

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves, *Econometrica*, 75, 1613-1669.

Blundell, R. and J.L. Horowitz (2007). A non-parametric test of exogeneity, *Review of Economic Studies*, 74, 1035-1058.

Blundell, R., P. Pashardes, and G. Weber (1993). What do we learn about consumer demand patterns from micro data? *American Economic Review*, 83, 570-597.

Carrasco, M., J.-P. Florens, and E. Renault (2007): Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization, in *Handbook of Econometrics*, Vol. 6, ed. by E.E. Leamer and J.J. Heckman. Amsterdam: North-Holland, pp. 5634-5751.

Cavalier, L. and N.W. Hengartner (2005). Adaptive estimation for inverse problems with noisy operators, *Inverse Problems*, 21, 1345-1361.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals, *Journal of Econometrics*, 152, 46-60.

Chen, X. and D. Pouzo (2008). Estimation of nonparametric conditional moment models with possibly nonsmooth moments, working paper, Department of Economics, Yale University, New Haven, CT.

Chen, X. and M. Reiss (2011). On rate optimality for ill-posed inverse problems in econometrics, *Econometric Theory*, 27, 472-521.

Chernozhukov, V., G.W. Imbens, and W.K. Newey (2007). Instrumental variable identification and estimation of nonseparable models via quantile conditions, *Journal of Econometrics*, 139, 4-14.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): Nonparametric instrumental regression, *Econometrica*, 79, 1541-1565.

Efromovich, S. and V. Koltchinskii (2001). On inverse problems with unknown operators, *IEEE Transactions on Information Theory*, 47, 2876-2894.

Engl, H.W., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic Publishers.

Florens, J.-P. and A. Simoni (2010). Nonparametric estimation of an instrumental regression: a quasi-Bayesian approach based on regularized posterior, working paper, Department of Decision Sciences, Bocconi University, Milan, Italy.

Gagliardini, P. and O. Scaillet (2012). Nonparametric instrumental variable estimation of structural quantile effects, *Econometrica*, 80, 1533-1562.

Hall, P. and J.L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables, *Annals of Statistics*, 33, 2904-2929.

Horowitz, J.L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables, *Econometrica*, 74, 521-538.

Horowitz, J.L. (2007). Asymptotic normality of a nonparametric instrumental variables estimator, *International Economic Review*, 48, 1329-1349.

Horowitz, J.L. (2012). Specification testing in nonparametric instrumental variables estimation, *Journal of Econometrics*, 167, 383-396.

Horowitz, J.L. and S. Lee (2007). Nonparametric instrumental variables estimation of a quantile regression model, *Econometrica*, 75, 1191-1208.

Horowitz, J.L. and S. Lee (2012). Uniform confidence bands for functions estimated nonparametrically with instrumental variables, *Journal of Econometrics*, 168, 175-188.

Johannes, J. and M. Schwarz (2010). Adaptive nonparametric instrumental regression by model selection, working paper, Université Catholique de Louvain.

Kress, R. (1999). *Linear Integral Equations*, 2nd edition, New York: Springer-Verlag.

Loubes, J.-M. and C. Ludeña (2008). Adaptive complexity regularization for inverse problems, *Electronic Journal of Statistics*, 2, 661-677.

Loubes, J.-M. and C. Marteau (2009). Oracle Inequality for Instrumental Variable Regression, working paper, Institute of Mathematics, University of Toulouse 3, France.

Lukas, M.A. (1993). Asymptotic optimality of generalized cross-validation for choosing the regularization parameter, *Numerische Mathematik*, 66, 41-66.

Lukas, M.A. (1998). Comparisons of parameter choice methods for regularization with discrete, noisy data, *Inverse Problems*, 14, 161-184.

Marteau, C. (2006). Regularization of inverse problems with unknown operator, *Mathematical Methods of Statistics*, 15, 415-443.

Marteau, C. (2009). On the stability of the risk hull method, *Journal of Statistical Planning and Inference*, 139, 1821-1835.

Mathé, P. and S.V. Pereverzev (2003). Geometry of ill posed problems in variable Hilbert scales, *Inverse Problems*, 19, 789-803.

Newey, W.K. and J.L. Powell (2003). Instrumental variables estimation of nonparametric models, *Econometrica*, 71, 1565-1578.

Newey, W.K., J.L. Powell, and F. Vella (1999): Nonparametric estimation of triangular simultaneous equations models, *Econometrica*, 67, 565-603.

Spokoiny, V. and C. Vial (2009). Parameter tuning in pointwise adaptation using a propagation approach, *Annals of Statistics*, 37, 2783-2807.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy, *SIAM Journal of Numerical Analysis*, 14, 651-666.