A Parametric Approach to Flexible Nonlinear Inference*

James D. Hamilton

Department of Economics, 0508

University of California, San Diego

La Jolla, CA 92093-0508

May 1997

Revised: May 1999

*This paper is based on research supported by the NSF under Grant No. SBR-9707771, and has benefited from comments on earlier drafts by Christian Dahl, Robert Engle, Steve Gordon, Ana Herrera, Bruce Lehmann, Halbert White, Richard Blundell, and anonymous referees.. All data and software developed for this paper can be downloaded free of charge from http://weber.ucsd.edu/~jhamilto.

ABSTRACT

This paper proposes a new framework for determining whether a given relationship is nonlinear, what the nonlinearity looks like, and whether it is adequately described by a particular parametric model. The paper studies a regression or forecasting model of the form $y_t = \mu(\mathbf{x}_t) + \varepsilon_t$ where the functional form of $\mu(.)$ is unknown. We propose viewing $\mu(.)$ itself as the outcome of a random process. The paper introduces a new stationary random field $m(.)$ that generalizes finite-differenced Brownian motion to a vector field and whose realizations could represent a broad class of possible forms for $\mu(.)$. We view the parameters that characterize the relation between a given realization of $m(.)$ and the particular value of $\mu(.)$ for a given sample as population parameters to be estimated by maximum likelihood or Bayesian methods. We show that the resulting inference about the functional relation also yields consistent estimates for a broad class of deterministic functions $\mu(.)$. The paper further develops a new test of the null hypothesis of linearity based on the Lagrange multiplier principle and small-sample confidence intervals based on numerical Bayesian methods. An empirical application suggests that properly accounting for the nonlinearity of the inflation-unemployment tradeoff may explain the previously reported uneven empirical success of the Phillips Curve.

1 Introduction

There has been a lot of interest recently in whether nonlinear statistical models can improve on linear forecasts or shed light on particular economic hypotheses. Parametric approaches to nonlinear dynamics include time-varying parameter models (for example, Sims 1993), ARCH-related specifications (surveyed by Bollerslev, Chou, and Kroner, 1992, and Hamilton, 1994), threshold autoregressions (Tong, 1983; Tsay, 1989; Potter, 1995), regime-switching models (Hamilton, 1989), and smooth transition autoregressions (Granger, Terasvirta, and Anderson, 1993). A problem with any of these parametric approaches is deciding which model to use, that is, deciding in what way the data might be nonlinear.

There is much to be said for flexible nonparametric methods that offer consistent estimates within a broad class of nonlinear relations. Popular flexible nonparametric approaches include kernel methods (Nadaraya, 1964; Watson, 1964; Härdle, 1990; Robinson, 1983; Diebold and Nason, 1990; Fan, et. al., 1996), series expansions (e.g., Gallant and Nychka, 1987), wavelets (Donoho, et. al., 1995), nearest neighbor (Yakowitz, 1987; Mizrach, 1992); smoothing splines (Reinsch, 1967; Eubank, 1988; Wahba, 1990), and local polynomials (Seifert and Gasser, 1996). Unfortunately, these nonparametric approaches sacrifice many of the benefits of parametric methods. First, one needs some system for adjusting a bandwidth or series expansion length as the sample size grows. Second, it is unclear how a classical or Bayesian statistician should interpret the inferences that result from the procedure for a given sample of fixed size. Third, the methods are not readily adapted for the hypothesis testing and model simplification that are quite necessary in order to make

sense of a multivariate nonlinear relation.

This paper proposes a flexible parametric framework for investigating nonlinear relations that combines the advantages of the two approaches.

The object we seek to estimate is the expectation of a scalar y_t conditional on an observed vector $\mathbf{x}_t$:

$$E(y_t|\mathbf{x}_t) = \mu(\mathbf{x}_t). \tag{1.1}$$

The proposal is to view the underlying conditional expectation function $\mu(\mathbf{x})$ as itself the outcome of a stochastic process that associates any possible value for $\mathbf{x}$ with a scalar $\mu(\mathbf{x})$. We think of nature as having generated a single realization of $\mu(.)$ prior to generating the observed data $\{\mathbf{x}_t, y_t\}_{t=1}^T$. The econometrician's task is then to form an inference about the nature of the realized value for $\mu(.)$ based on the properties of the observed data.

For $\mathbf{x}_t$ a scalar, one possibility would be to view $\mu(x)$ as an unobserved realization of Brownian motion. As noted by Lauritzen (1981), this idea goes back over a century to work by T. N. Thiele in 1880. Wahba (1978) showed that such a structure implies that inference about the unknown function takes the form of a smoothing polynomial spline, and most of the literature has pursued such models from the perspective of choosing the smoothing parameter (or an implicit variance of the Brownian motion) by means of cross-validation or other nonparametric methods. By contrast, the approach followed here more closely follows Wecker and Ansley's (1983) view that the latent stochastic process is part of the true data-generating process, with the properties of the latent process regarded as population parameters to be estimated by maximum likelihood or Bayesian methods.

More generally, for $\mathbf{x}_t$ a vector, if $\mu(\mathbf{x})$ is viewed as a realization of a particular Gaussian random field, then the optimal inference about the unobserved nonlinear relation appears to take the form of a thin-plate spline for some specification of the smoothness penalty (see Wahba, 1990, Section 2.5). One key unanswered question is, what Gaussian random field is appropriate to employ, or what is the logical way to generalize univariate Brownian motion to k dimensions?

This paper introduces a new Gaussian random field that appears to provide a sensible answer to this question. The estimator that results from this specification could be interpreted as an example of a thin-plate spline, albeit a particular thin-plate spline that does not appear to have been used previously by empirical researchers. The key difference from most of the existing nonparametric literature is the perspective adopted throughout this paper that the estimator represents the optimal inference for a maintained parametric model as opposed to an atheoretical data-smoothing device. The claimed benefits of this focus are the following. (1) The paper develops a test of the hypothesis of linearity against a broad class of nonlinear alternatives based on the Lagrange multiplier principle; no such result appears in the nonparametric literature. (2) Fixed values of the parameters in the framework proposed here would be associated with different values of the implicit smoothness or bandwidth parameters for different sample sizes, with the result that the inference procedure proposed here automatically adjusts what would correspond to smoothness or bandwidth parameters in a nonparametric formulation as the sample size increases so as to obtain consistent estimates. (3) The framework here is specifically designed to identify which

variables contribute to the nonlinearity. (4) The framework here allows ready calculation of exact small-sample confidence intervals for the nonlinear relation. (5) The framework here allows immediate testing of whether a conventional parametric nonlinear model adequately describes any nonlinearities in the data. In sum, the paper thus proposes a new tool-kit for the task of modelling a nonlinear relation, with new methods to see if the relation is nonlinear, what that nonlinearity looks like, and whether it is correctly described by some particular model, and do all this within a single encompassing framework.

The plan of the paper is as follows. Section 2 describes the stochastic process assumed for $\mu(\mathbf{x})$. Section 3 describes algorithms for optimal statistical inference about $\mu(.)$ conditional on the population parameters and for estimation of population parameters by maximum likelihood. Section 4 discusses asymptotic properties of the inference. Section 5 develops procedures for small-sample Bayesian inference. Section 6 develops the Lagrange multiplier test of the null hypothesis that $\mu(\mathbf{x})$ is linear. Applications are provided in Section 7.

2 The stochastic process assumed for the conditional expectation function

2.1 The case of a single explanatory variable

We first describe a latent stochastic process $m(x)$ which will be used to characterize the conditional expectation function $\mu(x)$ when the explanatory variable x is a scalar. Consider $[a, b]$ a closed interval in $\Re^1$. Let ω be a parameter to be described shortly, and partition the

interval $[a-\omega, b+\omega]$ as $\{x_1, ..., x_N\}$ where $x_1 = a-\omega$, $x_N = b+\omega$, and $x_i = x_{i-1} + \Delta_N$ for $i = 2, ..., N$. We imagine generating for each point x_i a standard Normal variable $e(x_i)$ with $e(x_i)$ independent of $e(x_j)$ for $i \neq j$. Figure 1 displays an illustrative example for $\Delta_N = 0.5$.

For each node x_i such that $a \leq x_i \leq b$, we further construct a random variable $m_N(x_i)$ which is proportional to the average value of $e(x_j)$ for all x_j whose distance from x_i is less than or equal to ω. The constant of proportionality is the square root of the number of values of $e(x_j)$ that are averaged. For example, when ω/Δ_N is an integer,

$$m_N(x_i) = (1 + 2\omega/\Delta_N)^{-1/2} \sum_{j=-\omega/\Delta_N}^{\omega/\Delta_N} e(x_{i+j}).$$

This is illustrated in Figures 1 and 2 for $\omega = 1$. The constant of proportionality ensures that $m_N(x_i) \sim N(0,1)$, though $m_N(x_i)$ is correlated with $m_N(x_j)$ whenever $|x_i - x_j| \leq 2\omega$.

We consider successive refinements of the partition by letting $N \to \infty$ and $\Delta_N \to 0$, thus arriving at a stochastic process defined over a continuum of possible values for x. A single realization of this process associates each $x \in [a, b]$ with a value $m(x) \in \Re^1$. The function can be characterized as

$$m(x) = (2\omega)^{-1/2}[W(x+\omega) - W(x-\omega)] \tag{2.1}$$

for $W(.)$ a standard Wiener process. Note that, like the Wiener process, any given realization of $m(.)$ is continuous in x but not differentiable using standard calculus. It has the further properties that, for any x, $m(x) \sim N(0,1)$ and

$$E[m(x_1)m(x_2)] = \begin{cases} 1 - |x_2 - x_1|/(2\omega) & \text{if } |x_2 - x_1| \leq 2\omega \\ 0 & \text{otherwise} \end{cases}.$$

We find it convenient to normalize the distance parameter $\omega = 1$ and consider the nonlinear component of the mean function $\mu(x)$ to be governed by two other scalar parameters. The first is a constant g which multiplies the value of x, and the second is a constant λ which multiplies the value of $m(.)$:

$$\mu(x) = \alpha_0 + \alpha_1 x + \lambda m(gx). \tag{2.2}$$

Thus, prior to generating any data on y_t or x_t, nature is presumed to have generated a single realization of the stochastic process described by (2.1) for $\omega = 1$, and thus to have settled on a value for $\mu(x)$ for any $x \in [a/g, b/g]$. Finally, nature generates observed values for x_t and y_t according to

$$y_t = \mu(x_t) + \varepsilon_t \tag{2.3}$$

where x_t could be either a lagged value of y_{t-j} or else an exogenous variable that is independent of the realization of the stochastic process $\mu(.)$, and ε_t is i.i.d. with mean zero and independent of both $\mu(.)$ and x_τ for $\tau = t, t-1, ..., 1$.

The parameter λ in (2.2) governs how big a contribution the nonlinear component $m(.)$ makes to the conditional expectation function $\mu(.)$. When $\lambda = 0$, the conditional expectation is linear and (2.3) would describe a standard regression model. The parameter λ also reflects the scale of the dependent variable y_t; a doubling of the units in which y_t is measured would be associated with a doubling in the value of λ. The parameter g in (2.2) governs the curvature of $\mu(x)$:

$$E[\mu(x_t) - \alpha_0 - \alpha_1 x_t][\mu(x_s) - \alpha_0 - \alpha_1 x_s]$$

$$= \begin{cases} 1 - g|x_t - x_s|/2 & \text{if } g|x_t - x_s| \leq 2 \\ 0 & \text{otherwise} \end{cases}.$$

A doubling of the units in which x is measured would be associated with cutting the value of g by 1/2. As $g \to \infty$, the contribution of $m(gx_t)$ to the value of y_t becomes indistinguishable from that of ε_t, while when $g \to 0$, the contribution becomes indistinguishable from that of α_0.

2.2 The case of k explanatory variables

Define a grid in $\Re^k$ by the nodes $\{\mathbf{x}(i_1, i_2, ..., i_k)\}$ where the index $i_j \in \{1, ..., N\}$ for $j = 1, ..., k$ and where, for a given set of indexes $(i_1, i_2, ..., i_k)$, the variable $\mathbf{x}(i_1, i_2, ..., i_k)$ denotes a k-dimensional vector whose jth element is given by

$$x_j(i_1, i_2, ..., i_N) = \begin{cases} a_j + \Delta_{i_j,N}(i_j - 1) & \text{for } i_j = 1, ..., N-1 \\ b_j & \text{for } i_j = N \end{cases}.$$

Let A_N be the set consisting of the N^k distinct points in $\Re^k$ covered by this grid,

$$A_N = \{\mathbf{x}(i_1, i_2, ..., i_k),\ i_j = 1, ..., N,\ j = 1, ..., k\}.$$

For each $\mathbf{x} \in A_N$, let $e(\mathbf{x}) \sim N(0, 1)$ with $e(\mathbf{x})$ independent of $e(\mathbf{z})$ for all $\mathbf{x} \neq \mathbf{z}$. Also associated with each $\mathbf{x} \in A_N$ we define $B_N(\mathbf{x}) \subset A_N$ to be the set of all points in A_N whose distance from $\mathbf{x}$ is less than or equal to unity:

$$B_N(\mathbf{x}) = \{\mathbf{z} \in A_N : (\mathbf{x} - \mathbf{z})'(\mathbf{x} - \mathbf{z}) \leq 1\}.$$

Let $n_N(\mathbf{x})$ denote the number of points in $B_N(\mathbf{x})$. Associated with any point $\mathbf{x}$ in A_N, we then calculate a scalar $m_N(\mathbf{x})$ which is defined as $\sqrt{n_N(\mathbf{x})}$ times the average value of $e(\mathbf{z})$ for all $\mathbf{z}$ contained in $B_N(\mathbf{x})$:

$$m_N(\mathbf{x}) = [n_N(\mathbf{x})]^{-1/2} \sum_{\mathbf{z} \in B_N(\mathbf{x})} e(\mathbf{z}). \tag{2.4}$$

Figure 3 illustrates this for $k = 2$ and $\Delta_{1N} = \Delta_{2N} = 0.5$.

Taking the limit as the partition becomes arbitrarily fine ($\Delta_{i_j,N} \to 0$) gives the probability law for $m(.)$, $m : \mathbf{x} \in \Re^k \to \Re^1$, where $m(.)$ represents a continuous-valued k-dimensional random field. For any $\mathbf{x}$, the scalar $m(\mathbf{x})$ is distributed $N(0, 1)$. For $\mathbf{x}$ and $\mathbf{z}$ arbitrary elements of $\Re^k$, the correlation between $m(\mathbf{x})$ and $m(\mathbf{z})$ is zero if $(\mathbf{x} - \mathbf{z})'(\mathbf{x} - \mathbf{z}) > 2$ and otherwise is given by the ratio of the volume of the overlap of k-dimensional unit spheroids centered at $\mathbf{x}$ and $\mathbf{z}$ to the volume of a single k-dimensional unit spheroid. An expression for this correlation is given in the following results, proved in Appendix A.

Lemma 2.1. Let r and h be scalars satisfying $r \geq h \geq 0$ and define

$$G_k(h, r) = \int_h^r (r^2 - z^2)^{k/2} dz. \tag{2.5}$$

Then $G_k(h, r)$ can be calculated recursively for $k = 2, 3, ...$ as

$$G_k(h, r) = -\frac{h}{1+k}(r^2 - h^2)^{k/2} + \frac{kr^2}{1+k} G_{k-2}(h, r) \tag{2.6}$$

with initial values

$$G_0(h, r) = r - h \tag{2.7}$$

$$G_1(h, r) = (\pi/4)r^2 - (1/2)h(r^2 - h^2)^{1/2} - (r^2/2)\sin^{-1}(h/r) \tag{2.8}$$

where $\theta = \sin^{-1}(\omega)$ indicates that $\theta \in [-\pi/2, \pi/2]$ and $\sin(\theta) = \omega$.

Theorem 2.2. Let $\mathbf{x} \in \Re^k$ and $\mathbf{z} \in \Re^k$ and let $m(\mathbf{x})$ and $m(\mathbf{z})$ be the random field generated as the limit of (2.4) as $\Delta_{i_j,N} \to 0$ evaluated at the fixed points $\mathbf{x}$ and $\mathbf{z}$. Define $h \equiv (1/2)[(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})]^{1/2}$. Then $E[m(\mathbf{x})m(\mathbf{z})] = H_k(h)$ where

$$H_k(h) = \begin{cases} G_{k-1}(h,1)/G_{k-1}(0,1) & \text{if } h \leq 1 \\ 0 & \text{if } h > 1 \end{cases} . \tag{2.9}$$

Closed-form expressions for $H_k(h)$ for $k = 1, ..., 5$ are tabulated in Table 1 for convenience.

Again we view nature as generating a single realization $m(.)$ of this random field, from which a conditional expectation function $\mu(\mathbf{x})$ is determined according to

$$\mu(\mathbf{x}) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x} + \lambda m(\mathbf{g} \odot \mathbf{x}) \tag{2.10}$$

where $\odot$ indicates element-by-element multiplication. Here α_0 and λ are scalars and $\boldsymbol{\alpha}$ and $\mathbf{g}$ are $(k \times 1)$ vectors of population parameters. A zero value for the ith element of $\mathbf{g}$ implies that the conditional expectation function is linear in x_i. Observed data are then viewed as if generated by

$$y_t = \mu(\mathbf{x}_t) + \varepsilon_t \tag{2.11}$$

where $\mathbf{x}_t$ and ε_t are independent of the realization of the random field $m(.)$. Furthermore, ε_t has mean zero and is independent of $\mathbf{x}_t$ and of lagged values of y_{t-j} or $\mathbf{x}_{t-j}$.

3 Inference about the conditional expectation function

3.1 A recursive formulation

Suppose we have observed data on $\{\mathbf{x}_t, y_t\}_{t=1}^T$ generated by (2.10) and (2.11) where $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. The next subsection will explain how to estimate the vector of population parameters, $(\alpha_0, \boldsymbol{\alpha}', \sigma, \mathbf{g}', \lambda)'$. In this subsection, however, we proceed as if these parameters were known with certainty, and the goal is to form an optimal inference about the properties of the unobserved conditional expectation function $\mu(\mathbf{x})$ given the data using an iteration. Like the Kalman filter, the algorithm is a straightforward application of the following well-known result; (see, for example, Hamilton, 1994, p. 102).

Lemma 3.1. If $(\mathbf{y}_1^{'}, \mathbf{y}_2^{'})'$ is multivariate nonsingular Normal with $\mathbf{y}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_{11})$ and $\mathbf{y}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_{22})$ with $\boldsymbol{\Omega}_{12}$ the covariance, then $\mathbf{y}_2|\mathbf{y}_1$ is $N(\mathbf{m}, \mathbf{H})$ where $\mathbf{m} = \boldsymbol{\mu}_2 + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$ and $\mathbf{H} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$.

Consider evaluating the function $\mu(\mathbf{x})$ at a finite set of N values for $\mathbf{x}$ that might be of particular interest, denoted $\mathbf{x} = \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ...,$ or $\boldsymbol{\tau}_N$. Collect these in an $(N \times 1)$ vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\boldsymbol{\tau}_1) \\ \mu(\boldsymbol{\tau}_2) \\ \vdots \\ \mu(\boldsymbol{\tau}_N) \end{bmatrix}. \tag{3.1}$$

Notice from equation (2.10) and Theorem 2.2 that

$$\boldsymbol{\mu} \sim N(\boldsymbol{\xi}_0, \mathbf{P}_0) \tag{3.2}$$

where the ith element of $\boldsymbol{\xi}_0$ is given by $\alpha_0 + \boldsymbol{\alpha}'\boldsymbol{\tau}_i$, while the row i, column j element of $\mathbf{P}_0$ is given by

$$p_{ij}^{(0)} = \begin{cases} \lambda^2 H_k(h_{ij}) & \text{if } h_{ij} < 1 \\ 0 & \text{if } h_{ij} \geq 1 \end{cases} \tag{3.3}$$

for

$$h_{ij} = (1/2)\{[\mathbf{g} \odot (\boldsymbol{\tau}_i - \boldsymbol{\tau}_j)]'[\mathbf{g} \odot (\boldsymbol{\tau}_i - \boldsymbol{\tau}_j)]\}^{1/2} \tag{3.4}$$

and $H_k(h)$ the function given in Theorem 2.2 or Table 1.

It will turn out that a particular value of $\boldsymbol{\tau}_j$ only matters for evaluating the likelihood function if $\boldsymbol{\tau}_j$ corresponds to an observed value of $\mathbf{x}_t$ for some t. For now we simply assume that the grid $\{\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_N\}$ is sufficiently dense that for every value of $\mathbf{x}_t$ that is observed in this particular sample, there exists a grid index j_t such that $\mathbf{x}_t = \boldsymbol{\tau}_{j_t}$. Let $\mathbf{i}_t$ denote column j_t of the $(N \times N)$ identity matrix, so that

$$y_t = \mathbf{i}_t'\boldsymbol{\mu} + \varepsilon_t. \tag{3.5}$$

Notice that (3.2) and (3.5) imply

$$y_1|\mathbf{x}_1 \sim N(\mathbf{i}_1'\boldsymbol{\xi}_0,\ \mathbf{i}_1'\mathbf{P}_0\mathbf{i}_1 + \sigma^2)$$

$$\begin{aligned} \text{Cov}(y_1, \boldsymbol{\mu}'|\mathbf{x}_1) &= E[(y_1 - \mathbf{i}_1'\boldsymbol{\xi}_0)(\boldsymbol{\mu} - \boldsymbol{\xi}_0)'] \\ &= E[(y_1 - \mathbf{i}_1'\boldsymbol{\mu} + \mathbf{i}_1'\boldsymbol{\mu} - \mathbf{i}_1'\boldsymbol{\xi}_0)(\boldsymbol{\mu} - \boldsymbol{\xi}_0)'] \\ &= E[\varepsilon_1(\boldsymbol{\mu} - \boldsymbol{\xi}_0)'] + \mathbf{i}_1'E[(\boldsymbol{\mu} - \boldsymbol{\xi}_0)(\boldsymbol{\mu} - \boldsymbol{\xi}_0)'] \\ &= \mathbf{i}_1'\mathbf{P}_0. \end{aligned}$$

It follows from Lemma 3.1 that[1]

$$\boldsymbol{\mu}|y_1, \mathbf{x}_1 \sim N(\boldsymbol{\xi}_1, \mathbf{P}_1)$$

where

$$\boldsymbol{\xi}_1 = \boldsymbol{\xi}_0 + \frac{\mathbf{P}_0\mathbf{i}_1(y_1 - \mathbf{i}_1'\boldsymbol{\xi}_0)}{\mathbf{i}_1'\mathbf{P}_0\mathbf{i}_1 + \sigma^2} \tag{3.6}$$

$$\mathbf{P}_1 = \mathbf{P}_0 - \frac{\mathbf{P}_0\mathbf{i}_1\mathbf{i}_1'\mathbf{P}_0}{\mathbf{i}_1'\mathbf{P}_0\mathbf{i}_1 + \sigma^2}. \tag{3.7}$$

This same principle can be used to generate a recursion analogous to the Kalman filter. Let the $(N \times 1)$ vector $\boldsymbol{\xi}_t$ represent our inference as to the value of $\boldsymbol{\mu}$ on the basis of observation of $\mathbf{Y}_t = (y_t, \mathbf{x}_t', y_{t-1}, \mathbf{x}_{t-1}', ..., y_1, \mathbf{x}_1')'$ and let the $(N \times N)$ matrix $\mathbf{P}_t$ represent the mean squared error of this inference:

$$\boldsymbol{\xi}_t = E(\boldsymbol{\mu}|\mathbf{Y}_t)$$

$$\mathbf{P}_t = E(\boldsymbol{\mu} - \boldsymbol{\xi}_t)(\boldsymbol{\mu} - \boldsymbol{\xi}_t)'.$$

Suppose that prior to the tth step of the iteration we have established that

$$\boldsymbol{\mu}|\mathbf{Y}_{t-1} \sim N(\boldsymbol{\xi}_{t-1}, \mathbf{P}_{t-1}). \tag{3.8}$$

We assume that $\mathbf{x}_t$ contains no information about the realization of $\mu(.)$ beyond that contained in $\mathbf{Y}_{t-1}$, which would be true if $\mathbf{x}_t$ contains either lagged values of y or variables that are strictly exogenous:

$$\boldsymbol{\mu}|\mathbf{x}_t, \mathbf{Y}_{t-1} \sim N(\boldsymbol{\xi}_{t-1}, \mathbf{P}_{t-1}). \tag{3.9}$$

[1] Here $\mathbf{y}_1 = y_1$, $\mathbf{y}_2 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_1 = \mathbf{i}_1'\boldsymbol{\xi}_0$, $\boldsymbol{\mu}_2 = \boldsymbol{\xi}_0$, $\boldsymbol{\Omega}_{11} = \mathbf{i}_1'\mathbf{P}_0\mathbf{i}_1 + \sigma^2$, $\boldsymbol{\Omega}_{22} = \mathbf{P}_0$, and $\boldsymbol{\Omega}_{12} = \mathbf{i}_1'\mathbf{P}_0$.

It then follows from exactly the same calculations that produced (3.6) and (3.7) that

$$\boldsymbol{\mu}|\mathbf{Y}_t \sim N(\boldsymbol{\xi}_t, \mathbf{P}_t)$$

where

$$\boldsymbol{\xi}_t = \boldsymbol{\xi}_{t-1} + \frac{\mathbf{P}_{t-1}\mathbf{i}_t(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} \tag{3.10}$$

$$\mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\mathbf{i}_t\mathbf{i}_t'\mathbf{P}_{t-1}}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2}. \tag{3.11}$$

Thus one iterates on (3.10) and (3.11) for $t = 1, 2, ..., T$ starting with $\boldsymbol{\xi}_0$ and $\mathbf{P}_0$ as given in (3.2) and (3.3). The end result of this iteration (for $t = T$) is an inference as to the value of $\boldsymbol{\mu}$, that is, an inference as to the value of the conditional mean function $\mu(\mathbf{x})$ evaluated at the set of N particular values of $\mathbf{x}$ represented by the values $\{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_N\}$.

The above calculations require only that each observed $\mathbf{x}_t$ corresponds to some point $\boldsymbol{\tau}_j$ at which the function $\mu(\mathbf{x})$ is to be evaluated. Furthermore, if $\boldsymbol{\tau}_i = \boldsymbol{\tau}_j$ for some i and j, then the ith and jth rows of $\boldsymbol{\xi}_t$ calculated from the above recursion will be identical and will be perfectly correlated with each other through the matrix $\mathbf{P}_t$ for each $t = 0, 1, ..., T$. Thus without loss of generality we can take $N = T$ and define $\boldsymbol{\mu}$ to be the vector $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), ..., \mu(\mathbf{x}_N))'$ so that $\mathbf{i}_t$ in equation (3.5) is the tth column of the $(T \times T)$ identity matrix. In this case the initial conditions for $\boldsymbol{\xi}_0$ and $\mathbf{P}_0$ could be written

$$\boldsymbol{\xi}_0 = \alpha_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix} \boldsymbol{\alpha} \tag{3.12}$$

$$\mathbf{P}_0 = [\lambda^2 H_k(h_{ij})]_{i,j=1,...,T} \tag{3.13}$$

$$h_{ij} = (1/2)\{[\mathbf{g} \odot (\mathbf{x}_i - \mathbf{x}_j)]'[\mathbf{g} \odot (\mathbf{x}_i - \mathbf{x}_j)]\}^{1/2}. \tag{3.14}$$

3.2 Evaluating the likelihood function

It follows from equations (3.5) and (3.9) that

$$y_t|\mathbf{x}_t, \mathbf{Y}_{t-1} \sim N(\mathbf{i}_t'\boldsymbol{\xi}_{t-1}, \mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2). \tag{3.15}$$

Thus iterating on (3.10) and (3.11) allows us to calculate the log of the conditional likelihood of the tth observation from

$$\ln f(y_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \alpha_0, \boldsymbol{\alpha}', \sigma, \mathbf{g}', \lambda) = -(1/2)\ln(2\pi)$$

$$-(1/2)\ln(\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2) - (1/2)\frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2}. \tag{3.16}$$

We can then estimate the value of the vector of unknown population parameters $(\alpha_0, \boldsymbol{\alpha}', \sigma, \mathbf{g}', \lambda)'$ by numerical maximization of

$$\sum_{t=1}^{T} \ln f(y_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \alpha_0, \boldsymbol{\alpha}', \sigma, \mathbf{g}', \lambda). \tag{3.17}$$

3.3 Relation to GLS

We derived the above formulas by considering the distribution of y_t conditional on its own lagged values and on current and past values of $\mathbf{x}_t$. It is possible to perform the identical calculations in a single pass by regarding this as a GLS regression problem. Define

$$\underset{(T\times 1)}{\mathbf{y}} = (y_1, y_2, ..., y_T)'$$

$$\underset{[T\times(k+1)]}{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1' \\ 1 & \mathbf{x}_2' \\ \vdots & \vdots \\ 1 & \mathbf{x}_T' \end{bmatrix}$$

$$\underset{[(k+1)\times 1]}{\boldsymbol{\beta}} = (\alpha_0, \boldsymbol{\alpha}')'$$

$$\underset{(T\times 1)}{\boldsymbol{\varepsilon}} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_T)'.$$

Return for the moment to regarding $\boldsymbol{\mu}$ to be the function $\mu(\mathbf{x})$ evaluated at an arbitrary set of T points, rather than evaluated at observed data. Suppose $\mathbf{X}$ were a deterministic $[T \times (k+1)]$ matrix summarizing these particular points. Since $m(\mathbf{x})$ in (2.10) has unconditional expectation zero for any $\mathbf{x}$, we would in this case regard the vector $\boldsymbol{\mu}$ as having unconditional expectation $\mathbf{X}\boldsymbol{\beta}$ with variance $\mathbf{P}_0$. Since $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, the vectors $\mathbf{y}$ and $\boldsymbol{\mu}$ then have the following unconditional joint distribution:

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\mu} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} (\mathbf{P}_0 + \sigma^2\mathbf{I}_T) & \mathbf{P}_0 \\ \mathbf{P}_0 & \mathbf{P}_0 \end{bmatrix} \right). \tag{3.18}$$

Applying Lemma 3.1 to (3.18) yields immediately

$$\boldsymbol{\mu}|\mathbf{y} \sim \mathbf{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}) \tag{3.19}$$

where

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.20}$$

$$\hat{\mathbf{V}} = \mathbf{P}_0 - \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}\mathbf{P}_0. \tag{3.21}$$

Furthermore, from the first block of (3.18) we would write the unconditional log likelihood of $\mathbf{y}$ as

$$\begin{aligned} \ln f(\mathbf{y}) &= -(T/2)\ln(2\pi) - (1/2)\ln|\mathbf{P}_0 + \sigma^2\mathbf{I}_T| \\ &\quad -(1/2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{3.22}$$

If the explanatory variables $\{\mathbf{x}_t\}_{t=1}^T$ were purely deterministic, then (3.18) is accurate and results (3.19)-(3.22) hold exactly. When $\mathbf{x}_t$ contains lagged values of y_{t-j}, it is not the case that $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, (\mathbf{P}_0 + \sigma^2\mathbf{I}_T))$ as implied by the first block of (3.18). Nevertheless, it turns out that the derived formulas (3.19)-(3.22) are perfectly valid even when $\mathbf{x}_t$ includes lagged values of y_{t-j}, indeed, they are identical to those arrived at by following the recursion proposed in Subsection 3.1. An analogous result occurs in the familiar problem of estimating the parameters of an autoregression, $y_t = \phi y_{t-1} + \varepsilon_t$. If one writes this autoregression in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, even though it is no longer true that $\mathbf{y}|\mathbf{X} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_T)$, the formulas derived for OLS with deterministic regressors also maximize the true conditional log likelihood. The result for the present case is provided by the following theorem.

Theorem 3.2. Let $\boldsymbol{\xi}_T$ and $\mathbf{P}_T$ denote the terminal values resulting from iteration on (3.10) and (3.11) for $t = 1, ..., T$ starting from $\boldsymbol{\xi}_0 = \mathbf{X}\boldsymbol{\beta}$. Then (a) $\mathbf{P}_T$ is numerically identical to $\hat{\mathbf{V}}$ in equation (3.21); (b) $\boldsymbol{\xi}_T$ is numerically identical to $\hat{\boldsymbol{\mu}}$ in equation (3.20); and (c) expression (3.17) and (3.22) are identical.

To take full advantage of the GLS representation, it is convenient to define $\zeta \equiv \lambda/\sigma$ to be the ratio of the standard deviation of the nonlinear component $\lambda m(\mathbf{x})$ to that of the regression residual ε. Let $\boldsymbol{\psi} = (\alpha_0, \boldsymbol{\alpha}', \sigma^2)'$ denote the vector of parameters characterizing

the linear part of the model and let $\boldsymbol{\theta} = (\mathbf{g}', \zeta)'$ denote the nonlinear parameters. For each pair of observations t and s, calculate $\tilde{\mathbf{x}}_t = \mathbf{g} \odot \mathbf{x}_t$ and $h_{ts}(\mathbf{g}) = (1/2)[(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_s)'(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_s)]^{1/2}$. Let $\mathbf{H}(\mathbf{g})$ denote the $(T \times T)$ matrix whose row t, column s element is $H_k(h_{ts}(\mathbf{g}))$ for $H_k(.)$ given by Theorem 2.2 and define

$$\mathbf{W}(\mathbf{X}; \boldsymbol{\theta}) \equiv \zeta^2 \mathbf{H}(\mathbf{g}) + \mathbf{I}_T. \tag{3.23}$$

Note from (3.22) that the log likelihood can be written

$$\begin{aligned} \ln f(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\theta}) &= -(T/2)\ln(2\pi) - (T/2)\ln\sigma^2 - (1/2)\ln|\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})| \\ &\quad -[1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{3.24}$$

For given $\boldsymbol{\theta}$, the value of $\boldsymbol{\psi}$ that maximizes (3.24) can be calculated analytically as

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{X}'\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})^{-1}\mathbf{y}] \tag{3.25}$$

$$\tilde{\sigma}^2(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})^{-1}[\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})]/T. \tag{3.26}$$

This allows us to concentrate the log likelihood (3.22) as

$$\begin{aligned} \eta(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) &= \sum_{t=1}^{T} \ln f(y_t | \mathbf{x}_t, \mathbf{Y}_{t-1}; \tilde{\boldsymbol{\psi}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \\ &= -(T/2)\ln(2\pi) - (T/2)\ln\tilde{\sigma}^2(\boldsymbol{\theta}) - (1/2)\ln|\mathbf{W}(\mathbf{X}; \boldsymbol{\theta})| - (T/2). \end{aligned} \tag{3.27}$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can then be found by maximizing (3.27) with respect to $\boldsymbol{\theta}$ using numerical methods. The maximum likelihood estimate $\hat{\boldsymbol{\psi}}$ is found by plugging this value of $\hat{\boldsymbol{\theta}}$ into (3.25) and (3.26).

3.4 General inference about $\mu(.)$

So far we have discussed forming an inference about the value of $\mu(\mathbf{x})$ evaluated only at those points $\mathbf{x}_t$ observed in the sample. The general framework, however, allows us to make statements about the value of $\mu(\mathbf{x}^*)$ for arbitrary $\mathbf{x}^*$. The simplest derivation of the appropriate formulas follows the deterministic regressor framework of (3.18), though the results can again be shown to be perfectly appropriate for the case of lagged dependent variables as well. Let $\mathbf{y}$ be the $(T \times 1)$ vector of observations on the dependent variable and let $\mathbf{X}$ be a $[T \times (k+1)]$ matrix whose first column contains all ones and whose other columns contain observations on the explanatory variables. Let $\mathbf{X}^* = (1, \mathbf{x}^{*\prime})$ be a $[1 \times (k+1)]$ vector whose first element is unity and whose next k elements are the values $\mathbf{x}^*$ for the explanatory variables at which one would like to evaluate the function $\mu(\mathbf{x}^*)$. Let $\mathbf{q}$ be a $(T \times 1)$ vector whose tth element is the covariance between $\mu(\mathbf{x}^*)$ and $\mu(\mathbf{x}_t)$:

$$q_t = \begin{cases} \lambda^2 H_k(h_t^*) & \text{if } h_t^* \leq 1 \\ 0 & \text{if } h_t^* > 1 \end{cases}$$

where

$$h_t^* = (1/2)[(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}^*)'(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}^*)]^{1/2}$$

$$\tilde{\mathbf{x}}_t = \mathbf{g} \odot \mathbf{x}_t$$

$$\tilde{\mathbf{x}}^* = \mathbf{g} \odot \mathbf{x}^*.$$

We then simply replace $\boldsymbol{\mu}$ in the system (3.18) with $\mu^* = \mu(\mathbf{x}^*)$:

$$\begin{bmatrix} \mathbf{y} \\ \mu^* \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}^*\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} (\mathbf{P}_0 + \sigma^2 \mathbf{I}_T) & \mathbf{q} \\ \mathbf{q}' & \lambda^2 \end{bmatrix} \right)$$

from which it follows as in (3.19)-(3.21) that, given knowledge of the population parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$,

$$\mu^*|\mathbf{y}, \mathbf{X} \sim N(\hat{\mu}^*, \hat{V}^*) \tag{3.28}$$

where

$$\hat{\mu}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{q}'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.29}$$

$$\hat{V}^* = \lambda^2 - \mathbf{q}'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}\mathbf{q}. \tag{3.30}$$

4 Consistent estimation of the conditional mean

We next turn to the behavior of the inference proposed in (3.29) when the sample size T becomes large. It is first interesting to comment on the properties of this algorithm when the explanatory variable $\mathbf{x}_t$ can only take on one of N discrete values.

4.1 Discrete-valued explanatory variables

Theorem 4.1. Suppose that the true data are generated according to

$$y_t = \ell(\mathbf{x}_t) + \varepsilon_t \tag{4.1}$$

where $\{\mathbf{x}_t\}_{t=1}^T$ is a deterministic sequence with $\mathbf{x}_t \in \{\mathbf{x}(1), ..., \mathbf{x}(N)\}$ for all t and where $\ell(\mathbf{x}(1)), ..., \ell(\mathbf{x}(N))$ are N arbitrary numbers. Let $T_i = \sum_{t=1}^T \delta_{\mathbf{x}_t = \mathbf{x}(i)}$ be the number of times that $\mathbf{x}_t$ assumes the value $\mathbf{x}(i)$ within the given sample of size T, so that $T_1 + T_2 + \cdots + T_N = T$. Let $\boldsymbol{\ell}$ be the $(N \times 1)$ vector whose ith element is $\ell(\mathbf{x}(i))$. Let $\boldsymbol{\xi}_0$ be an arbitrary $(N \times 1)$ vector and $\mathbf{P}_0$ an arbitrary $(N \times N)$ positive definite matrix. Let $\hat{\mu}_{it}$ denote the ith element of the vector $\boldsymbol{\xi}_t$ as determined from the recursion (3.10).

(a) If

$$T_i^{-1} \sum_{t=1}^{T} \varepsilon_t \delta_{\mathbf{x}_t = \mathbf{x}(i)} \xrightarrow{p} 0 \tag{4.2}$$

as $T \to \infty$, then

$$\hat{\mu}_{iT} \xrightarrow{p} \ell(\mathbf{x}(i)) \tag{4.3}$$

as $T \to \infty$.

(b) If (4.2) holds for $i = 1, ..., N$, then the log likelihood in (3.17) converges to ℓ^*,

$$\left[\sum_{t=1}^{T} \ln f(y_t | \mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\psi}, \boldsymbol{\theta}) \right] - \ell^* \xrightarrow{p} 0 \tag{4.4}$$

where

$$\begin{aligned} \ell^* &= -(T/2)\ln(2\pi) - [(T-N)/2]\ln(\sigma^2) - (1/2)\sum_{i=1}^{N} \ln(T_i) \\ &\quad -[1/(2\sigma^2)]\sum_{i=1}^{N} T_i s_i^2 - (1/2)\ln|\mathbf{P}_0| - (1/2)(\mathbf{\mu} - \boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}(\mathbf{\mu} - \boldsymbol{\xi}_0) \end{aligned} \tag{4.5}$$

for $s_i^2 = T_i^{-1} \sum_{t=1}^{T} (y_t - h_i)^2 \delta_{\mathbf{x}_t = \mathbf{x}(i)}$ and $h_i = T_i^{-1} \sum_{t=1}^{T} y_t \delta_{\mathbf{x}_t = \mathbf{x}(i)}$.

(c) If (4.2) holds for $i = 1, ..., N$, then the maximum likelihood estimate $\hat{\sigma}^2$ satisfies

$$\hat{\sigma}^2 \xrightarrow{p} T^{-1} \sum_{t=1}^{T} \varepsilon_t^2 \tag{4.6}$$

and the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{g}}$, and $\hat{\lambda}$ are asymptotically equivalent to the values that maximize

$$-(1/2)\ln|\mathbf{P}_0(\mathbf{g},\lambda)| - (1/2)(\mathbf{\mu} - \mathbf{X}\boldsymbol{\beta})'[\mathbf{P}_0(\mathbf{g},\lambda)]^{-1}(\mathbf{\mu} - \mathbf{X}\boldsymbol{\beta}). \tag{4.7}$$

Although the estimator $\hat{\mu}_{iT}$ was motivated by assuming that the error ε_t is i.i.d. Normal, Theorem 4.1 shows that $\hat{\boldsymbol{\xi}}_T$ is consistent for the population mean $\mathbf{\mu}$ under much more general

conditions. For example, ε_t could come from a stationary ARMA process, $\phi(L)\varepsilon_t = \theta(L)a_t$, with roots of $\phi(z) = 0$ outside the unit circle and a_t a non-Gaussian white noise process. All that is required for consistency of $\hat{\mu}_{iT}$ is that the residuals associated with the observations $\mathbf{x}_t = \mathbf{x}(i)$ have population mean zero and obey a law of large numbers (expression (4.2)).

Note further that the estimate $\hat{\boldsymbol{\xi}}_T$ is consistent for *any* values of $\mathbf{P}_0$ and $\boldsymbol{\xi}_0$– any assumed latent process for $\boldsymbol{\mu}$ works equally well. The reason for this result is that the latent process for $\boldsymbol{\mu}$ functions basically as a Bayesian prior for $\boldsymbol{\mu}$. Regardless of the values of the prior, it is eventually dominated by the inference from the observed sample means for any given i.

Result (4.6) establishes that the maximum likelihood estimate $\hat{\sigma}^2$ gives a consistent estimate of the true variance of ε_t. If moreover the true $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$, then

$$\sqrt{T}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{L} N(0, V)$$

where V is consistently estimated from the element of the negative of the inverse of the matrix of second derivatives of (3.17) corresponding to σ^2. In other words, the standard formula for an asymptotic Wad test about a maximum likelihood estimate (e.g., Hamilton, 1994, p. 143) is perfectly appropriate.

By contrast, the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ do not give consistent estimates of any population magnitudes in this case. This is because the process is nonergodic for these parameters. The most the data $\mathbf{Y}_T$ can ever tell us about the process whereby $\boldsymbol{\mu}$ was generated is the realized value of $\boldsymbol{\mu}$ governing the sample. Nevertheless, maximization of the observed log likelihood is a perfectly reasonable way to estimate these parameters. In particular, (4.7) establishes that the MLE $\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to a GLS regression

of the $(N \times 1)$ vector of population means $\mathbf{\mu}$ on $\mathbf{X}\boldsymbol{\beta}$, for $\mathbf{X}$ the $[N \times (k+1)]$ matrix whose ith row is given by $(1,\mathbf{x}(i)')$:

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} (\mathbf{X}'\mathbf{P}_0^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P}_0^{-1}\mathbf{\mu}). \tag{4.8}$$

In other words, $\boldsymbol{\beta}$ is chosen so as to minimize the GLS distance between $\mathbf{X}\boldsymbol{\beta}$ and the true value of the vector of population means, $\mathbf{\mu}$. Furthermore, the information matrix is asymptotically block diagonal between σ^2 and $(\hat{\boldsymbol{\beta}}', \hat{\mathbf{g}}', \hat{\lambda})'$, so that the standard estimate of the variance-covariance matrix of the latter parameters should be a reasonable approximation. In particular, note from (4.7) that if $\mathbf{g}$ and λ were known, the standard formula would imply $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{P}_0^{-1}\mathbf{X})^{-1})$, which for deterministic $\mathbf{x}_t$ and stochastic Gaussian $\mathbf{\mu}$ would in fact be the exact small-sample distribution for a GLS regression of $\mathbf{\mu}$ on $\mathbf{X}\boldsymbol{\beta}$.

4.2 Continuous-valued explanatory variables

We now suppose that $\mathbf{x}_t$ is observed over a continuum in $\Re^k$. A critical condition is that all relevant regions of $\Re^k$ get repeatedly sampled. We state and develop convergence results for the case of deterministic regressors, though comparable results for stochastic regressors are presumably obtainable.

Definition 4.2. Let $A \subset \Re^k$ be a closed rectangular region and let $\{\mathbf{x}_t\}$ be a deterministic sequence in A $\{\mathbf{x}_t \in A \text{ for } t = 1, 2, ...\}$. The sequence is said to be *dense* for A if there exists a continuous function f: $A \to \Re^1$ such that $f(\mathbf{x}) > 0$ for all $\mathbf{x} \in A$ and such that, for any $\varepsilon > 0$ and any continuous function θ: $A \to \Re^1$, there exists an N such that

$$\left| T^{-1} \sum_{t=1}^{T} \theta(\mathbf{x}_t) - \int_A \theta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \right| < \varepsilon \tag{4.9}$$

for all $T \geq N$.

Note that if $\mathbf{x}_t$ were a stochastic i.i.d. sequence with density $f(\mathbf{x})$, then denseness (with deterministic convergence replaced by convergence in probability) would be a simple consequence of the law of large numbers along with a requirement that the density of $\mathbf{x}$ be everywhere positive. However, denseness is a much weaker condition than i.i.d. The essential requirement is that, given any measurable subset of A, one can obtain an arbitrarily large number of observations on $\mathbf{x}_t$ within that subset as the sample size T grows.

In describing the asymptotic properties for the continuous case, it will be helpful to replace the $(T \times 1)$ vector $\boldsymbol{\xi}_t$ in (3.10) with the function $\xi_t : A \to \Re^1$ having the interpretation

$$\xi_t(\mathbf{x}) = E[\mu(\mathbf{x})|\mathbf{Y}_t]. \tag{4.10}$$

We likewise replace the $(T \times T)$ matrix $\mathbf{P}_t$ with the function $p_t : A \times A \to \Re^1$ where

$$p_t(\mathbf{z}, \mathbf{w}) = E[\xi_t(\mathbf{z}) - \mu(\mathbf{z})][\xi_t(\mathbf{w}) - \mu(\mathbf{w})] \tag{4.11}$$

for $\mathbf{z}$ and $\mathbf{w}$ arbitrary elements of A. The recursions (3.10) and (3.11) are then replaced by functional recursions:

$$\xi_t(\mathbf{z}) = \xi_{t-1}(\mathbf{z}) - \frac{p_{t-1}(\mathbf{z}, \mathbf{x}_t)[y_t - \xi_{t-1}(\mathbf{x}_t)]}{p_{t-1}(\mathbf{x}_t, \mathbf{x}_t) + \sigma^2} \tag{4.12}$$

$$p_t(\mathbf{z}, \mathbf{w}) = p_{t-1}(\mathbf{z}, \mathbf{w}) - \frac{p_{t-1}(\mathbf{z}, \mathbf{x}_t)p_{t-1}(\mathbf{x}_t, \mathbf{w})}{p_{t-1}(\mathbf{x}_t, \mathbf{x}_t) + \sigma^2}. \tag{4.13}$$

Notice that if $\xi_{t-1}(\mathbf{z})$ and $p_{t-1}(\mathbf{z}, \mathbf{w})$ are continuous, then so are $\xi_t(\mathbf{z})$ and $p_t(\mathbf{z}, \mathbf{w})$.

Definition 4.3. A continuous function $p : A \times A \to \Re^1$ is said to be positive semidefinite if for any continuous function $\theta : A \to \Re^1$ it is the case that

$$\int_{\mathbf{z} \in A} \int_{\mathbf{w} \in A} \theta(\mathbf{z}) p(\mathbf{z}, \mathbf{w}) \theta(\mathbf{w}) \, d\mathbf{w} \, d\mathbf{z} \geq 0. \tag{4.14}$$

Theorem 4.4. Let $\sigma^2 > 0$ be an arbitrary positive number and let $A \subset \Re^k$ be a closed rectangular region. Let $p_0 : A \times A \to \Re^1$ be an arbitrary initial positive semidefinite function and let $p_t : A \times A \to \Re^1$ be given by (4.13). Suppose that $\{\mathbf{x}_t\}$ is dense for A. then

$$\lim_{T \to \infty} p_T(\mathbf{z}, \mathbf{w}) = 0 \tag{4.15}$$

for all $\mathbf{z}$ and $\mathbf{w}$ in A.

Recall that if the data were really generated from the maintained model with $p_0(\mathbf{z}, \mathbf{w})$ the true covariance of the nonlinear component, then $p_T(\mathbf{x}, \mathbf{x})$ is the MSE of the optimal inference about the unobserved function $\mu(\mathbf{x})$ based on the observed data $\mathbf{Y}_T$. Theorem 4.4 thus implies that the unobserved mean function for this class of processes can be consistently estimated using this algorithm.

It is also easy to show that if the true relation is linear, then the algorithm will consistently uncover this linear relation regardless of the population parameters used to describe the nonlinear portion, that is, regardless of the values used for $\mathbf{P}_0$ and σ^2. The key to this result is the following lemma.

Lemma 4.5. Let $\{\mathbf{x}_t\}$ be dense and let $\mathbf{P}_0$ be a $(T \times T)$ matrix whose row t, column s matrix is given by $p_0(\mathbf{x}_t, \mathbf{x}_s)$ for $p_0 : A \times A \to \Re^1$ a continuous function satisfying (4.14). Let $\mathbf{q}_T$ be a $(T \times 1)$ vector whose tth element is $p_0(\mathbf{x}, \mathbf{x}_t)$ for some $\mathbf{x} \in A$. Let $\mathbf{a} = (a_1, a_2, ..., a_T)'$ where $\{a_t\}$ is a white noise sequence not depending on $\{\mathbf{x}_t\}$:

$$E(a_t a_s) = \begin{cases} \nu^2 & \text{if } t = s \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\lim_{T\to\infty} E\left[\mathbf{q}_T'(\mathbf{P}_0+\sigma^2\mathbf{I}_T)^{-1}\mathbf{a}\right]^2 = 0. \tag{4.16}$$

It follows immediately from Lemma 4.5 that if the true relation is linear,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a},$$

then the inferred value for the conditional expectation of y given $\mathbf{x}^*$ as calculated from (3.29) converges in mean square to $\mathbf{X}^*\boldsymbol{\beta}$,

$$\hat{\mu}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{q}_T'(\mathbf{P}_0+\sigma^2\mathbf{I}_T)^{-1}\mathbf{a} \xrightarrow{m.s.} \mathbf{X}^*\boldsymbol{\beta},$$

regardless of the values used for $\mathbf{P}_0$ and $\sigma^2 > 0$.

We next describe a general class of nonlinear models for which our algorithm would also lead to consistent estimation of the conditional mean. To do so we first introduce the concept of representability.

Definition 4.6. Let A be a closed rectangular region of $\Re^k$ and let $\ell : A \to \Re^1$ and $p : A \times A \to \Re^1$ be arbitrary continuous functions. The function $\ell(.)$ is said to be representable with respect to $p(.,.)$ if there exists a continuous function $\lambda : A \to \Re^1$ such that

$$\ell(\mathbf{x}) = \int_A p(\mathbf{x},\mathbf{z})\lambda(\mathbf{z})\ d\mathbf{z}. \tag{4.17}$$

Theorem 4.7. Let $p_0 : A \times A \to \Re^1$ denote the particular positive semidefinite function from which the iteration on (4.13) is to be started, and let $\theta : A \to \Re^1$ be an arbitrary continuous function. For a given sample of T observations on the explanatory variables

$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ construct the function $\ell_T : A \to \Re^1$ from

$$\ell_T(\mathbf{x}) = T^{-1} \sum_{t=1}^{T} p_0(\mathbf{x}, \mathbf{x}_s)\theta(\mathbf{x}_s). \tag{4.18}$$

Consider a sequence of samples of size $T = 1, 2, ...$ where the sample of size T is generated according to

$$y_t = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_t + \ell_T(\mathbf{x}_t) + a_t \qquad t = 1, 2, ..., T \tag{4.19}$$

where

$$E(a_t a_s) = \begin{cases} \nu^2 & \text{if } t = s \\ \\ 0 & \text{otherwise} \end{cases} . \tag{4.20}$$

Let $\xi_T(\mathbf{z})$ and $p_T(\mathbf{z}, \mathbf{w})$ be the values obtained by iterating on (4.12) and (4.13) where $\sigma^2 > 0$ is an arbitrary constant. If $\{\mathbf{x}_T\}$ is dense for A, then

$$T^{-1} \sum_{t=1}^{T} E\{\xi_T(\mathbf{x}_t) - [\alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_t + \ell_T(\mathbf{x}_t)]\}^2 \to 0 \tag{4.21}$$

as $T \to \infty$.

For given functions $p_0(.,.)$ and $\theta(.)$, expression (4.18) describes a sequence of continuous functions $\ell_T(.)$. The claim of Theorem 4.7 is that the algorithm (4.12)-(4.13) will converge to the limit of this sequence of functions. From (4.9), the sequence of functions has a limiting continuous function described by

$$\lim_{T \to \infty} \ell_T(\mathbf{x}) = \int_A p_0(\mathbf{x}, \mathbf{z})\theta(\mathbf{z}) f(\mathbf{z})\, d\mathbf{z}. \tag{4.22}$$

By varying $\theta(.)$, a class of functions that can be consistently estimated by using this particular $p_0(.,.)$ is thus generated. Comparing (4.22) with (4.17), it appears that if the data were

generated from

$$y_t = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_t + \ell(\mathbf{x}_t) + a_t,$$

then the conditional mean can be consistently estimated provided that the conditional mean function $\ell(\mathbf{x})$ is representable in terms of the $p_0(\mathbf{x}, \mathbf{z})$ covariance function that is used to start the iteration.

For the particular function $p_0(\mathbf{x}, \mathbf{z})$ proposed in (3.3), we obtain the following alternative characterization of representability.

Lemma 4.8. Let $\mathbf{x}$ and $\mathbf{z}$ be elements of $A \subset \Re^k$ and let

$$p_0(\mathbf{x}, \mathbf{z}) = \lambda^2 H_k(h(\mathbf{x}, \mathbf{z}))$$

where $\lambda^2 > 0$ and $H_k(h(\mathbf{x}, \mathbf{z}))$ is the function described in Theorem 2.2 with $h(\mathbf{x}, \mathbf{z}) = (1/2)\{[\mathbf{g} \odot (\mathbf{x} - \mathbf{z})]'[\mathbf{g} \odot (\mathbf{x} - \mathbf{z})]\}^{1/2}$ and $\mathbf{g}$ is a $(k \times 1)$ vector of nonzero constants. Then $\ell : \mathbf{x} \in A \to \Re^1$ is representable with respect to $p_0(\mathbf{x}, \mathbf{z})$ if there exists a continuous function $\eta : \mathbf{z} \in A \to \Re^1$ such that

$$\ell(\mathbf{x}) = \int_{\mathbf{y} \in W(\mathbf{x})} \int_{\mathbf{z} \in W(\mathbf{y}) \cap A} \eta(\mathbf{z}) \; d\mathbf{z} \; d\mathbf{y} \tag{4.23}$$

for $W(\mathbf{x}) = \{\mathbf{y} \in \Re^k\colon \; [\mathbf{g} \odot (\mathbf{x} - \mathbf{y})]'[\mathbf{g} \odot (\mathbf{x} - \mathbf{y})] \leq 1\}$.

Lemma 4.8 suggests that we can think of a representable function $\ell(\mathbf{x})$ as having been arrived at from an underlying continuous function $\eta(\mathbf{z})$ through two steps. First, for any point $\mathbf{y}$ such that the distance between $(\mathbf{g} \odot \mathbf{x})$ and $(\mathbf{g} \odot \mathbf{y})$ is less than unity, we find all the vectors $\mathbf{z}$ that both are within A and are such that the distance between $(\mathbf{g} \odot \mathbf{x})$ and $(\mathbf{g} \odot \mathbf{y})$ is less than unity, and take the average value of $\eta(\mathbf{z})$ for all such values $\mathbf{z}$. Second,

we take an average of the resulting function of $\mathbf{y}$ over all such vectors $\mathbf{y}$ in order to calculate the value $\ell(\mathbf{x})$. By choosing a different function $\eta(\mathbf{z})$ we will arrive through this process at a different function $\ell(\mathbf{x})$ and the set of all such possible functions $\ell(\mathbf{x})$ that could result from starting with any continuous $\eta(\mathbf{z})$ is the set of functions $\ell(\mathbf{x})$ that will be consistently estimated with our procedure.

The set of functions $\ell(\mathbf{x})$ that are representable in terms of (4.23) is a broad and flexible class. For the case of a single explanatory variable with $k = 1$ and $A = [a, b]$, condition (4.23) becomes

$$\ell(x) = \int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=\max\{a,y-g^{-1}\}}^{\min\{b,y+g^{-1}\}} \eta(z)\ dz\ dy. \tag{4.24}$$

Suppose that the true functional form is either an rth-order Taylor series,

$$\ell(x) = \sum_{p=0}^{r} c_p x^p, \tag{4.25}$$

or an rth-order Fourier sine series,

$$\ell(x) = \sum_{p=0}^{r} c_p \sin(\omega_p x). \tag{4.26}$$

For values of x that are far enough from the boundaries, specifically, for values of $x \in [a + 2g^{-1}, b - 2g^{-1}]$, either of these functions are representable by a suitable choice for $\eta(z)$, as the following lemmas demonstrate.

Lemma 4.9. If $\ell(x)$ is given by (4.25), then for any constant $g > 0$, there exists a function $\eta(z) = \sum_{p=0}^{r} \gamma_p z^p$ such that

$$\ell(x) = \int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=y-g^{-1}}^{y+g^{-1}} \eta(z)\ dz\ dy. \tag{4.27}$$

Lemma 4.10. If $\ell(x)$ is given by (4.26), then for any constant $g > 0$ such that $\sin(\omega_p/g) \neq 0$ for $p = 0, 1, ..., r$, then (4.27) holds for the function $\eta(z) = \sum_{p=0}^{r} \gamma_p \sin(\omega_p z)$ where

$$\gamma_p = \frac{c_p \omega_p^2}{4 \sin^2(\omega_p/g)}. \tag{4.28}$$

These results suggest that the class of nonlinear models that one can estimate consistently with this procedure is quite general and flexible. Note moreover that these results hold for any value of the smoothing parameter g. Thus a major advantage of this approach over nonparametric methods is that one does not need to adjust a bandwidth parameter as a function of the sample size; the algorithm given by (4.12) and (4.13) automatically adjusts the inference about $\mu(\mathbf{x})$ as the information from a growing sample accumulates.

Having said this, a few qualifications are in order. First, these lemmas do not imply that the function $\ell(x) = \sum_{p=0}^{r} c_p x^p$ is representable at all $x \in [a, b]$, but only that the representability condition (4.24) holds for interior points $x \in [a + 2g^{-1}, b - 2g^{-1}]$. The actual result proven is that the conditional mean could be consistently estimated if it takes the form of $\ell(x) = \sum_{p=0}^{r} c_p x^p$ for $x \in [a + 2g^{-1}, b - 2g^{-1}]$ but has a different characterization near the boundaries, namely

$$\ell(x) = \int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=\max\{a,y-g^{-1}\}}^{\min\{b,y+g^{-1}\}} \sum_{p=0}^{r} \gamma_p z^p \, dz \, dy$$

where the coefficients γ_p are given in the proof of Lemma 4.9. If instead $\ell(x)$ took the form of $\sum_{p=0}^{r} c_p x^p$ for all x, then there is no guarantee that our algorithm will provide consistent estimates of the conditional mean $\ell(x)$ for values of x less than $a + 2g^{-1}$ or greater than $b - 2g^{-1}$.

Second, no claim has been made about the rate of convergence. For example, suppose that the true function is $\ell(x) = \sin(\omega x)$. If $g = \omega/\pi$, then the region over which one is averaging in (4.27) (namely, $x \pm g^{-1}$) would be exactly the same as the interval needed for the functional form to complete a cycle ($x \pm \pi/\omega$), and the condition for representability in Lemma 4.10 ($\sin(\omega/g) \neq 0$) would fail to hold. Suppose instead that $\sin(\omega/g)$ is close but not equal to 0. The function $\eta(z)$ of which Lemma 4.10 demonstrates the existence is given by $\{\omega/[2\sin(\omega/g)]\}^2 \sin(\omega z)$, which could assume quite large values at its peaks and troughs if $\sin(\omega/g)$ is near zero. Suppose for illustration that x is uniformly distributed over $[a, b]$, so that $f(z) = (b-a)^{-1}$. Then the function $\theta(z)$ in (4.22) is proportional to $\eta(z)$. The proof of Theorem 4.7 is based on the assumption that there exists a sample size T such that $T^{-1}\sum_{t=1}^{T} p_T(x_t, x_s)\theta(x_s)$ is negligible, where $p_T(x_t, x_s)$ is the result of T iterations on (4.13). as $\sin(\omega/g)$ approaches 0, the magnitude of $\theta(x_s)$ becomes larger and the necessary value of T becomes bigger. Obviously for a given finite sample, looking at averages over a region that roughly corresponds to the periodicity of the functional form is not going to be a very good way to find out about the function.

More generally, the function $\theta(\mathbf{x})$ will be proportional to $\eta(\mathbf{x})/f(\mathbf{x})$ where $f(.)$ is the density of the independent variable $\mathbf{x}_t$. If some region of the $\mathbf{x}$-space is sparsely sampled (so that $f(\mathbf{x})$ is small), a large number of observations will be necessary in order to estimate the value of the function in that region, since again a large value of T will be needed to make $T^{-1}\sum_{t=1}^{T} p_T(\mathbf{x}_t, \mathbf{x}_s)\theta(\mathbf{x}_s)$ small. Obviously the curse of dimensionality applies as well; the larger k, the larger T must be to ensure adequate coverage of any given neighborhood.

Finally, we note the critical role of the assumption that $\{\mathbf{x}_t\}$ is a dense sequence for a compact region of support A. It might be technically possible to relax the assumption of conpactness through the device of allowing the boundaries of the region to grow at the proper rate as the sample size increases. However, the theory that any nonzero value of $\mathbf{g}$ would work is based on the assumption that, for any $\mathbf{x}^*$ of interest, one has an arbitrarily large number of observations such that $\sum_{i=1}^{k} g_i^2(x_{it} - x_i^*)^2 < 1$. For any given finite sample, there exists a value for $\mathbf{g}$ sufficiently large that *no* observations fall in this region, in which case the expectation that we *would* have an infinite number of such observations in an infinite sample offers little practical comfort. On the other hand, as $\mathbf{g}$ becomes smaller, more observations will be included in the averaging region, but one would have growing concerns that the function is roughly periodic over such regions, that convergence toward a given polynomial will be slower, and that a larger number of observations will fall near the boundaries for which the estimates may be unreliable.

The approach suggested in this paper of choosing $\mathbf{g}$ and λ so as to maximize the likelihood function (3.22) or by the Bayesian methods suggested in the following section seems a sensible way of dealing with these issues. When $\mathbf{g} \to \infty$, the nonlinear component $\mathbf{P}_0$ becomes $\lambda^2\mathbf{I}_T$ which is indistinguishable from the disturbance covariance $\sigma^2\mathbf{I}_T$, and the value achieved for the likelihood function would be identical to that of a simple linear regression. Likewise, when $\mathbf{g} \to \mathbf{0}$, the nonlinear component $\mathbf{P}_0$ becomes $\lambda^2\mathbf{1}\mathbf{1}'$, which makes a contribution to the likelihood function identical to that of the constant term in $\mathbf{X}\boldsymbol{\beta}$, and the fit achieved would again be no better than that for the linear regression. Treating these as population

parameters to be estimated by maximum likelihood or Bayesian methods thus avoids both extremes by construction, and uses values that are most appropriate given such evidence of nonlinearity as appears in the data.

Ultimately, the question of whether the convergence rate is satisfactory depends on the practical experience of applied users. On this score, the examples presented in Section 6 of this paper and the extensive Monte Carlo investigations by Dahl (1998) suggest that the method holds a great deal of promise.

5 Bayesian analysis

One benefit of the parametric approach to flexible inference is that it allows one to evaluate the small-sample properties of any statistic of interest using numerical Bayesian methods, as described in this section.

5.1 Priors

As in Subsection 3.3, let $\boldsymbol{\psi} = (\boldsymbol{\beta}', \sigma^{-2})'$ denote the vector of parameters for the linear part of the model and $\boldsymbol{\theta} = (\mathbf{g}', \zeta)'$. The algorithms described below require using nondiffuse prior distributions in order to be valid.[2] We adopt a standard prior for the linear components; (see for example DeGroot, 1970, p. 251). Specifically, we employ a gamma prior for σ^{-2}:

$$p(\sigma^{-2}) = \frac{\xi^{\nu}}{\Gamma(\nu)} \sigma^{-2(\nu-1)} \exp[-\xi\sigma^{-2}]. \tag{5.1}$$

[2] The elements of $\boldsymbol{\theta}$ become unidentified (have no marginal effect on the likelihood) as $g_i \to \infty$, and hence a nondiffuse prior is necessary in order for the posterior distribution to be well-defined. The calculations leading to (5.8) below also require a nondiffuse prior for $\boldsymbol{\beta}$.

The applications presented below used $\nu = 0.25$ and $\xi = (\nu s_y^2/2)$ for s_y^2 the sample variance of y. These values imply that $E(\sigma^{-2}) = 1/(s_y^2/2)$, meaning the analyst anticipates a residual variance that is about (1/2) the variance of the original series, and puts a weight on this prior equivalent to one-half of an observation on ε^2 itself.

We suppose that the prior distribution of $\boldsymbol{\beta}$ conditional on σ^{-2} is Gaussian:

$$p(\boldsymbol{\beta}|\sigma^{-2}) = \frac{1}{(2\pi\sigma^2)^{(k+1)/2}}|\mathbf{M}|^{-1/2}\exp\left[\left(\frac{-1}{2\sigma^2}\right)(\boldsymbol{\beta}-\mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta}-\mathbf{m})\right]. \tag{5.2}$$

The applications below set the first element of $\mathbf{m}$ to the sample mean of y_t and all other elements of $\mathbf{m}$ to zero. We further follow Jeffreys' invariance principle[3] in taking $\mathbf{M} = T(\mathbf{X}'\mathbf{X})^{-1}$, so that the prior has the weight of a single observation on $(y_t, \mathbf{x}_t')$.

We use a lognormal prior for each element of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{k+1}\frac{1}{\sqrt{2\pi}\tau_i\theta_i}\exp\left[\frac{-[\ln(\theta_i)-\vartheta_i]^2}{2\tau_i^2}\right]. \tag{5.3}$$

Note that the prior for θ_i is taken to be independent of that for $\boldsymbol{\psi}$ and θ_j, $j \neq i$. We specify $\tau_i = 1$ and, for $i = 1, ..., k$, we take ϑ_i to depend on the standard deviation of variable i,

$$\vartheta_i = -\ln\left(\sqrt{ks_i^2}\right) \tag{5.4}$$

with $s_i^2 = T^{-1}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2$ and $\bar{x}_i$ the sample mean of the ith explanatory variable. The rationale for these choices is as follows. This prior implies that with 95% probability, $\ln(g_i) + \ln\left(\sqrt{ks_i^2}\right)$ falls in the interval [-2,+2], or that g_i falls in the interval $\left[e^{-2} \div \left(\sqrt{ks_i^2}\right), e^2 \div \left(\sqrt{ks_i^2}\right)\right]$. Note that at the lower bound, the argument of the function $H_k(h_{ts})$ in (3.14) would be

$$h_{ts} = (1/2)[g_1^2(x_{1t}-x_{1s})^2 + g_2^2(x_{2t}-x_{2s})^2 + \cdots + g_k^2(x_{kt}-x_{ks})^2]^{1/2}$$

[3] For discussion see Jeffreys (1961, p. 179), Zellner (1971, p. 47), or Phillips (1991, p. 342).

$$= \left[(e^{-4}/4)\sum_{\ell=1}^{k}\frac{(x_{\ell t}-x_{\ell s})^2}{ks_\ell^2}\right]^{1/2}.$$

Since $e^{-4}/4 \simeq (1/15)^2$, expression (5.4) implies that $\mu(\mathbf{x}_t)$ is uncorrelated with $\mu(\mathbf{x}_s)$ if each element of $\mathbf{x}_t$ differs from the corresponding element of $\mathbf{x}_s$ by more than 15 standard deviations; in other words, the function $\mu(\mathbf{x})$ is essentially a constant with respect to x_ℓ when g_ℓ is at this lower bound. At the upper bound, $e^4/4 \simeq 4^2$, implying that that $\mu(\mathbf{x}_t)$ is uncorrelated with $\mu(\mathbf{x}_s)$ if each element of $\mathbf{x}_t$ differs from the corresponding element of $\mathbf{x}_s$ by more than 1/4 of a standard deviation. Values of g_i above or below these bounds are regarded as unlikely a priori. We take $\vartheta_{k+1} = 0$, implying a prior distribution for ζ with mode at unity and 95% confidence interval (0.14, 7.4).

5.2 A useful decomposition of the posterior distribution

Continuing to regard $\mathbf{X}$ as deterministic, the joint distribution of $\mathbf{y}, \boldsymbol{\theta}$, and $\boldsymbol{\psi}$ is obtained from the product of (3.24), (5.1), (5.2), and (5.3):

$$f(\boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{y}|\mathbf{X}) = f(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{X}) \cdot p(\boldsymbol{\psi}) \cdot p(\boldsymbol{\theta}). \tag{5.5}$$

The product of the first two terms in (5.5) has the following well-known structure.

Lemma 5.1. The product $f(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{X}) \cdot p(\boldsymbol{\psi})$ can be factored as

$$f(\mathbf{y}, \boldsymbol{\psi}|\boldsymbol{\theta}, \mathbf{X}) = f(\boldsymbol{\beta}|\boldsymbol{\sigma}^{-2}, \mathbf{Y}_T, \boldsymbol{\theta}) \cdot f(\sigma^{-2}|\mathbf{Y}_T, \boldsymbol{\theta}) \cdot f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$$

where: (i) the posterior conditional distribution of $\boldsymbol{\beta}$ is $N(\mathbf{m}^*, \mathbf{M}^*)$,

$$f(\boldsymbol{\beta}|\boldsymbol{\sigma}^{-2}, \mathbf{Y}_T, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{(k+1)/2}}|\mathbf{M}^*|^{-1/2}\exp\left[\left(\frac{-1}{2\sigma^2}\right)(\boldsymbol{\beta}-\mathbf{m}^*)'\mathbf{M}^{*-1}(\boldsymbol{\beta}-\mathbf{m}^*)\right] \tag{5.6}$$

for $\mathbf{M}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$ and $\mathbf{m}^* = \mathbf{M}^*(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{W}^{-1}\mathbf{y})$; (ii) the posterior conditional distribution of σ^{-2} is $\Gamma(\nu^*, \xi^*)$,

$$f(\sigma^{-2}|\mathbf{Y}_T, \boldsymbol{\theta}) = \frac{\xi^{*\nu^*}}{\Gamma(\nu^*)}\sigma^{-2(\nu^*-1)}\exp[-\xi^*\sigma^{-2}] \tag{5.7}$$

for $\nu^* = \nu + (T/2)$ and

$$\xi^* = \xi + (1/2)(\mathbf{y} - \mathbf{X}\mathbf{m})'[\mathbf{W}(\mathbf{X};\boldsymbol{\theta}) + \mathbf{X}\mathbf{M}\mathbf{X}']^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m});$$

(iii) the marginal distribution of $\mathbf{y}$ is multivariate Student t:

$$f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) = \frac{\Gamma(\nu^*)\xi^{\nu}}{(2\pi)^{T/2}\Gamma(\nu)\xi^{*\nu^*}}\left|\mathbf{W}(\mathbf{X};\boldsymbol{\theta}) + \mathbf{X}\mathbf{M}\mathbf{X}'\right|^{-1/2}. \tag{5.8}$$

One implication of Lemma 5.1 is that the joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ is known analytically from the product of (5.3) and (5.8):

$$f(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X}) = f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\cdot p(\boldsymbol{\theta}). \tag{5.9}$$

5.3 Importance sampling

The goal in this subsection is to infer the posterior expected value of some function $\ell(\boldsymbol{\theta})$ of the nonlinear parameters,

$$E[\ell(\boldsymbol{\theta})|\mathbf{Y}_T] = \int \ell(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{Y}_T)\, d\boldsymbol{\theta} \tag{5.10}$$

where

$$f(\boldsymbol{\theta}|\mathbf{Y}_T) = \frac{f(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})}{\int f(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})\, d\boldsymbol{\theta}}.$$

Following Geweke (1989), one can in principle infer the value of (5.10) with any desired accuracy by generating an artificial i.i.d. sample $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$ drawn from an essentially

arbitrary "importance" density $I(\boldsymbol{\theta})$ and calculating the value of

$$\frac{\sum_{j=1}^{N} \ell(\boldsymbol{\theta}^{(j)}) w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}{\sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)} \tag{5.11}$$

where

$$w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) = \frac{f(\boldsymbol{\theta}^{(j)}, \mathbf{y}|\mathbf{X})}{I(\boldsymbol{\theta}^{(j)})}. \tag{5.12}$$

The key property that $I(\boldsymbol{\theta})$ must satisfy in order for the estimate (5.11) to be acceptably close to (5.10) for a feasible value of N is that $w(\boldsymbol{\theta}, \mathbf{Y}_T)$ approaches zero for large or small values of $\boldsymbol{\theta}$. One can show that in the tails, (5.9) is dominated by the prior $p(\boldsymbol{\theta})$. Hence the key requirement is that the importance density should be more spread out than the prior. We have developed a reasonably efficient algorithm based on a truncated mixture density. With probability 0.5, we generate $\boldsymbol{\theta}$ from a multivariate Student t distribution with $\varphi = 2$ degrees of freedom, centered at the maximum likelihood estimate, and with precision matrix given by (-1/2) times the matrix of second derivatives of the log likelihood function, in other words, a distribution similar to the assumed asymptotic distribution of the MLE, though more spread out. With probability 0.5, the elements of $\ln(\boldsymbol{\theta})$ are drawn independently from $N(\vartheta_i, 4)$ distributions, so that the logs have the same mean but twice the variance as specified by the prior. The truncation was achieved by throwing out any draw for which some $\theta_i < 0$. Thus the importance density is proportional to

$$\begin{aligned} I(\boldsymbol{\theta}) \quad \propto \quad & (0.5)\frac{\Gamma[(k+1+\varphi)/2]}{\Gamma(\varphi/2)(\varphi\pi)^{(k+1)/2}} \left|\hat{\boldsymbol{\Omega}}\right|^{-1/2} \left[1 + \varphi^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \hat{\boldsymbol{\Omega}}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right]^{-(k+1+\varphi)/2} \\ & + (0.5) \prod_{i=1}^{k+1} \frac{1}{\sqrt{2\pi}(2\tau_i)\theta_i} \exp\left[\frac{-[\ln(\theta_i) - \vartheta_i]^2}{2(2\tau_i)^2}\right] \end{aligned} \tag{5.13}$$

$$\text{for } \theta_i \geq 0, i = 1, 2, ..., k+1$$

where the constant of proportionality reflects the truncation, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate, $\hat{\boldsymbol{\Omega}}$ is twice its asymptotic variance matrix, $\varphi = 2, \tau_i = 1$, $\vartheta_{k+1} = 0$, and ϑ_i is given by (5.4) for $i = 1, ..., k$. Other values for any of these parameters that characterize the importance density would have worked equally well; these values were chosen so as to satisfy the tails condition and still have a reasonable concentration of mass in the same region as does the unknown $f(\boldsymbol{\theta}|\mathbf{Y}_T)$.

5.4 Confidence intervals

Let $\boldsymbol{\zeta}$ be any random vector whose distribution conditional on $\boldsymbol{\theta}$ and $\mathbf{Y}_T$ is known ($f(\boldsymbol{\zeta}|\boldsymbol{\theta}, \mathbf{Y}_T)$). As shown in Appendix D, we can estimate the posterior probability that $\boldsymbol{\zeta}$ falls in some region C as follows. For each $\boldsymbol{\theta}^{(j)}$ generated from the importance density in (5.13), generate $\boldsymbol{\zeta}^{(j)}$ from $f(\boldsymbol{\zeta}|\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)$ and calculate

$$\widehat{\Pr}(\boldsymbol{\zeta} \in C|\mathbf{Y}_T) = \frac{\sum_{j=1}^{N} \delta_{[\boldsymbol{\zeta}^{(j)} \in C]} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}{\sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}. \tag{5.14}$$

Thus for example one can obtain the posterior mean of the vector of linear parameters $\boldsymbol{\psi}$ by generating $\boldsymbol{\psi}^{(j)}$ from $f(\boldsymbol{\psi}|\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)$ in (5.6) and (5.7) and calculating

$$\hat{E}(\boldsymbol{\psi}|\mathbf{Y}_T) = \frac{\sum_{j=1}^{N} \boldsymbol{\psi}^{(j)} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}{\sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}. \tag{5.15}$$

Similarly, we know from (4.12) and (4.13) that, conditional on $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$,

$$\mu(\mathbf{x})|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{Y}_T \sim N\left(\xi_T(\mathbf{x}|\boldsymbol{\psi}, \boldsymbol{\theta}), p_T(\mathbf{x}, \mathbf{x}|\boldsymbol{\psi}, \boldsymbol{\theta})\right). \tag{5.16}$$

Hence the posterior mean can be calculated from

$$\hat{E}[\mu(\mathbf{x})|\mathbf{Y}_T] = \frac{\sum_{j=1}^{N} \xi_T(\mathbf{x}|\boldsymbol{\psi}^{(j)}, \boldsymbol{\theta}^{(j)}) w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}{\sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T)}. \tag{5.17}$$

A $100(1-\alpha)\%$ confidence interval can be found as follows. Generate $\boldsymbol{\theta}^{(j)}$ from (5.13) and use this $\boldsymbol{\theta}^{(j)}$ to generate $\boldsymbol{\psi}^{(j)}$ from (5.6) and (5.7). From these then generate a scalar $\zeta^{(j)}$ from the density in (5.16). Sort the values of $\{\zeta^{(1)}, ..., \zeta^{(N)}\}$ from smallest to largest and find the observation indexes j_1 and j_2 for which $\sum_{j=1}^{j_1} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) \div \sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) = \alpha/2$ and $\sum_{j=j_2}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) \div \sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) = \alpha/2$. The confidence interval is then $[\zeta^{(j_1)}, \zeta^{(j_2)}]$.

5.5 Limitations of the parametric approach

Conditional on the data and the value of the population parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, the inference $\xi_T(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta})$ determined by (4.12) is the optimal estimate of $\mu(\mathbf{x})$ for a quadratic loss function. Likewise, with the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ stochastic but governed by the prior distributions (5.1)-(5.3), then the limit of (5.17) as $N \to \infty$ would be the optimal estimate of $\mu(\mathbf{x})$. These optimality properties, however, hold only if the data were truly generated from the model specified by (2.4), (2.10), and (2.11). Although the estimate $\xi_T(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta})$ converges to the true $\mu(\mathbf{x})$ even if (2.4) and (2.10) do not hold, we can not claim that $\xi_T(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta})$ would be the optimal estimate no matter what the true $\mu(.)$.

Indeed, the very definition of optimality over some broad class of possible $\mu(.)$ is problematic. The conventional nonparametric approach focuses on the rate of convergence for the most troubling possibility within the class of allowable $\mu(.)$. This min-max formulation of the problem, however, is not one that arises naturally from thinking about the researcher's ultimate objectives. For example, if what one really cares about is minimizing the value of $[\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})]^2$, then under the Friedman and Savage (1948) postulates, the optimal strategy is to place probability weights on all the possible values for $\mu(\mathbf{x})$ and integrate. The priors

for $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ imply a prior distribution for possible $\mu(.)$, and if one accepts this implicit assignment of prior probabilities, then the estimator (5.17) would be strictly superior to min-max methods for choosing an implicit bandwidth. Of course, as is a potential problem for any Bayesian estimator, not all researchers will embrace this particular prior.

Similar caveats apply to the confidence intervals generated by this approach. We claim to have a procedure that consistently estimates the value of $\mu(\mathbf{x})$ for $\mu(.)$ any function within a broad class, and claim to have small-sample 95% confidence intervals for the inference. However, we can not claim, as a classical econometrician might wish, that for any particular function $\mu_0(.)$ within this class, if the data were truly generated from $y_t = \mu_0(\mathbf{x}_t) + \varepsilon_t$, then the confidence intervals for $\mu(\mathbf{x})$ would include the true value $\mu_0(\mathbf{x})$ in 95% of the samples. Rather, the claim is that, if one assigns prior probabilities to various possible values for $\mu(.)$ as in (5.1)-(5.3), then the posterior probability that $\mu(\mathbf{x})$ falls within the calculated band is 95%.

There is, however, one hypothesis of special interest for which such qualifications and caveats do not apply. This is the hypothesis that the true relation is linear, to which test we now turn.

6 Testing for nonlinearity

One advantage of having an explicit parametric model of general nonlinearity is that it suggests a simple way to test the null hypothesis that the true relation is linear, namely, by testing whether the value of λ^2 in equation (3.13) is zero. Admittedly, if the true

relation is linear, then the parameters $\mathbf{g}$ that govern the scale of the nonlinearity in (3.14) are unidentified. Fortunately, it is quite natural for purposes of testing the null hypothesis of linearity simply to fix these values on the basis of the scale of the data, for example, by setting g_i equal to the mean of the prior distribution in (5.3).

Theorem 6.1. Let $\mathbf{H}_T$ be a known $(T \times T)$ positive semidefinite matrix and let

$$\mathbf{\Omega}_T = \lambda^2 \mathbf{H}_T + \sigma^2 \mathbf{I}_T. \tag{6.1}$$

Consider the likelihood function under the assumption that $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Omega}_T)$:

$$\ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\zeta}) = -(T/2)\ln(2\pi) - (1/2)\ln|\mathbf{\Omega}_T| - (1/2)\mathrm{tr}(\mathbf{\Omega}_T^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \tag{6.2}$$

for $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\zeta} = (\lambda^2, \sigma^2, \boldsymbol{\beta}')'$.

(a) The score is given by

$$\left.\frac{\partial \ \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\zeta})}{\partial \lambda^2}\right|_{\lambda^2=0} = (2\sigma^4)^{-1}[\boldsymbol{\varepsilon}'\mathbf{H}_T\boldsymbol{\varepsilon} - \sigma^2\mathrm{tr}(\mathbf{H}_T)]. \tag{6.3}$$

(b) Suppose that the data were actually generated from (6.2) with some true parameter vector $\boldsymbol{\zeta}_0 = (0, \sigma_0^2, \boldsymbol{\beta}_0')'$. Then the score in (6.3) evaluated at $\boldsymbol{\zeta}_0$ has expectation zero, and the information matrix is

$$\begin{aligned} & -E\left\{\left.\frac{\partial^2 \ \ln f(\mathbf{y}|\mathbf{X}; \boldsymbol{\psi}, \lambda^2)}{\partial \boldsymbol{\zeta}\, \partial \boldsymbol{\zeta}'}\right|_{\zeta=\zeta_0}\right\} \\ = & \begin{bmatrix} (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T^2) & (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T) & \mathbf{0}' \\ (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T) & (2\sigma^4)^{-1}T & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \sigma^{-2}\mathbf{X}'\mathbf{X} \end{bmatrix}. \end{aligned} \tag{6.4}$$

(c) The Lagrange multiplier test of the null hypothesis that $\lambda^2 = 0$ is given by

$$\aleph_T = \frac{\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \hat{\sigma}_T^2 \mathrm{tr}(\mathbf{H}_T)}{\hat{\sigma}_T^2\sqrt{2}\left\{\mathrm{tr}(\mathbf{H}_T^2) - T^{-1}\left[\mathrm{tr}(\mathbf{H}_T)\right]^2\right\}^{1/2}} \tag{6.5}$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\hat{\sigma}_T^2 = T^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

(d) Suppose that the data were actually generated from (6.2) with $\lambda^2 = 0$. Let $\mathbf{H}_T$ be a $(T \times T)$ positive semidefinite matrix whose row t, column s element is given by some function $h(\mathbf{x}_t, \mathbf{x}_s)$ where diagonal elements $h(\mathbf{x}_t, \mathbf{x}_t)$ are all unity. Define

$$\mathbf{A}_T \equiv \mathbf{H}_T - T^{-1}[\mathrm{tr}(\mathbf{H}_T)]\mathbf{I}_T \tag{6.6}$$

so that $\mathbf{A}_T$ is the same as $\mathbf{H}_T$ except that diagonal elements are all zero. Let $\{\mathbf{x}_t\}$ be a deterministic sequence where $h(.,.)$ and $\{\mathbf{x}_t\}$ satisfy

(i) $T^{-1}\mathbf{X}'\mathbf{X} \to \mathbf{Q}$, a positive definite matrix;

(ii) $T^{-1}\mathbf{X}'\mathbf{A}_T\mathbf{X} \to \mathbf{P}$, a positive semidefinite matrix;

(iii) $T^{-1}\mathbf{X}'\mathbf{A}_T^2\mathbf{X} \to \mathbf{R}$, a positive semidefinite matrix;

(iv) $\exists \delta > 0$ such that $T^{-1}\mathrm{tr}(\mathbf{A}_T^2) > \delta \ \forall T$.

Then as $T \to \infty$, $\aleph_T \xrightarrow{L} N(0,1)$.

The LM test statistic (6.5) has a small-sample bias which can be corrected using the following result.

Lemma 6.2. For $\mathbf{x}_t$ deterministic, the following score-based sample statistic,

$$\tilde{\aleph}_T = \frac{\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \tilde{\sigma}_T^2 \mathrm{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T)}{\left(2 \ \mathrm{tr}\left\{\left[\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T - (T-k-1)^{-1}\mathbf{M}_T\mathrm{tr}\left(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T\right)\right]^2\right\}\right)^{1/2}}, \tag{6.7}$$

has mean zero and variance σ_0^4, where $\mathbf{M}_T = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\tilde{\sigma}_T^2 = (T-k-1)^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

The recommended test of the null hypothesis that the true relation is a linear model of the form $y_t = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_t$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\sigma^2\mathbf{I}_T)$ against the alternative that $y_t = \mu(\mathbf{x}_t) + \varepsilon_t$ for $\mu(.)$ given by (2.10) is thus conducted as follows. Set $g_i = 2 \div \sqrt{ks_i^2}$ for s_i the standard deviation of variable i and k the number of explanatory variables, not counting the constant term.[4] Calculate the $(T \times T)$ matrix $\mathbf{H}$ whose row t, column s element is given by

$$H_k\left((1/2)\left[g_1^2(x_{1t} - x_{1s})^2 + g_2^2(x_{2t} - x_{2s})^2 + \cdots + g_k^2(x_{kt} - x_{ks})^2\right]^{1/2}\right) \tag{6.8}$$

for $H_k(h)$ the function given in Theorem 2.2 or Table 1. Next perform a linear OLS regression of y_t on $(1, \mathbf{x}_t')'$ with the usual $(T \times 1)$ residual vector $\hat{\boldsymbol{\varepsilon}}$, OLS squared standard error $\tilde{\sigma}^2 = (T - k - 1)^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, and $(T \times T)$ projection matrix $\mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for $\mathbf{X}$ the $[T \times (k+1)]$ matrix whose tth row is $(1, \mathbf{x}_t')$. Finally, calculate the value of

$$\nu^2 = \frac{\left[\hat{\boldsymbol{\varepsilon}}'\mathbf{H}\hat{\boldsymbol{\varepsilon}} - \tilde{\sigma}^2\text{tr}(\mathbf{MHM})\right]^2}{\tilde{\sigma}^4\left(2\text{tr}\left\{[\mathbf{MHM} - (T-k-1)^{-1}\mathbf{M}\text{tr}(\mathbf{MHM})]^2\right\}\right)}. \tag{6.9}$$

If $\nu^2 > 3.84$, the null hypothesis of linearity should be rejected at the 5% level.

Note that a further implication of Theorem 6.1 is that, if the true model is linear ($\lambda_0^2 = 0$) and one estimates a model of the class suggested here, then for fixed $\mathbf{g}$, the estimate $\hat{\lambda}^2$ is consistent for the true value of zero and, if not constrained to be nonnegative, $\hat{\lambda}^2$ is asymptotically Normal.

---

[4] This value for g_i is approximately the mean of the prior distribution in (5.3); the actual mean is $1.65 \div \sqrt{ks_i^2}$. This value for g_i implies that if each element of $\mathbf{x}_t$ differs from the corresponding element of $\mathbf{x}_s$ by one standard deviation, then $\mu(\mathbf{x}_t)$ is independent of $\mu(\mathbf{x}_s)$.

7 Illustrations.

7.1 Example 1

We generated $T = 100$ observations from the following threshold regression model,

$$y_t = 0.6x_{1t}\delta_{[x_{1t}>0]} + 0.2x_{2t} + \varepsilon_t \tag{7.1}$$

where $x_{it} \sim N(0, 100)$ and $\varepsilon_t \sim N(0, 1)$. Thus the true model is linear in x_2 and nonlinear in x_1. The nonlinearity is quite dramatic for this example; correct specification of the nonlinearity would produce a large improvement in the R^2 relative to a linear model. Not surprisingly, the LM test of the null hypothesis that the true relation is linear (expression (6.9)) yields a $\chi^2(1)$ test statistic of 232.93, so this test would leave the researcher no doubt that a nonlinear specification is needed. The challenge is to let the data guide us to choose the particular nonlinear form (7.1) if we know nothing a priori about the process.

Maximum likelihood estimates under the (false) assumption that the data were generated from (2.10) and (2.11) are as follows

$$\begin{aligned} y_t &= \underset{\substack{(1.27)\\[0.77]}}{4.70} + \underset{\substack{(0.06)\\[0.05]}}{0.33}\, x_{1t} + \underset{\substack{(0.01)\\[0.02]}}{0.21}\, x_{2t} \\ &\quad + \underset{\substack{(0.08)\\[0.09]}}{0.93}\, [\underset{\substack{(0.56)\\[0.53]}}{2.13}\, m(\underset{\substack{(0.009)\\[0.024]}}{0.060}\, x_{1t}, \underset{\substack{(0.0013)\\[0.0065]}}{0.00004}\, x_{2t}) + v_t]. \end{aligned} \tag{7.2}$$

The term v_t in (7.2) represents a variable distributed $N(0, 1)$ and $m(\tau_1, \tau_2)$ represents the value at (τ_1, τ_2) of an unobserved realization of a random field characterized as the limit of (2.4) as $N \to \infty$. The estimated population parameters $\hat{\zeta} = 1.85, \hat{g}_1 = 0.084, \hat{g}_2 = 0.00006$,and $\hat{\sigma} = 0.93$ characterize the relation between the unobserved $m(.)$ and v_t and the unobserved conditional mean function $\mu(x_1, x_2)$. Numbers in parentheses in (7.2) are

the square roots of diagonal elements of the negative of the inverse of the matrix of second derivatives of (3.24), in other words, the usual asymptotic standard errors for maximum likelihood estimation. Numbers in brackets are the square roots of the posterior Bayesian variances $E\{[\theta_i - E(\theta_i|\mathbf{Y}_T)]^2|\mathbf{Y}_T\}$ as calculated from (5.11) or (5.14) with $N = 20,000$ Monte Carlo simulations. The estimate $\hat{g}_2$ is essentially zero; in other words, x_2 is correctly inferred to play no role in the nonlinearity. By contrast, the coefficients $\hat{g}_1$ and $\hat{\zeta}$ are highly statistically significant, consistent with the strong evidence of nonlinearity found from the LM test.

Given this finding of linearity with respect to x_2, the next task would be to plot the conditional expectation function with respect to x_1 holding x_2 constant. Figure 4 plots $\hat{E}[\mu(x_1, \bar{x}_2)|\mathbf{Y}_T]$ as calculated from (5.17) as a function of x_1 for $\bar{x}_2$ the sample mean of the second explanatory variable, along with 95% confidence intervals. The true function is also shown. Figure 5 shows the analogous plot of $\hat{E}[\mu(\bar{x}_1, x_2)|\mathbf{Y}_T]$ as a function of x_2. Based on these results, a researcher would have little trouble in correctly inferring that the true specification is linear in x_2 and a threshold-linear relation in x_1.

7.2 Example 2

In the second example, the nonlinearity is more complicated and slightly harder to detect statistically:

$$y_t = 5 + 2x_{1t}x_{2t}\delta_{(x_{1t}>0)}\delta_{(x_{2t}>0)} + 0.7x_{3t} + \varepsilon_t \tag{7.3}$$

where $x_{it} \sim N(0, 4)$, $\varepsilon_t \sim N(0, 1)$, and again $T = 100$. Thus $\mu(x_1, x_2, x_3)$ is linear in x_3 and depends on x_1 and x_2 through their product, but only if both are positive. The LM $\chi^2(1)$

test statistic of 64.39 again overwhelmingly rejects the null hypothesis that $\mu(.)$ is linear in all three variables. The maximum likelihood estimates are as follows:

$$\begin{aligned} y_t &= \underset{\substack{(1.21)\\[0.66]}}{7.72} + \underset{\substack{(0.30)\\[0.26]}}{0.85}\, x_{1t} + \underset{\substack{(0.39)\\[0.31]}}{1.01}\, x_{2t} + \underset{\substack{(0.08)\\[0.13]}}{0.70}\, x_{3t} \\ &\quad + \underset{\substack{(0.20)\\[0.13]}}{0.58}\, [\underset{\substack{(2.15)\\[0.89]}}{5.23}\, m(\underset{\substack{(0.04)\\[0.08]}}{0.26}\, x_{1t}, \underset{\substack{(0.03)\\[0.10]}}{0.43}\, x_{2t}, \underset{\substack{(0.017)\\[0.023]}}{0.017}\, x_{3t}) + v_t]. \end{aligned} \tag{7.4}$$

The insignificant value for $\hat{g}_3$ would correctly lead us to conclude that $\mu(x_1, x_2, x_3)$ is linear in x_3. To see what the maximum likelihood estimates reveal about the nature of the nonlinearity in x_1 and x_2, Figure 6 plots contours of the function $\xi_T(x_1, x_2, \bar{x}_3|\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$ for $\bar{x}_3$ the sample mean of the third explanatory variable, $\xi_T(.)$ the result of the calculation in (3.29) or (4.12), and $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimates. This plot would correctly instruct the econometrician that values of x_1 or x_2 outside the northeast quadrant are irrelevant for $\mu(.)$. The figure would also lead us to infer correctly that contours in the northeast quadrant take the form of rectangular hyperbolas; in other words, that $\mu(.)$ is of the form

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 x_2 \delta_{(x_1>0)} \delta_{(x_2>0)} + \beta_3 x_3. \tag{7.5}$$

The model specified by (7.5) and (2.3) could then be estimated by maximum likelihood. In this case, maximum likelihood is achieved by simple OLS estimation, whose results turn out to be

$$y_t = \underset{(0.11)}{4.95} + \underset{(0.07)}{2.02}\, z_{1t} + \underset{(0.05)}{0.69}\, x_{3t} + \hat{\varepsilon}_t \tag{7.6}$$

for $z_{1t} = x_{1t} x_{2t} \delta_{(x_{1t}>0)} \delta_{(x_{2t}>0)}$. One can now use the LM statistic to test whether the form of the nonlinear function $\mu(x_1, x_2, x_3)$ has been correctly identified in (7.5), by taking $\hat{\boldsymbol{\varepsilon}}$ and $\mathbf{M}$ in (6.9) to be the vector of residuals and orthogonal projection matrix from (7.6)

while continuing to calculate **H** from the original explanatory variables (x_{1t}, x_{2t}, x_{3t}). The resulting $\chi^2(1)$ test statistic from (6.9) is now 0.54, so that we would (correctly) accept the null hypothesis that the true nonlinear relation is of the form we guessed in (7.5).

This example illustrates a four-step procedure that may be useful for choosing a conventional parametric nonlinear model. First, we use the LM test (6.9) to see whether the relation between y_t and $\mathbf{x}_t$ is nonlinear. Second, if the relation appears to be nonlinear, we then estimate a flexible model of the form of (7.4). Third, we use this flexible nonlinear inference to learn about the nature of the nonlinearity and to suggest a conventional parametric model of the nonlinearity such as (7.5). Finally, we estimate this parametric model and now use the LM statistic (6.9) as a specification test to see whether the nonlinearity has been successfully modeled.

7.3 Example 3

Our third example is a re-examination of the structural stability of the Phillips Curve, a subject recently addressed by Cooley and Ohanian (1991), King and Watson (1994), King, Stock, and Watson (1995), Raun and Sola (1995), Staiger, Stock and Watson (1997), and Gordon (1997), among others. We use annual data for the inflation rate (π_t) and unemployment rate (u_t) in year t. An OLS regression estimated for $t = 1949$ to 1997,

$$\pi_t = \underset{(54)}{-65} \underset{(0.27)}{-0.44}\, u_t + \underset{(0.12)}{0.73}\, \pi_{t-1} + \underset{(0.028)}{0.035}\, t + \hat{\varepsilon}_t, \tag{7.7}$$

reveals statistically insignificant evidence of a negative inflation-unemployment tradeoff in postwar U.S. data. Note that the large negative intercept of this relation is a consequence

of the large positive values for t.

We are interested in whether a nonlinear relation of the form

$$\pi_t = \mu(u_t, \pi_{t-1}, t) + \varepsilon_t \tag{7.8}$$

might be an improvement over the model in (7.7). The $\chi^2(1)$ LM test in (6.9) yields a value of 5.14, leading us to reject the null hypothesis of a linear relation with a p-value of 0.02. Maximum likelihood estimates of the nonlinear alternative are as follows:

$$\begin{aligned} \pi_t \;=\; & \underset{\substack{(127)\\[115]}}{-88} - \underset{\substack{(0.46)\\[0.44]}}{0.92}\, u_t + \underset{\substack{(0.23)\\[0.22]}}{0.44}\, \pi_{t-1} + \underset{\substack{(0.065)\\[0.058]}}{0.049}\, t \\ & + \underset{\substack{(0.44)\\[0.31]}}{1.24}\, [\underset{\substack{(1.29)\\[1.05]}}{2.05}\, m(\underset{\substack{(0.17)\\[0.25]}}{0.14}\, u_t, \underset{\substack{(0.08)\\[0.12]}}{0.16}\, \pi_{t-1}, \underset{\substack{(0.03)\\[0.05]}}{0.14}\, t) + v_t]. \end{aligned} \tag{7.9}$$

The variable making the most important contribution to the nonlinear part of the conditional expectation in (7.9) is clearly the time trend. Figure 7 plots the value of $\xi_T(\bar{u}, \bar{\pi}, t|\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$ as a function of t, which represents the rate of inflation we would have expected at any date in the sample if the unemployment rate in that year and the inflation rate for the previous year had been equal to their historical average values. The model characterizes the 1970s as a period of unusually high inflation that is not solely attributable to the unemployment rate or the lagged inflation rate.

Figure 8 offers a second way of summarizing the estimated historical relation. It displays a scatter plot of the pair (π_t, u_t) for all dates in the sample; solid boxes denote observations from 1948 to 1972, empty boxes denote observations during 1973-1983, and boxes with x's denote observations during 1984-1997. One can think of the model (7.9) as implying a different "Phillips Curve" for each date in the sample, namely the function $\mu(u, \pi_{t-1}, t)$

plotted as a function of u with π_{t-1} and t the actual historical values for year t. Figure 8 shows three representative Phillips Curves corresponding to $t = 1955, 1975$, and 1985. The diagram is consistent with the standard textbook account according to which there is a short-run negative tradeoff between inflation and unemployment with the intercept of this relation shifting over time. The statistical significance of the estimate $\hat{\alpha}_1$ in equation (7.9) confirms that, once one has allowed for a shifting intercept, there is indeed an important short-run effect of the unemployment rate on the rate of inflation; a 1% increase in the unemployment rate typically lowers the inflation rate for that year by 0.92%.

These findings are, of course, entirely consistent with the conventional textbook account and several of the recent econometric inquiries noted above. The value added by the present exercise is to confirm that interpreting the scatter plot in Figure 8 as the outcome of a shifting Phillips Curve is not the result of sticking obstinately to an ideological prior, but rather is exactly the kind of relation that one would have arrived at using a flexible, atheoretical investigation of the data.

References

Bollerslev, Tim, Ray Y. Chou, and Kenneth F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence", *Journal of Econometrics*, 52: 5-59.

Cooley, Thomas F., and Lee E. Ohanian (1991), "The Cyclical Behavior of Prices," *Journal of Monetary Economics*, 25 (August): 25-60.

Dahl,Christian M. (1998), "An Investigation of Tests for Linearity and the Accuracy of Flexible Nonlinear Inference," working paper, University of Aarhus.

DeGroot, Morris H. (1970), *Optimal Statistical Decisions,* New York: McGraw-Hill.

Diebold, Francis X., and James A. Nason (1990), "Nonparametric Exchange Rate Prediction?", *Journal of International Economics*, 28: 315-332.

Donoho, D. L., I.M. Johnstone, G. Kerkyacarian, and D. Picard (1995), "Wavelet Shrinkage: Asymptopia," *Journal of the Royal Statistical Society, SeriesB*, 57: 301-360.

Eubank, Randall L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, Jianqing, Peter Hall, Michael A. Martin, and Prakash Patil (1996), "On Local Smoothing of Nonparametric Curve Estimators," *Journal of the American Statistical Association*, 91: 258-266.

Friedman, M., and L. J. Savage (1948), "The Utility Analysis of Choice Involving Risk," *Journal of Political Economy,* 56: 279-304.

Gallant, A. Ronald and D. W. Nychka (1987), "Semi-nonparametric Maximum Likeli-

hood Estimation," *Econometrica*, 55: 363-390.

Geweke, John (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica,* 57: 1317-1339.

Gordon, Robert J. (1997), "The Time-Varying NAIRU and its Implications for Economic Policy," *Journal of Economic Perspectives*, 11 (no. 1, Winter): 11-32.

Granger. Clive W. J., Timo Terasvirta, and Heather M. Anderson (1993), "Modeling Nonlinearity over the Business Cycle," in James H. Stock and Mark W. Watson, eds., *Business Cycles, Indicators, and Forecasting*, Chicago: University of Chicago Press.

Hamilton, James D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, 57: 357-384.

____ (1994), *Time Series Analysis*, Princeton, N.J.: Princeton University Press.

Härdle, Wolfgang (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Jeffreys, Harold (1961), *Theory of Probability*, Oxford: Clarendon.

King, Robert G., and Mark W. Watson (1994), "The Postwar U.S. Phillips Curve: A Revisionist Econometric History," *Carnegie-Rochester Conference Series on Public Policy*, 41: 157-219.

____, James H. Stock, and Mark W. Watson (1995), "Temporal Instability of the Unemployment-Inflation Relationship," *Economic Perspectives of the Federal Reserve Bank of Chicago*, May/June 1995, 19: 2-12.

Lauritzen, S. L. (1981), "Time Series Analysis in 1880: A Discussion of Contributions

Made by T. N. Thiele," *International Statistical Review*, 49: 319-331.

Lütkepohl, Helmut (1996), *Handbook of Matrices*, New York: John Wiley & Sons.

Maddala, G. S. (1977), *Econometrics*, New York: McGraw-Hill.

Magnus, Jan R., and Heinz Neudecker (1988), *Matrix Differential Calculus*, New York: John Wiley & Sons.

Mizrach, Bruce (1992), "Multivariate Nearest-Neighbor Forecasts of EMS Exchange Rates," *Journal of Applied Econometrics*, 7 (supplement): S151-S163.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications*, 9: 141-142.

Phillips, Peter C. B. (1991), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics* 6:333-364.

Potter, Simon (1995), "A Nonlinear Approach to U.S. GNP," *Journal of Applied Econometrics*, 10: 109-125.

Raun, Morten O., and Martin Sola (1995), "Stylized Facts and Regime Changes: Are Prices Procyclical?", *Journal of Monetary Economics* 36: 497-526.

Reinsch, C. H. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10: 177-183.

Robinson, Peter M. (1983), "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4: 185-207.

Seifert, Burkhardt, and Theo Gasser (1996), "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91: 267-275.

Sims, Christopher A. (1993), "A Nine-Variable Probabilistic Macroeconomic Forecasting Model," in James H. Stock and Mark W. Watson, eds., *Business Cycles, Indicators, and Forecasting*, Chicago: University of Chicago Press.

Staiger, Douglas, James H. Stock, and Mark W. Watson (1997), "How Precise are Estimates of the Natural Rate of Unemployment? In Christina Romer and David Romer, eds., *Reducing Inflation: Motivation and Strategy*, Chicago: University of Chicago Press.

Terasvirta, Timo, and Heather M. Anderson (1992), "Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models," *Journal of Applied Econometrics*, 7 (supplement): S119-S136.

Tong, Howell (1983), *Threshold Models in Non-linear Time Series Analysis*, New York: Springer-Verlag.

Tsay, Ruey S. (1989), "Testing and Modeling Threshold Autoregressive Processes," *Journal of the American Statistical Association*, 84: 231-240.

Wahba, Grace (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding against Model Errors in Regression," *Journal of the Royal Statistical Society, Series B*, 40: 364-372.

____ (1990), *Spline Models for Observational Data,* Philadelphia: Society for Industrial and Applied Mathematics.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya A*, 26: 359-372.

Wecker, William E., and Craig F. Ansley (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the American Statistical Associa-*

tion, 78: 81-89.

White, Halbert (1994), *Estimation, Inference, and Specification Analysis,* Cambridge, U.K.: Cambridge University Press.

Yakowitz, S. J. (1987), "Nearest Neighbor Methods for Time Series Analysis," *Journal of Time Series Analysis*, 8: 235-247.

Zellner, Arnold (1971), *An Introduction to Bayesian Inference in Econometrics,* New York: John Wiley & Sons, Inc.

Table 1

Covariance between $m(\mathbf{x})$ and $m(\mathbf{z})$ as a function of k (the dimension of $\mathbf{x}$) and h (one-half the distance between $\mathbf{x}$ and $\mathbf{z}$), where $0 \leq h \leq 1$.

k	$H_k(h) = \mathrm{Cov}\Big(m(\mathbf{x}),m(\mathbf{z}) \Big\vert [(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})]^{1/2} = 2h\Big)$
1	$1-h$
2	$1-(2/\pi)\left[h(1-h^2)^{1/2}+\sin^{-1}(h)\right]$
3	$1-(3h/2)+(h^3/2)$
4	$1-(2/\pi)\left[(2/3)h(1-h^2)^{3/2}+h(1-h^2)^{1/2}+\sin^{-1}(h)\right]$
5	$1-(3/2)h+(h^3/2)-(3h/8)(1-h^2)^2$

Notes to Table 1: For any k, the covariance is unity when $h = 0$ and is zero when $h \geq 1$.

Figure Captions

Figure 1. Sample realizations of $e(x_i)$ for $a = 1$, $b = 3$, $\omega = 1$, and $\Delta_N = 0.5$.

Figure 2. Sample realizations of $m_N(x_i)$ for $a = 1$, $b = 3$, $\omega = 1$, and $\Delta_N = 0.5$.

Figure 3. Grid of values for $\mathbf{x}(i_1, i_2)$ for $k = 2$ and $\Delta_{1N} = \Delta_{2N} = 0.5$ and illustration of nodes whose values for $e(\mathbf{x})$ get averaged to determine $m_N(2.0, 1.5)$ and $m_N(2.5, 1.5)$.

Figure 4. Solid line: posterior mean (as estimated from (5.17)) with $N = 10,000$ Monte Carlo draws for a fixed sample of size $T = 100$ generated from the model (7.1). The figure plots $\hat{E}[\mu(x_1, \bar{x}_2)|\mathbf{Y}_T]$ as a function of x_1 for $\bar{x}_2$ the sample mean for variable 2 and $\mathbf{Y}_T$ the given sample of observations on y_t, x_{1t}, and x_{2t}. Dashed lines: 95% confidence intervals. Dotted line: true relation.

Figure 5. Solid line: posterior mean for sample of size 100 generated from the model (7.1). The figure plots $\hat{E}[\mu(\bar{x}_1, x_2)|\mathbf{Y}_T]$ as a function of x_2 for $\bar{x}_1$ the sample mean for variable 1. Dashed lines: 95% confidence intervals.

Figure 6. Contour lines for estimated $\mu(\mathbf{x})$ function for sample of size 100 generated from the model (7.3). The figure plots combinations of x_1 and x_2 such that $\xi_T(x_1, x_2, \bar{x}_3; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$ is constant for $\bar{x}_3$ the sample mean for variable 3 and $\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}}$ the maximum likelihood estimates.

Figure 7. Posterior mean and 95% confidence intervals for the intercept of the historical Phillips Curve as estimated from (5.17). The figure plots $\hat{E}[\bar{u}, \bar{\pi}, t|\mathbf{Y}_T]$ as a function of t for $\bar{u}$ and $\bar{\pi}$ the historical average unemployment and inflation rates, respectively.

Figure 8. Scatter plot for inflation and unemployment data used in estimation; solid boxes correspond to data from 1948-72, empty boxes 1973-83, and boxes with x's 1984-

97. Line labeled "1955" plots $\xi_T(u, \pi_{1954}, 1955; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$ as a function of u for π_{1954} the actual inflation rate in 1954 and $\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}}$ the maximum likelihood estimates. Line labeled "1975" plots $\xi_T(u, \pi_{1974}, 1975; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$ while "1985" plots $\xi_T(u, \pi_{1984}, 1985; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$.

Figure 1

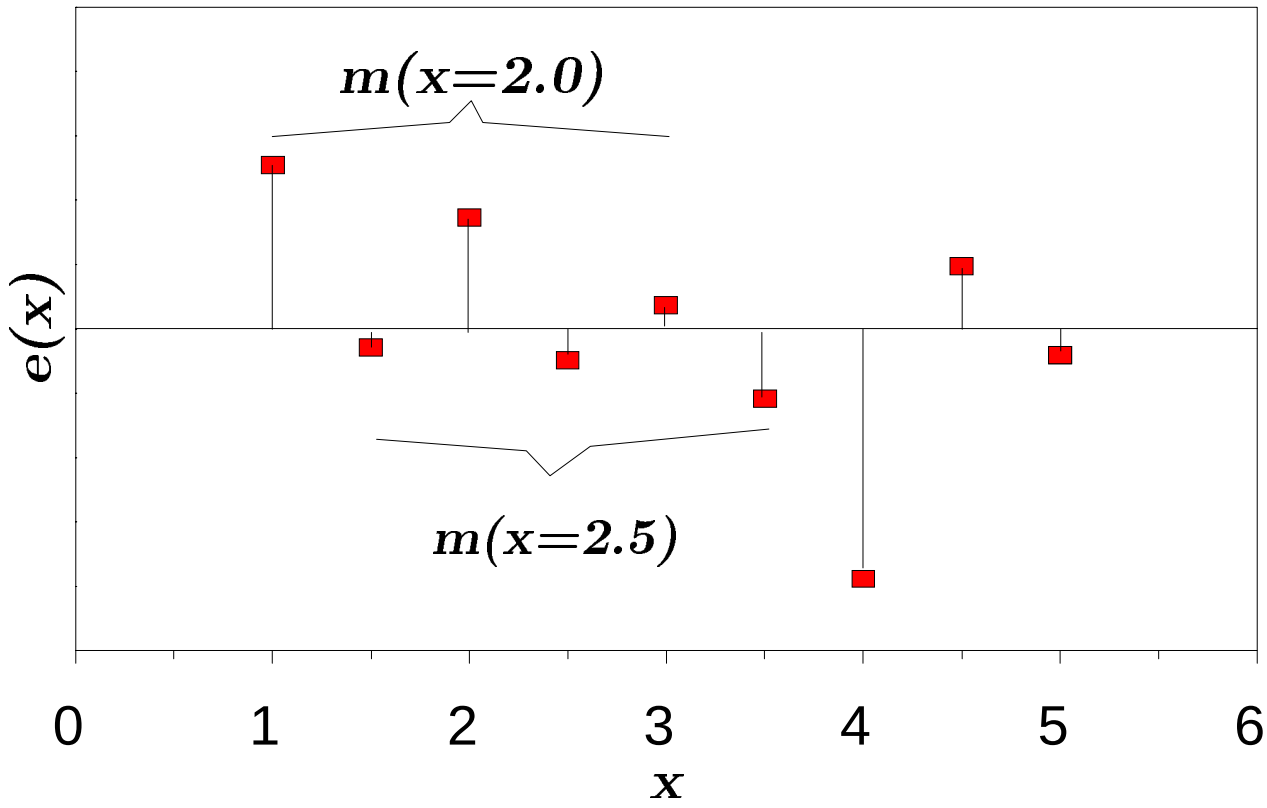


Figure 2

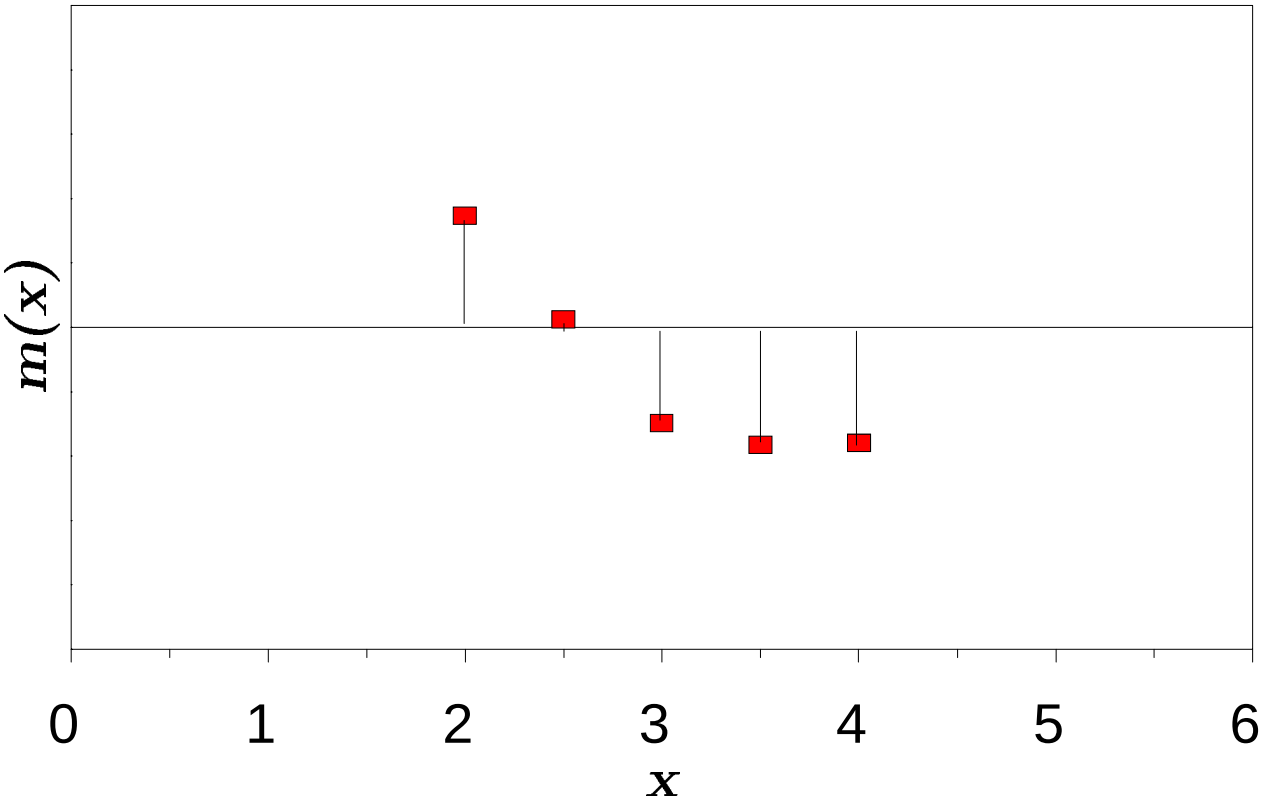


Figure 3

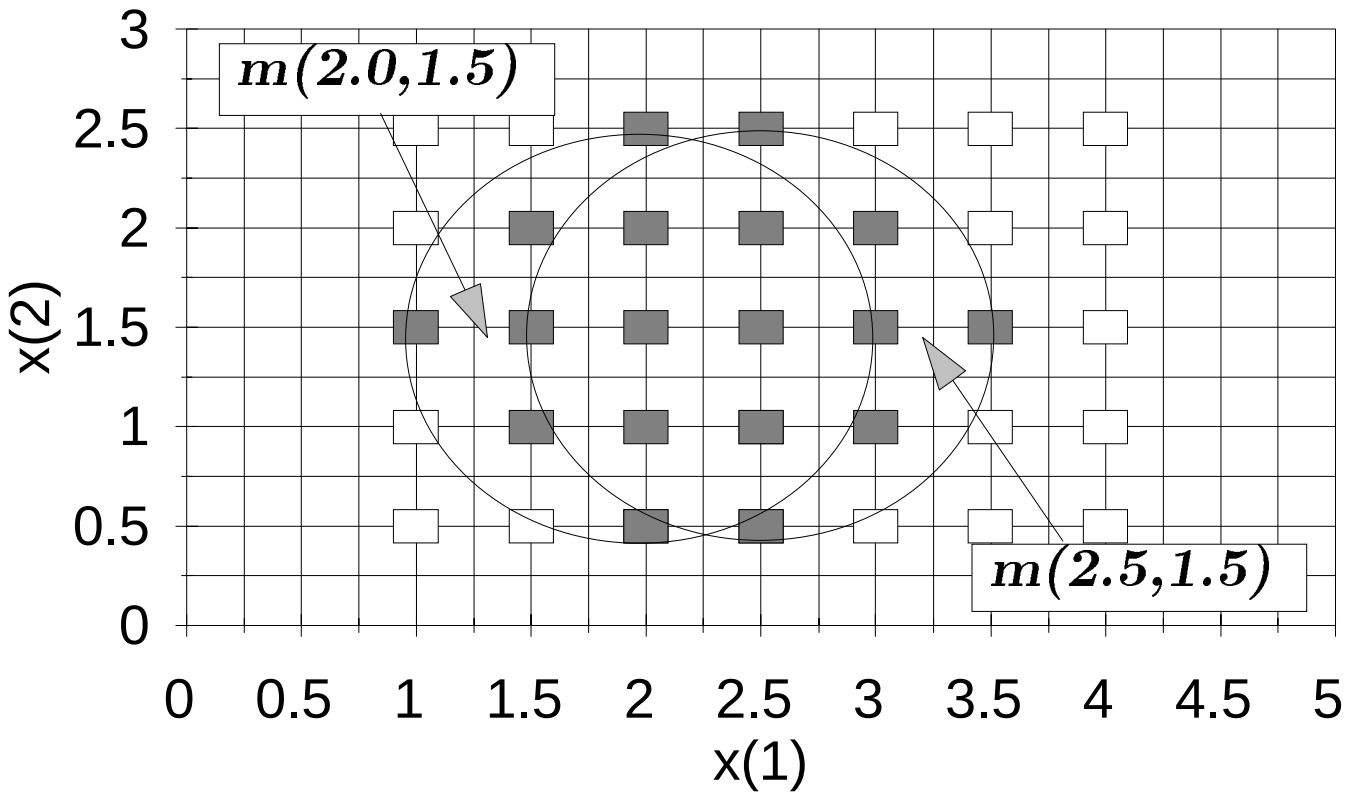


Figure 4

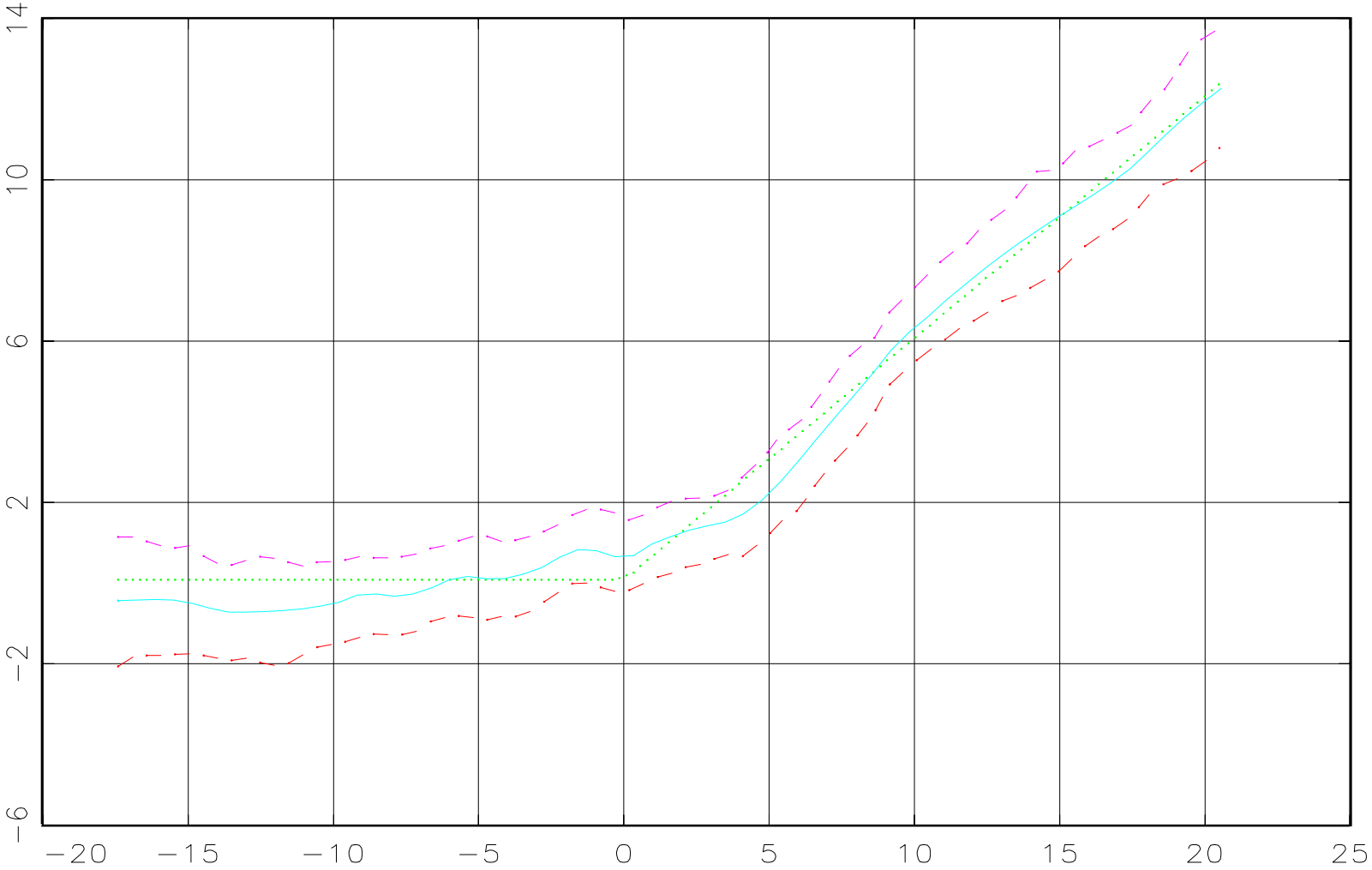


Figure 5

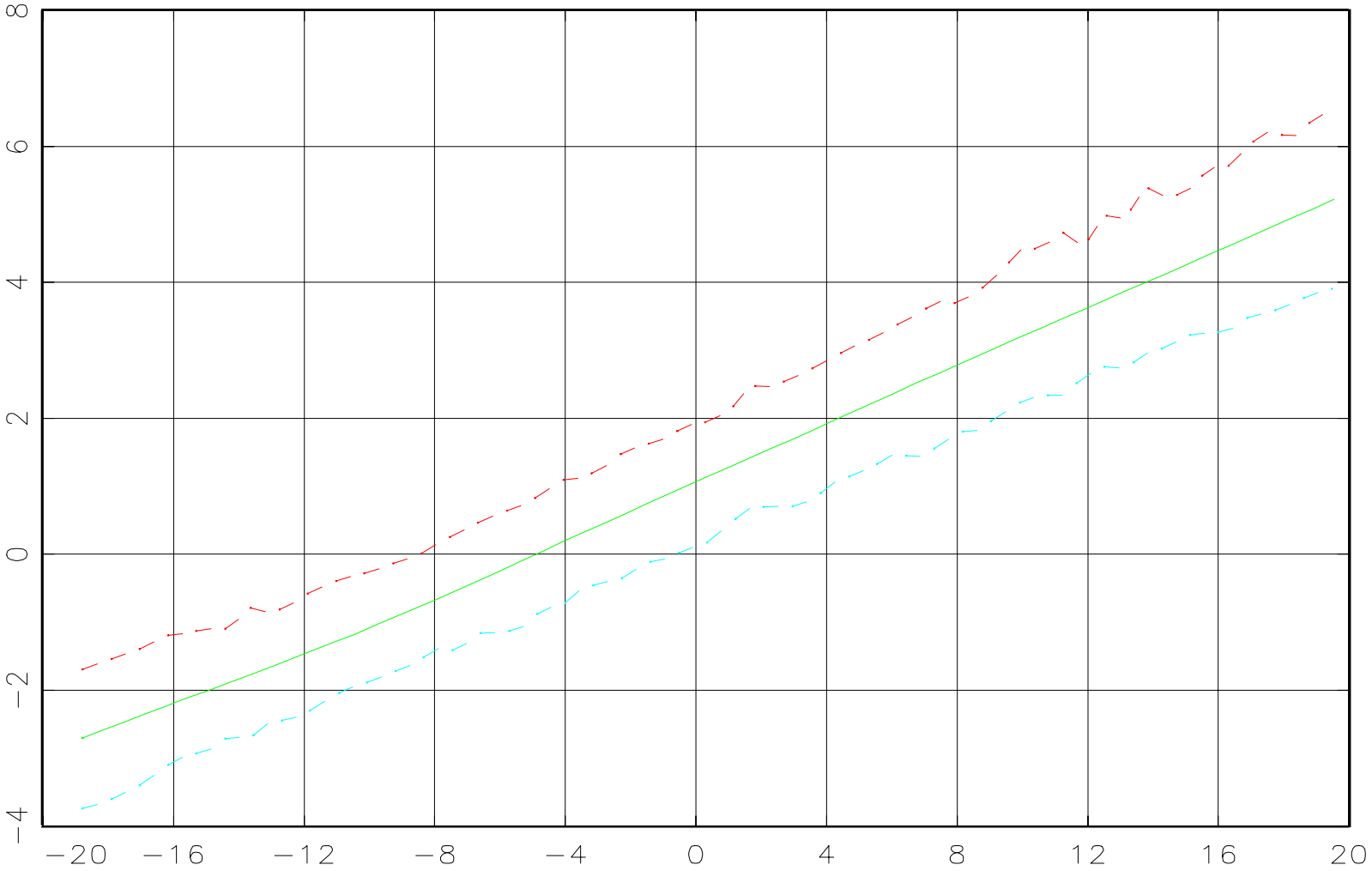


Figure 6

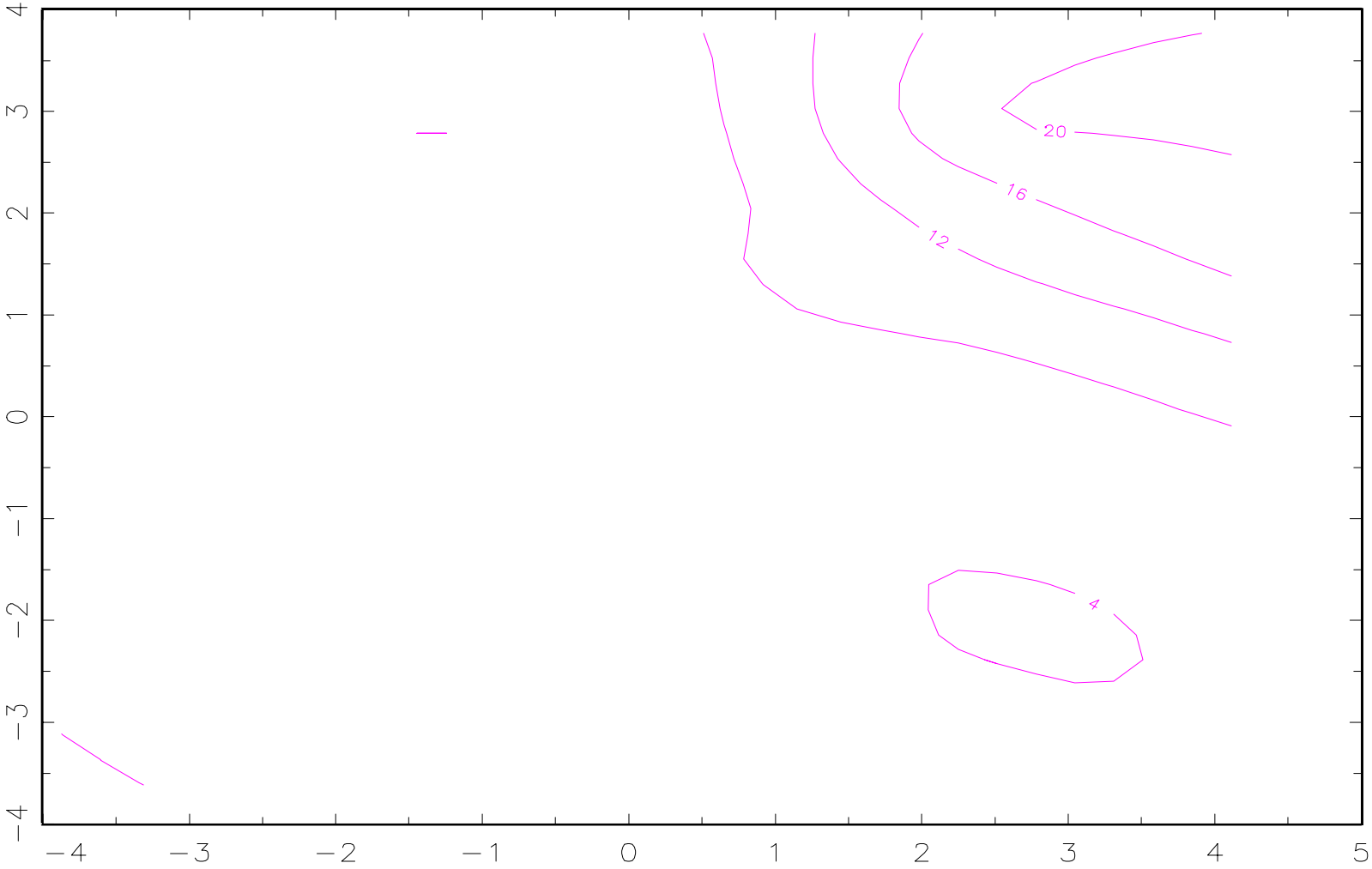


Figure 7

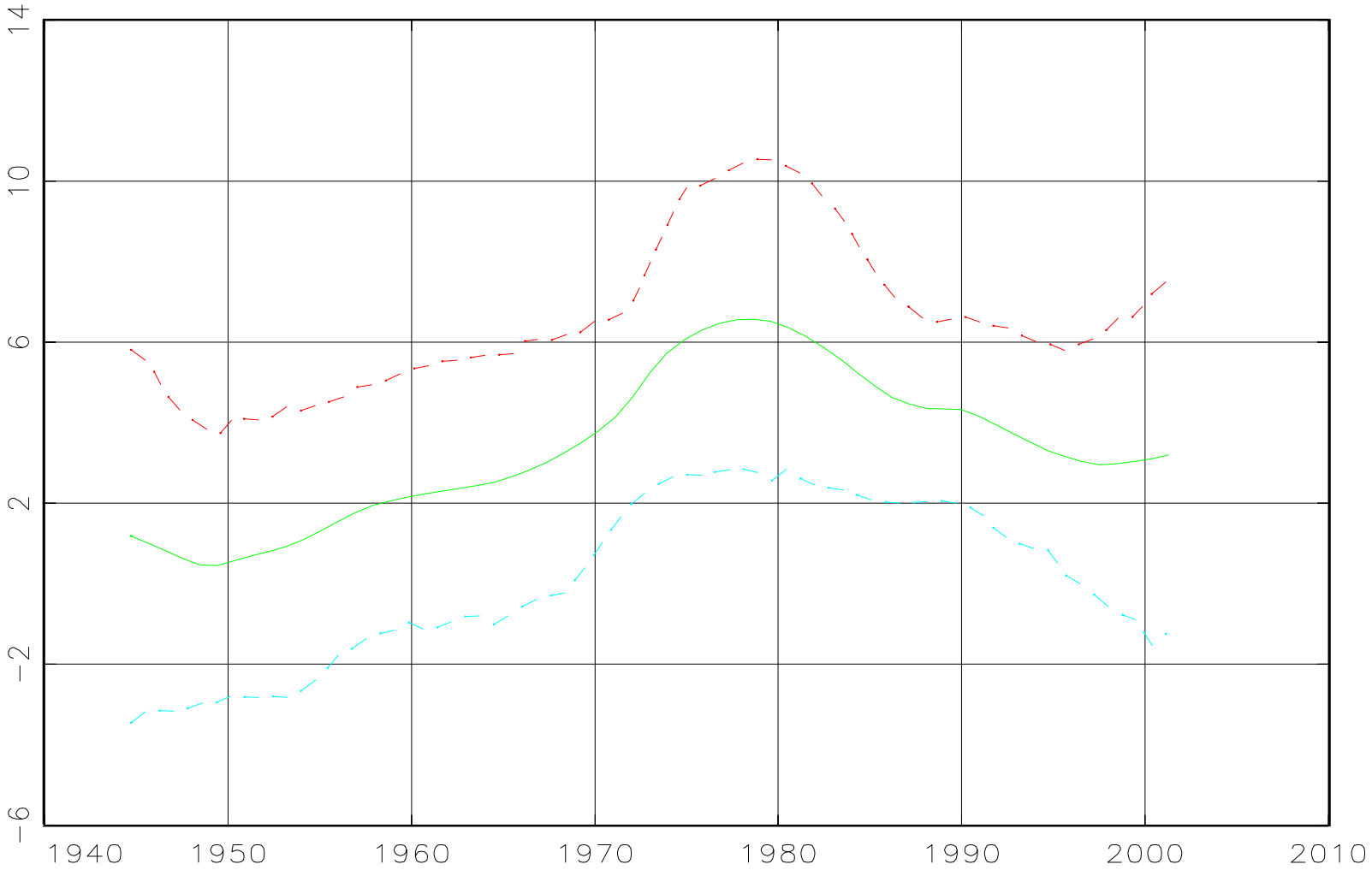
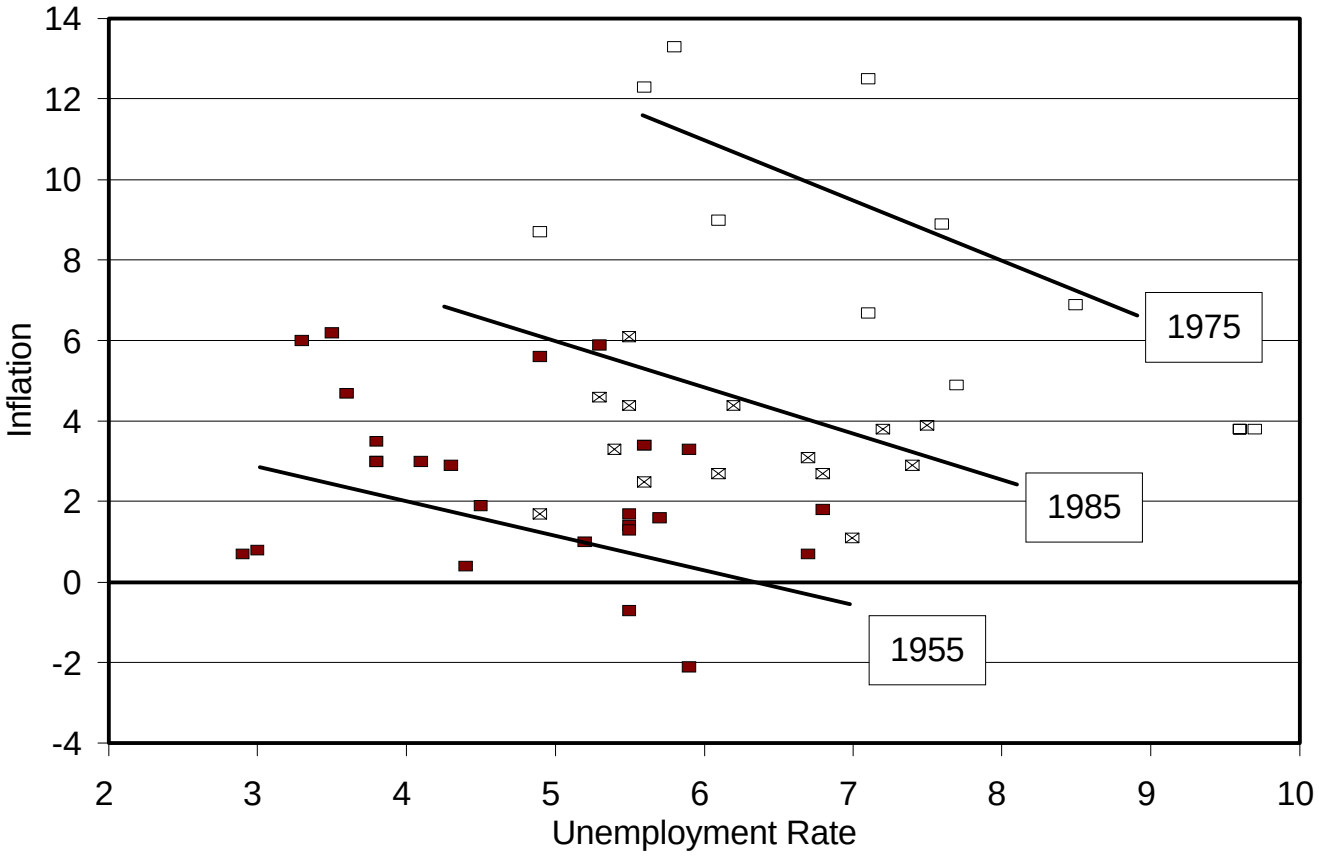


Figure 8


14
12
10
8
6
4
2
0
-2
-4
Inflation
2
3
4
5
6
7
8
9
10
Unemployment Rate
1975
1985
1955

Appendix A: Proofs for Section 2

Proof of Lemma 2.1.

Recall the formula for integration by parts: $\int udv = uv - \int vdu$. Interpreting $u = (r^2 - z^2)^{k/2}$ and $dv = dz$, we have

$$\int (r^2 - z^2)^{k/2} dz = z(r^2 - z^2)^{k/2} + k \int z^2 (r^2 - z^2)^{(k-2)/2} dz. \tag{A.1}$$

Also

$$\begin{aligned} \int (r^2 - z^2)^{k/2} dz &= \int (r^2 - z^2)^{(k-2)/2} (r^2 - z^2) dz \\ &= r^2 \int (r^2 - z^2)^{(k-2)/2} dz - \int z^2 (r^2 - z^2)^{(k-2)/2} dz. \end{aligned} \tag{A.2}$$

Multiplying (A.2) by k and adding the result to (A.1) establishes

$$(1 + k) \int (r^2 - z^2)^{k/2} dz = z(r^2 - z^2)^{k/2} + kr^2 \int (r^2 - z^2)^{(k-2)/2} dz \tag{A.3}$$

and so

$$(1 + k) \int_h^r (r^2 - z^2)^{k/2} dz = -h(r^2 - h^2)^{k/2} + kr^2 \int_h^r (r^2 - z^2)^{(k-2)/2} dz.$$

Dividing by $(1 + k)$ produces (2.6). Equations (2.7) and (2.8) were obtained by direct evaluation of (2.5) using standard tables of integrals.

To prove Theorem 2.2, we first establish three lemmas.

Lemma A.1. The sequence $G_k(0, r)$ of Lemma 2.1 satisfies

$$G_k(0, r) = r^{k+1} G_k(0, 1). \tag{A.4}$$

Proof of Lemma A.1.

From (2.7) and (2.8) we see that (A.4) holds for $k = 0$ and $k = 1$:

$$G_0(0, r) = r = rG_0(0, 1)$$

$$G_1(0, r) = (\pi/4)r^2 = r^2 G_1(0, 1).$$

By induction, given that (A.4) holds for $k - 2$, it follows from (2.6) that it holds for k:

$$G_k(0, r) = \frac{kr^2}{1 + k} r^{k-1} G_{k-2}(0, 1) = r^{k+1} G_k(0, 1). \tag{A.5}$$

Lemma A.2. Let $\mathbf{x} = (x_1, ..., x_k)' \in \Re^k$ and define the k-dimensional spheroid of radius r to be $A_k(r) = \{\mathbf{x} : \mathbf{x}'\mathbf{x} \leq r^2\}$. Let $V_k(r)$ denote the volume of $A_k(r)$. Then

$$V_k(r) = r^k V_k(1) \tag{A.6}$$

where $V_1(1) = 2$ and $V_k(1)$ can be found recursively from

$$V_k(1) = 2V_{k-1}(1)G_{k-1}(0, 1) \tag{A.7}$$

and where $G_k(h, r)$ is given by Lemma 2.1.

Proof of Lemma A.2.

Notice that for any x_k^* such that $|x_k^*| < r$, the values $\mathbf{x}$ such that $\mathbf{x} \in A_k(r)$ and $x_k = x_k^*$ can be characterized as

$$\begin{aligned} \{\mathbf{x} \quad &: \quad (x_1^2 + x_2^2 + \cdots + x_k^2) \leq r^2 \quad \text{and} \quad x_k = x_k^*\} \\ &= \quad \{\mathbf{x} : (x_1^2 + x_2^2 + \cdots + x_{k-1}^2) \leq (r^2 - x_k^{*2}) \quad \text{and} \quad x_k = x_k^*\}. \end{aligned}$$

For given x_k^* this is the description of a $(k-1)$-dimensional spheroid of radius $(r^2 - x_k^{*2})^{1/2}$, denoted $A_{k-1}((r^2 - x_k^{*2})^{1/2})$. Consider a k-dimensional cylindroid defined as the set of all points whose first $k-1$ coordinates are in $A_{k-1}((r^2 - x_k^{*2})^{1/2})$ and whose kth coordinate is in the interval $[x_k^*, x_k^* + dx_k]$. The volume of this cylindroid is $V_{k-1}((r^2 - x_k^{*2})^{1/2})dx_k$. The volume of $A_k(r)$ can be found by integrating these volumes over $x_k^* \in [-r, r]$, or from symmetry

$$V_k(r) = 2\int_0^r V_{k-1}((r^2 - x_k^2)^{1/2})dx_k. \tag{A.8}$$

Result (A.6) can then be shown by induction. We know directly that $V_1(r) = 2r = rV_1(1)$. Given that $V_{k-1}(r) = r^{k-1}V_{k-1}(1)$, it follows from (A.8) that

$$\begin{aligned} V_k(r) &= 2\int_0^r (r^2 - x_k^2)^{(k-1)/2} V_{k-1}(1)dx_k \\ &= 2V_{k-1}(1)G_{k-1}(0, r) \end{aligned}$$

from the definition of $G_{k-1}(0, r)$ in (2.5). From Lemma A.1, then,

$$V_k(r) = 2V_{k-1}(1)r^k G_{k-1}(0, 1), \tag{A.9}$$

establishing both (A.6) and (A.7).

Lemma A.3. Let $A_k(1)$ be the k-dimensional unit spheroid centered at the origin and let $B_k(h, 1)$ be a k-dimensional unit spheroid centered at $(0, 0, ..., 0, 2h)$ with $0 \le h \le 1$:

$$A_k(1) = \{\mathbf{x} : x_1^2 + \cdots + x_k^2 \le 1\}$$

$$B_k(h, 1) = \{\mathbf{x} : x_1^2 + \cdots + x_{k-1}^2 + (x_k - 2h)^2 \le 1\}.$$

Let $C_k(h) = A_k(1) \cap B_k(h,1)$. Then the volume of $C_k(h)$ is given by

$$\int_{\mathbf{x}\in C_k(h)} d\mathbf{x} = 2V_{k-1}(1)G_{k-1}(h,1). \tag{A.10}$$

Proof of Lemma A.3.

Notice that if $\mathbf{x} \in C_k(h)$, then it must be the case that both

$$x_1^2 + \cdots + x_{k-1}^2 \leq 1 - x_k^2 \tag{A.11}$$

and

$$x_1^2 + \cdots + x_{k-1}^2 \leq 1 - (x_k - 2h)^2. \tag{A.12}$$

If $x_k \geq h$, then (A.11) implies (A.12), whereas if $x_k \leq h$, then (A.12) implies (A.11). Thus for given x_k satisfying $h \leq x_k \leq 1$, the set of values of $(x_1, ..., x_k)$ contained in $C_k(h)$ are fully characterized by (A.11), which describes a $(k-1)$-dimensional spheroid of radius $(1-x_k^2)^{1/2}$. By symmetry, the points in $C_k(h)$ satisfying $h \leq x_k \leq 1$ comprise half the total of $C_k(h)$. The volume of $C_k(h)$ is thus

$$\begin{aligned}
\int_{\mathbf{x}\in C_k(h)} d\mathbf{x} &= 2\int_h^1 V_{k-1}((1-x_k^2)^{1/2})dx_k \\
&= 2\int_h^1 (1-x_k^2)^{(k-1)/2} V_{k-1}(1)dx_k \\
&= 2V_{k-1}(1)G_{k-1}(h,1)
\end{aligned}$$

as claimed.

Proof of Theorem 2.2.

If $\mathbf{x}$ and $\mathbf{z}$ are separated by a distance of $2h$, then the correlation between $m(\mathbf{x})$ and $m(\mathbf{z})$ is given by the ratio of the volume of the overlap of unit spheroids $2h$ units apart,

$2V_{k-1}(1)G_{k-1}(h,1)$, to the volume of a single unit spheroid, $V_k(1)$. From (A.7), this ratio is

$$2V_{k-1}(1)G_{k-1}(h,1)/V_k(1) = G_{k-1}(h,1)/G_{k-1}(0,1).$$

Appendix B: Proofs for Section 3

Proof of Theorem 3.2.

(a) Recall (e.g. Maddala, 1977, p. 446) that

$$(\mathbf{A}+\mathbf{BDB}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}+\mathbf{D}^{-1})^{-1}\mathbf{B}'\mathbf{A}^{-1}. \tag{B.1}$$

Letting $\mathbf{A}^{-1} = \mathbf{P}_{t-1}, \mathbf{D}^{-1} = \sigma^2$, and $\mathbf{B} = \mathbf{i}_t$ for $\mathbf{i}_t$ the tth column of $\mathbf{I}_T$, this means

$$(\mathbf{P}_{t-1}^{-1} + \sigma^{-2}\mathbf{i}_t\mathbf{i}_t')^{-1} = \mathbf{P}_{t-1} - \mathbf{P}_{t-1}\mathbf{i}_t(\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2)^{-1}\mathbf{i}_t'\mathbf{P}_{t-1}. \tag{B.2}$$

Equation (B.2) allows the recursion (3.11) to be written

$$\mathbf{P}_t = (\mathbf{P}_{t-1}^{-1} + \sigma^{-2}\mathbf{i}_t\mathbf{i}_t')^{-1}$$

or

$$\mathbf{P}_t^{-1} = \mathbf{P}_{t-1}^{-1} + \sigma^{-2}\mathbf{i}_t\mathbf{i}_t'. \tag{B.3}$$

Recursive evaluation of (B.3) for $t = 1, 2, ..., T$ establishes

$$\mathbf{P}_T^{-1} = \mathbf{P}_0^{-1} + \sigma^{-2}\sum_{t=1}^{T}\mathbf{i}_t\mathbf{i}_t' = \mathbf{P}_0^{-1} + \sigma^{-2}\mathbf{I}_T$$

so

$$\mathbf{P}_T = (\mathbf{P}_0^{-1} + \sigma^{-2}\mathbf{I}_T)^{-1}. \tag{B.4}$$

Appealing to another matrix result from Lütkepohl (1996, p. 29), expression (B.4) can be written

$$\begin{aligned}\mathbf{P}_T &= \sigma^2\mathbf{I}_T(\mathbf{P}_0+\sigma^2\mathbf{I}_T)^{-1}\mathbf{P}_0 \qquad \text{(B.5)}\\ &= (\mathbf{P}_0+\sigma^2\mathbf{I}_T-\mathbf{P}_0)(\mathbf{P}_0+\sigma^2\mathbf{I}_T)^{-1}\mathbf{P}_0\\ &= \mathbf{P}_0-\mathbf{P}_0(\mathbf{P}_0+\sigma^2\mathbf{I}_T)^{-1}\mathbf{P}_0\end{aligned}$$

which reproduces (3.21).

(b) Since $y_t=\mathbf{i}_t'\mathbf{y}$, the recursion (3.10) can be written

$$\begin{aligned}\boldsymbol{\xi}_t &= \boldsymbol{\xi}_{t-1}+\frac{\mathbf{P}_{t-1}\mathbf{i}_t\mathbf{i}_t'(\mathbf{y}-\boldsymbol{\xi}_{t-1})}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t+\sigma^2} \qquad \text{(B.6)}\\ &= \boldsymbol{\xi}_{t-1}+\frac{\mathbf{P}_{t-1}\mathbf{i}_t\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{P}_{t-1}^{-1}(\mathbf{y}-\boldsymbol{\xi}_{t-1})}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t+\sigma^2}.\end{aligned}$$

Substituting (3.11) into (B.6),

$$\begin{aligned}\boldsymbol{\xi}_t &= \boldsymbol{\xi}_{t-1}+(\mathbf{P}_{t-1}-\mathbf{P}_t)\mathbf{P}_{t-1}^{-1}(\mathbf{y}-\boldsymbol{\xi}_{t-1})\\ &= \boldsymbol{\xi}_{t-1}+(\mathbf{y}-\boldsymbol{\xi}_{t-1})-\mathbf{P}_t\mathbf{P}_{t-1}^{-1}(\mathbf{y}-\boldsymbol{\xi}_{t-1})\end{aligned}$$

implying

$$\mathbf{y}-\boldsymbol{\xi}_t=\mathbf{P}_t\mathbf{P}_{t-1}^{-1}(\mathbf{y}-\boldsymbol{\xi}_{t-1}). \qquad \text{(B.7)}$$

Recursive evaluation of (B.7) for $t=1,...,T$ establishes

$$\begin{aligned}\mathbf{y}-\boldsymbol{\xi}_T &= (\mathbf{P}_T\mathbf{P}_{T-1}^{-1})(\mathbf{P}_{T-1}\mathbf{P}_{T-2}^{-1})\cdots(\mathbf{P}_1\mathbf{P}_0^{-1})(\mathbf{y}-\boldsymbol{\xi}_0) \qquad \text{(B.8)}\\ &= \mathbf{P}_T\mathbf{P}_0^{-1}(\mathbf{y}-\boldsymbol{\xi}_0)\end{aligned}$$

or

$$\boldsymbol{\xi}_T=\boldsymbol{\xi}_0+(\mathbf{I}_T-\mathbf{P}_T\mathbf{P}_0^{-1})(\mathbf{y}-\boldsymbol{\xi}_0). \qquad \text{(B.9)}$$

Substituting (B.5) into (B.9) gives

$$\boldsymbol{\xi}_T = \boldsymbol{\xi}_0 + \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \boldsymbol{\xi}_0) \tag{B.10}$$

which for $\boldsymbol{\xi}_0 = \mathbf{X}\boldsymbol{\beta}$ is identical to (3.20).

(c) We first show that

$$\sum_{t=1}^{T} \frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{B.11}$$

Recall that $y_t = \mathbf{y}'\mathbf{i}_t$ so that

$$\frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} = \frac{(\mathbf{y} - \boldsymbol{\xi}_{t-1})'\mathbf{i}_t\mathbf{i}_t'(\mathbf{y} - \boldsymbol{\xi}_{t-1})}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2}. \tag{B.12}$$

It further follows as in (B.8) that

$$\mathbf{y} - \boldsymbol{\xi}_{t-1} = \mathbf{P}_{t-1}\mathbf{P}_0^{-1}(\mathbf{y} - \boldsymbol{\xi}_0) \tag{B.13}$$

so that (B.12) becomes

$$\begin{aligned} \frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} &= \frac{(\mathbf{y} - \boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}\mathbf{P}_{t-1}\mathbf{i}_t\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{P}_0^{-1}(\mathbf{y} - \boldsymbol{\xi}_0)}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} \\ &= (\mathbf{y} - \boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}(\mathbf{P}_{t-1} - \mathbf{P}_t)\mathbf{P}_0^{-1}(\mathbf{y} - \boldsymbol{\xi}_0) \end{aligned} \tag{B.14}$$

with the last equality following from (3.11). Summing (B.14) over t establishes

$$\begin{aligned} \sum_{t=1}^{T} \frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2} &= (\mathbf{y} - \boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}\left[\sum_{t=1}^{T}(\mathbf{P}_{t-1} - \mathbf{P}_t)\right]\mathbf{P}_0^{-1}(\mathbf{y} - \boldsymbol{\xi}_0) \\ &= (\mathbf{y} - \boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}(\mathbf{P}_0 - \mathbf{P}_T)\mathbf{P}_0^{-1}(\mathbf{y} - \boldsymbol{\xi}_0). \end{aligned} \tag{B.15}$$

But (B.5) implies that

$$\mathbf{P}_0^{-1}(\mathbf{P}_0 - \mathbf{P}_T)\mathbf{P}_0^{-1} = (\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}. \tag{B.16}$$

Substituting (B.16) and $\boldsymbol{\xi}_0 = \mathbf{X}\boldsymbol{\beta}$ into (B.15) reproduces (B.11).

Next we show that

$$\sum_{t=1}^{T} \ln(\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2) = \ln|\mathbf{P}_0 + \sigma^2\mathbf{I}_T|. \tag{B.17}$$

Define $\boldsymbol{\Omega}_t$ to be the $(T-t+1)\times(T-t+1)$ matrix consisting of rows t through T and columns t through T of the matrix $\mathbf{P}_{t-1}+\sigma^2\mathbf{I}_T$. Define $n \equiv T-t$ and partition this matrix as

$$\boldsymbol{\Omega}_t = \begin{bmatrix} \underset{(1\times1)}{\Omega_{11}} & \underset{(1\times n)}{\boldsymbol{\Omega}_{12}} \\ \underset{(n\times1)}{\boldsymbol{\Omega}_{21}} & \underset{(n\times n)}{\boldsymbol{\Omega}_{22}} \end{bmatrix}.$$

The triangular factorization (e.g. Hamilton, 1994, equation [4.5.26]) of $\boldsymbol{\Omega}_t$ is $\boldsymbol{\Omega}_t = \mathbf{ADA}'$ where

$$\mathbf{A} = \begin{bmatrix} 1 & \mathbf{0}' \\ \boldsymbol{\Omega}_{21}\Omega_{11}^{-1} & \mathbf{I}_n \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} \Omega_{11} & \mathbf{0}' \\ \mathbf{0} & (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\Omega_{11}^{-1}\boldsymbol{\Omega}_{12}) \end{bmatrix}.$$

But since $|\mathbf{A}| = 1$, we see

$$|\boldsymbol{\Omega}_t| = |\mathbf{A}|\cdot|\mathbf{D}|\cdot|\mathbf{A}'| = |\mathbf{D}| = \Omega_{11}\cdot|\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\Omega_{11}^{-1}\boldsymbol{\Omega}_{12}|. \tag{B.18}$$

Now, Ω_{11} is defined as the row t, column t element of $\mathbf{P}_{t-1}+\sigma^2\mathbf{I}_T$, which can be written

$$\Omega_{11} = \mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2. \tag{B.19}$$

Furthermore, $\boldsymbol{\Omega}_{22}$ is the submatrix consisting of the last n rows and columns of $\mathbf{P}_{t-1} + \sigma^2\mathbf{I}_T$. Finally, $\boldsymbol{\Omega}_{21}$ comprises the last n elements of the vector $\mathbf{P}_{t-1}\mathbf{i}_t$. It follows that

$\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\Omega_{11}^{-1}\mathbf{\Omega}_{12}$ is identical to the submatrix comprising the last n rows and columns of the matrix

$$\mathbf{P}_{t-1} + \sigma^2\mathbf{I}_T - \frac{\mathbf{P}_{t-1}\mathbf{i}_t\mathbf{i}_t'\mathbf{P}_{t-1}}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2}.$$

But from (3.11), this is identical to the submatrix consisting of the last n rows and columns of $\mathbf{P}_t + \sigma^2\mathbf{I}_T$. But the latter submatrix was earlier defined to be $\mathbf{\Omega}_{t+1}$. Thus from (B.18) and (B.19) we conclude

$$|\mathbf{\Omega}_t| = (\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2) \cdot |\mathbf{\Omega}_{t+1}|. \tag{B.20}$$

Recursive evaluation of (B.20) for $t = 1, .., T-1$ establishes

$$|\mathbf{\Omega}_1| = (\mathbf{i}_1'\mathbf{P}_0\mathbf{i}_1 + \sigma^2)(\mathbf{i}_2'\mathbf{P}_1\mathbf{i}_2 + \sigma^2)\cdots(\mathbf{i}_T'\mathbf{P}_{T-1}\mathbf{i}_T + \sigma^2). \tag{B.21}$$

But since $\mathbf{\Omega}_1$ is defined to equal $\mathbf{P}_0 + \sigma^2\mathbf{I}_T$, taking logs of (B.21) implies

$$\ln|\mathbf{P}_0 + \sigma^2\mathbf{I}_T| = \sum_{t=1}^{T} \ln(\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2)$$

as claimed in (B.17).

Combining (B.11) and (B.17), we conclude that

$$-(1/2)\sum_{t=1}^{T} \ln(\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2) - (1/2)\sum_{t=1}^{T} \frac{(y_t - \mathbf{i}_t'\boldsymbol{\xi}_{t-1})^2}{\mathbf{i}_t'\mathbf{P}_{t-1}\mathbf{i}_t + \sigma^2}$$

$$= -(1/2)\ln|\mathbf{P}_0 + \sigma^2\mathbf{I}_T| - (1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

establishing the numerical equivalence of (3.17) and (3.22).

Appendix C: Proofs for Section 4

Proof of Theorem 4.1.

(a) It is easier to prove this result using an alternative expression for $\boldsymbol{\xi}_T$. Notice that $\boldsymbol{\xi}_T$ would correspond to the conditional expectation $E(\boldsymbol{\mu}|\mathbf{y})$ for $\mathbf{y}$ the $(T \times 1)$ vector of observations on the dependent variable if $\boldsymbol{\mu} \sim N(\boldsymbol{\xi}_0, \mathbf{P}_0)$ and $\mathbf{y}|\boldsymbol{\mu}$ had the following density,

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\mu}) &= (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{t=1}^{T}\sum_{i=1}^{N}(y_t-\mu_i)^2\delta_{\mathbf{x}t=\mathbf{x}(i)}\right\} \qquad \text{(C.1)} \\ &= (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{N}\left[T_i s_i^2 + T_i(h_i-\mu_i)^2\right]\right\}. \end{aligned}$$

Expression (C.1) can further be written

$$f(\mathbf{y}|\boldsymbol{\mu}) = \kappa|\boldsymbol{\Lambda}_T|^{-1/2}\exp[-(1/2)(\mathbf{h}-\boldsymbol{\mu})'\boldsymbol{\Lambda}_T^{-1}(\mathbf{h}-\boldsymbol{\mu})] \qquad \text{(C.2)}$$

where $\mathbf{h} = (h_1, ..., h_N)'$ and

$$\underset{(N\times N)}{\boldsymbol{\Lambda}_T} = \begin{bmatrix} \sigma^2/T_1 & 0 & \cdots & 0 \\ 0 & \sigma^2/T_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2/T_N \end{bmatrix} \qquad \text{(C.3)}$$

$$\kappa = (2\pi\sigma^2)^{-T/2}(\sigma^2)^{N/2}(T_1\cdots T_N)^{-1/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{N}T_i s_i^2\right\}. \qquad \text{(C.4)}$$

The joint density of $\mathbf{y}$ and $\boldsymbol{\mu}$ would in this case be

$$\begin{aligned} f(\mathbf{y},\boldsymbol{\mu}) &= \kappa|\boldsymbol{\Lambda}_T|^{-1/2}\exp[-(1/2)(\mathbf{h}-\boldsymbol{\mu})'\boldsymbol{\Lambda}_T^{-1}(\mathbf{h}-\boldsymbol{\mu})] \qquad \text{(C.5)} \\ &\quad \times(2\pi)^{-N/2}|\mathbf{P}_0|^{-1/2}\exp[-(1/2)(\boldsymbol{\mu}-\boldsymbol{\xi}_0)'\mathbf{P}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\xi}_0)] \end{aligned}$$

$$= \kappa(2\pi)^{-N/2} \left| \begin{array}{cc} \mathbf{\Lambda}_T + \mathbf{P}_0 & \mathbf{P}_0 \\ \mathbf{P}_0 & \mathbf{P}_0 \end{array} \right|^{-1/2} \exp \left\{ -(1/2) \begin{bmatrix} \mathbf{h} - \boldsymbol{\xi}_0 \\ \boldsymbol{\mu} - \boldsymbol{\xi}_0 \end{bmatrix}' \right.$$
$$\left. \times \begin{bmatrix} \mathbf{\Lambda}_T + \mathbf{P}_0 & \mathbf{P}_0 \\ \mathbf{P}_0 & \mathbf{P}_0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h} - \boldsymbol{\xi}_0 \\ \boldsymbol{\mu} - \boldsymbol{\xi}_0 \end{bmatrix} \right\}$$

Thus from Lemma 3.1,

$$E(\boldsymbol{\mu}|\mathbf{y}) = E(\boldsymbol{\mu}|\mathbf{h}) = \boldsymbol{\xi}_0 + \mathbf{P}_0(\mathbf{\Lambda}_T + \mathbf{P}_0)^{-1}(\mathbf{h} - \boldsymbol{\xi}_0). \tag{C.6}$$

We derived (C.6) as the estimator that one would use if the true mean $\boldsymbol{\mu}$ were distributed $N(\boldsymbol{\xi}_0, \mathbf{P}_0)$, and thus as an expression for $\boldsymbol{\xi}_T$ in this case of discrete-valued explanatory variables. Now that this expression for the estimator has been derived, however, one can note a key property that it will display for any assumptions about the true population mean $\boldsymbol{\xi}$. In particular, from (C.3), the ith row of $\mathbf{P}_0(\mathbf{\Lambda}_T + \mathbf{P}_0)^{-1}$ converges to the ith row of $\mathbf{I}_N$ as $T_i \to \infty$, from which the ith row of (C.6) becomes

$$\hat{\mu}_{iT} \xrightarrow{p} \hat{\mu}_{i0} + (h_i - \hat{\mu}_{i0}) = h_i.$$

Furthermore, under (4.2), $h_i \xrightarrow{p} \ell(\mathbf{x}(i))$, completing the demonstration of (4.3).

(b) Integrating (C.5) over all possible values for $\boldsymbol{\mu}$ gives the marginal density

$$f(\mathbf{y}) = \kappa|\mathbf{\Lambda}_T + \mathbf{P}_0|^{-1/2} \exp\left[-(1/2)(\mathbf{h} - \boldsymbol{\xi}_0)'(\mathbf{\Lambda}_T + \mathbf{P}_0)^{-1}(\mathbf{h} - \boldsymbol{\xi}_0)\right]$$

with log likelihood

$$\ln f(\mathbf{y}) = -(T/2)\ln(2\pi) - [(T-N)/2]\ln\sigma^2 - (1/2)\sum_{i=1}^{N} \ln(T_i) \tag{C.7}$$

$$-[1/(2\sigma^2)]\sum_{i=1}^{N} T_i s_i^2 - (1/2)\ln|\mathbf{\Lambda}_T + \mathbf{P}_0|$$
$$-(1/2)(\mathbf{h}-\boldsymbol{\xi}_0)'(\mathbf{\Lambda}_T + \mathbf{P}_0)^{-1}(\mathbf{h}-\boldsymbol{\xi}_0).$$

Expression (4.5) follows from (C.7) and the facts that $\mathbf{\Lambda}_T \to \mathbf{0}$ and $\mathbf{h} \xrightarrow{p}$ Ļ.

(c) By differentiating (C.7) we see that

$$\frac{\partial \ln f(\mathbf{y})}{\partial \sigma^2} + \frac{T-N}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N} T_i s_i^2 \xrightarrow{p} 0$$

and hence

$$\begin{aligned}
\hat{\sigma}^2 \xrightarrow{p} \quad & (T-N)^{-1}\sum_{i=1}^{N} T_i s_i^2 \\
= \quad & (T-N)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}(y_t - h_i)^2 \delta_{\mathbf{x}_t = \mathbf{x}(i)} \\
\xrightarrow{p} \quad & T^{-1}\sum_{t=1}^{T}(y_t - \ell(\mathbf{x}_t))^2
\end{aligned}$$

as claimed in (4.6). Expression (4.7) follows from (4.5) with $\boldsymbol{\xi}_0 = \mathbf{X}\boldsymbol{\beta}$.

Proof of Theorem 4.4. Observe from (4.11) that if some imaginary data for $\{\tilde{y}_t, \mathbf{x}_t\}$ had been generated from $\tilde{y}_t = \tilde{\mu}(\mathbf{x}_t) + \tilde{\varepsilon}_t$ where the function $\tilde{\mu}(\mathbf{z})$ had been generated from a Gaussian field with mean $\xi_0(\mathbf{z})$ and covariance $p_0(\mathbf{z}, \mathbf{w})$, then $p_t(\mathbf{z}, \mathbf{z})$ could be interpreted as the MSE of the optimal estimate of $\tilde{\mu}(\mathbf{z})$ based on a sample of size t for such data,

$$p_t(\mathbf{z}, \mathbf{z}) = E[\tilde{\mu}(\mathbf{z}) - \xi_t(\mathbf{z})]^2.$$

As such, $p_t(.,.)$ is positive semidefinite for all t and, for given $\mathbf{z}$, $\{p_t(\mathbf{z}, \mathbf{z})\}$ is a monotonically nonincreasing sequence which is bounded from below by zero. The sequence $\{p_t(\mathbf{z}, \mathbf{z})\}$

therefore converges to some constant $p(\mathbf{z}, \mathbf{z})$ as $t \to \infty$. From (4.13), this requires that for any $\varepsilon_1 > 0$, $\exists T_0$ such that $\forall t \geq T_0$,

$$\frac{[p_{t-1}(\mathbf{x}, \mathbf{z}_t)]^2}{p_{t-1}(\mathbf{x}_t, \mathbf{x}_t) + \sigma^2} < \varepsilon_1.$$

Since $\sigma^2 > 0$ and $p_{t-1}(\mathbf{x}_t, \mathbf{x}_t)$ is bounded, this requires that for any $\varepsilon_2 > 0$,

$$|p_{t-1}(\mathbf{x}_t, \mathbf{z}_t)| < \varepsilon_2 \tag{C.8}$$

for all $t \geq T_0$.

Next consider the random variable $\tilde{\mu}(\mathbf{x}_t) - \tilde{\mu}(\mathbf{z})$, whose variance would be

$$\begin{aligned} p_0(\mathbf{x}_t, \mathbf{x}_t) - \quad & 2p_0(\mathbf{x}_t, \mathbf{z}) + p_0(\mathbf{z}, \mathbf{z}) \\ \leq \quad & |p_0(\mathbf{x}_t, \mathbf{x}_t) - p_0(\mathbf{x}_t, \mathbf{z})| + |p_0(\mathbf{z}, \mathbf{z}) - p_0(\mathbf{x}_t, \mathbf{z})|. \end{aligned} \tag{C.9}$$

Let $\mathbf{g} \in \Re^k$ be an arbitrary vector of fixed weights and define

$$W_\delta(\mathbf{z}) = \{\mathbf{w} \in A\text{: } \|\mathbf{g} \odot (\mathbf{w} - \mathbf{z})\| < \delta\}.$$

Continuity of $p_0(\mathbf{z}, \mathbf{w})$ ensures that for any $\varepsilon_3 > 0, \exists \delta > 0$ such that the RHS of (C.9) is less than ε_3 whenever $\mathbf{x}_t \in W_\delta(\mathbf{z})$. Furthermore, the optimal inference about $\tilde{\mu}(\mathbf{x}_t) - \tilde{\mu}(\mathbf{z})$ based on observation of $\mathbf{Y}_{t-1}$ could have an MSE no greater than that based on no observations, requiring

$$p_{t-1}(\mathbf{x}_t, \mathbf{x}_t) - 2p_{t-1}(\mathbf{x}_t, \mathbf{z}) + p_{t-1}(\mathbf{z}, \mathbf{z}) < \varepsilon_3 \tag{C.10}$$

for all $\mathbf{x}_t \in W_\delta(\mathbf{z})$. Condition (C.8) thus ensures that

$$p_{t-1}(\mathbf{x}_t, \mathbf{x}_t) + p_{t-1}(\mathbf{z}, \mathbf{z}) < \varepsilon_3 + 2\varepsilon_2 \tag{C.11}$$

whenever $t \geq T_0$ and $\mathbf{x}_t \in W_\delta(\mathbf{z})$. But denseness of $\{\mathbf{x}_t\}$ guarantees the existence of some $\mathbf{x}_t$ with $t \geq T_0$ satisfying $\mathbf{x}_t \in W_\delta(\mathbf{z})$, and for this t, (C.11) implies that $p_{t-1}(\mathbf{z}, \mathbf{z})$ must be arbitrarily small. Indeed, from (C.8), denseness ensures that $|p_t(\mathbf{z}, \mathbf{w})|$ becomes arbitrarily small for any $\mathbf{z}$ and $\mathbf{w}$ as t becomes large.

Note that the assumption of artificial data $\tilde{y}_t$ was irrelevant for this result. The sequence (4.13) depends only on $p_0(.,.)$ and σ^2 and is not influenced by any data on y_t. Hence (4.15) must be a property of the recursion (4.13) itself rather than an outcome for a particular data set.

Proof of Lemma 4.5.

Imagine a different sample that had actually been generated from

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{m} + \tilde{\boldsymbol{\varepsilon}} \tag{C.12}$$

where $\mathbf{m} = (m(\mathbf{x}_1), ..., m(\mathbf{x}_T))', \lambda^2 E[m(\mathbf{z})m(\mathbf{w})] = p_0(\mathbf{z}, \mathbf{w})$, and $\tilde{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$ independent of $\mathbf{m}$. Then from (3.29),

$$\hat{\mu}^* - \mathbf{X}^*\boldsymbol{\beta} = \mathbf{h}_T'(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{h}_T'(\lambda\mathbf{m} + \tilde{\boldsymbol{\varepsilon}}) \tag{C.13}$$

where

$$\mathbf{h}_T' = \mathbf{q}_T'(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}. \tag{C.14}$$

Subtracting $\lambda m(\mathbf{x}^*)$ from both sides of (C.13), it follows that

$$\hat{\mu}^* - \mathbf{X}^*\boldsymbol{\beta} - \lambda m(\mathbf{x}^*) = \lambda[\mathbf{h}_T'\mathbf{m} - m(\mathbf{x}^*)] + \mathbf{h}_T'\tilde{\boldsymbol{\varepsilon}}$$

and

$$E[\hat{\mu}^* - \mathbf{X}^*\boldsymbol{\beta} - \lambda m(\mathbf{x}^*)]^2 = \lambda^2 E[\mathbf{h}_T'\mathbf{m} - m(\mathbf{x}^*)]^2 + E(\mathbf{h}_T'\tilde{\boldsymbol{\varepsilon}})^2. \tag{C.15}$$

Now, for data actually generated by (C.12), the left side of (C.15) would be described by $p_T(\mathbf{x}^*, \mathbf{x}^*)$, which from Theorem 4.4 converges to zero. Therefore, each term on the right side of (C.15) also converges to zero; in particular,

$$E(\mathbf{h}_T'\tilde{\boldsymbol{\varepsilon}})^2 = \sigma^2(\mathbf{h}_T'\mathbf{h}_T) \to 0. \tag{C.16}$$

Expression (C.16) was derived under the assumption that the data on $\mathbf{y}$ were actually generated from (C.12). However, the vector $\mathbf{h}_T$ can be calculated mechanically from an arbitrary $\mathbf{P}_0$ matrix as specified in (C.14), without using any data on $\mathbf{y}$. It follows that (C.16) is a property of the $\mathbf{h}_T$ vector so constructed rather than a property of the data. This means that if $\mathbf{h}_T$ is any vector constructed as in (C.14), then for $\mathbf{a}$ the vector in Lemma 4.5,

$$E(\mathbf{h}_T'\mathbf{a})^2 = \mathbf{h}_T'\nu^2\mathbf{I}_T\mathbf{h}_T \to 0$$

as $T \to \infty$, as claimed in (4.16).

Proof of Theorem 4.7.

Define the $(T\times 1)$ vectors $\boldsymbol{\xi}_T = (\xi_T(\mathbf{x}_1), \xi_T(\mathbf{x}_2), ..., \xi_T(\mathbf{x}_T))', \mathbf{l}_T = (\ell_T(\mathbf{x}_1), \ell_T(\mathbf{x}_2), ..., \ell_T(\mathbf{x}_T))'$, and $\boldsymbol{\theta}_T = (\theta_T(\mathbf{x}_1), \theta_T(\mathbf{x}_2), ..., \theta_T(\mathbf{x}_T))'$. Then (4.21) can be written

$$T^{-1}E(\boldsymbol{\xi}_T - \mathbf{X}\boldsymbol{\beta} - \mathbf{l}_T)'(\boldsymbol{\xi}_T - \mathbf{X}\boldsymbol{\beta} - \mathbf{l}_T) \to 0 \tag{C.17}$$

while (4.18) implies

$$\mathbf{l}_T = T^{-1}\mathbf{P}_0\boldsymbol{\theta}_T. \tag{C.18}$$

Notice also from (3.20) that

$$\boldsymbol{\xi}_T = \mathbf{X}\boldsymbol{\beta} + \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{C.19}$$

It follows from (C.19) that

$$\begin{aligned} \boldsymbol{\xi}_T - \mathbf{X}\boldsymbol{\beta} - \mathbf{l}_T &= \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{l}_T \\ &= \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}(\mathbf{l}_T + \mathbf{a}) - \mathbf{l}_T \end{aligned} \tag{C.20}$$

with the last equality coming from (4.19). Substituting (C.18) into (C.20) gives

$$\begin{aligned} \boldsymbol{\xi}_T - \mathbf{X}\boldsymbol{\beta} - \mathbf{l}_T &= T^{-1}[\mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}\mathbf{P}_0 - \mathbf{P}_0]\boldsymbol{\theta}_T + \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}\mathbf{a} \\ &= -T^{-1}\mathbf{P}_T\boldsymbol{\theta}_T + \mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}\mathbf{a} \end{aligned} \tag{C.21}$$

with the last equality following from (3.21). The tth row of (C.21) states that

$$\xi_T(\mathbf{x}_t) - \alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_t - \ell_T(\mathbf{x}_t) = -T^{-1}\sum_{s=1}^{T} p_T(\mathbf{x}_t, \mathbf{x}_s)\theta(\mathbf{x}_s) + [\mathbf{h}_T(\mathbf{x}_t)]'\mathbf{a} \tag{C.22}$$

where $[\mathbf{h}_T(\mathbf{x}_t)]'$ denotes the tth row of $\mathbf{P}_0(\mathbf{P}_0 + \sigma^2\mathbf{I}_T)^{-1}$. Squaring (C.22) and taking expectations results in

$$\begin{aligned} E[\xi_T(\mathbf{x}_t) - \quad & \alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_t - \ell_T(\mathbf{x}_t)]^2 = \\ & \left\{T^{-1}\sum_{s=1}^{T} p_T(\mathbf{x}_t, \mathbf{x}_s)\theta(\mathbf{x}_s)\right\}^2 + E\{[\mathbf{h}_T(\mathbf{x}_t)]'\mathbf{a}\}^2. \end{aligned} \tag{C.23}$$

By continuity of $\theta(.)$, Theorem 4.4 ensures that for any $\varepsilon > 0$ there exists a T sufficiently large to make $p_T(\mathbf{x}_t, \mathbf{x}_s)\theta(\mathbf{x}_s) < \varepsilon$ for all t, and thus to make

$$\left\{T^{-1}\sum_{s=1}^{T} p_T(\mathbf{x}_t, \mathbf{x}_s)\theta(\mathbf{x}_s)\right\}^2 < \varepsilon$$

as well. Likewise, Lemma 4.5 implies a T such that $E\{[\mathbf{h}_T(\mathbf{x}_t)]'\mathbf{a}\}^2 < \varepsilon$ as well. Hence (C.23) can be made less than 2ε by choosing T sufficiently large. The same is then true of

$$T^{-1}\sum_{t=1}^{T} E[\xi_T(\mathbf{x}_t) - \alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_t - \ell_T(\mathbf{x}_t)]^2,$$

which was to be shown.

Proof of Lemma 4.8.

Recall that $p_0(\mathbf{x}, \mathbf{z})$ is proportional to the ratio of the volume of the overlap of unit spheroids in $(\mathbf{g} \odot \mathbf{x})$-space centered at $\mathbf{x}$ and $\mathbf{z}$ to the volume of a unit spheroid

$$p_0(\mathbf{x}, \mathbf{z}) = (\lambda^2/b) \int_{\mathbf{y} \in [W(\mathbf{x}) \cap W(\mathbf{z})]} d\mathbf{y}$$

where

$$b = \int_{\mathbf{y} \in W(\mathbf{x})} d\mathbf{y}.$$

Note that b, the volume of a unit spheroid, does not depend on $\mathbf{x}$. Recall that the function $\ell(\mathbf{x})$ is representable with respect to $p_0(\mathbf{x}, \mathbf{z})$ if there exists a continuous function $\lambda(\mathbf{z})$ such that

$$\begin{aligned} \ell(\mathbf{x}) &= \int_{\mathbf{z} \in A} p_0(\mathbf{x}, \mathbf{z}) \lambda(\mathbf{z}) \, d\mathbf{z} \\ &= (\lambda^2/b) \int_{\mathbf{z} \in A} \left[\int_{\mathbf{y} \in [W(\mathbf{x}) \cap W(\mathbf{z})]} d\mathbf{y} \right] \lambda(\mathbf{z}) \, d\mathbf{z}. \end{aligned} \tag{C.24}$$

Note further that

$$\{\mathbf{y} \in \Re^k \colon \mathbf{y} \in [W(\mathbf{x}) \cap W(\mathbf{z})]\} = \{\mathbf{y} \in \Re^k \colon [\mathbf{y} \in W(\mathbf{x})] \text{ and } [\mathbf{z} \in W(\mathbf{y})]\}$$

so that (C.24) is equivalent to

$$\ell(\mathbf{x}) = (\lambda^2/b) \int_{\mathbf{y} \in W(\mathbf{x})} \left[\int_{\mathbf{z} \in [W(\mathbf{y}) \cap A]} \lambda(\mathbf{z}) \, d\mathbf{z} \right] \, d\mathbf{y}.$$

Equation (4.23) then follows with $\eta(\mathbf{z}) = (\lambda^2/b)\lambda(\mathbf{z})$.

Proof of Lemma 4.9.

We first establish that

$$\int_{z=y-g^{-1}}^{y+g^{-1}} z^p \, dz = \sum_{j=0}^{p} \alpha_{jp} y^j \tag{C.25}$$

where

$$\alpha_{jp} = \frac{p!}{(p+1-j)!j!} g^{-(p+1-j)}[1-(-1)^{p+1-j}]. \tag{C.26}$$

Result (C.25) follows directly by integrating and applying the binomial expansion:

$$\begin{aligned}
\int_{z=y-g^{-1}}^{y+g^{-1}} z^p \, dz &= (p+1)^{-1}[(y+g^{-1})^{p+1} - (y-g^{-1})^{p+1}] \\
&= \left[(p+1)^{-1} \sum_{j=0}^{p+1} \frac{(p+1)!}{(p+1-j)!j!} \, y^j g^{-(p+1-j)} \right. \\
&\quad \left. - \sum_{j=0}^{p+1} \frac{(p+1)!}{(p+1-j)!j!} \, y^j g^{-(p+1-j)} (-1)^{-(p+1-j)} \right].
\end{aligned}$$

Result (C.25) then follows by collecting terms on y^j and noticing that the coefficient on y^{p+1} is zero. We then have likewise that

$$\begin{aligned}
\int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=y-g^{-1}}^{y+g^{-1}} z^p \, dz \, dy &= \sum_{j=0}^{p} \alpha_{jp} \int_{y=x-g^{-1}}^{x+g^{-1}} y^j \, dy \\
&= \sum_{j=0}^{p} \alpha_{jp} \sum_{i=0}^{j} \alpha_{ij} x^i = \sum_{i=0}^{p} \beta_{ip} x^i
\end{aligned} \tag{C.27}$$

where

$$\beta_{ip} = \sum_{j=i}^{p} \alpha_{jp} \alpha_{ij}. \tag{C.28}$$

It follows from (C.28) that

$$\begin{aligned}
\int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=y-g^{-1}}^{y+g^{-1}} \sum_{p=0}^{r} \gamma_p z^p \, dz \, dy &= \sum_{p=0}^{r} \gamma_p \sum_{i=0}^{p} \beta_{ip} x^i \\
&= \sum_{i=0}^{r} \sum_{p=i}^{r} \gamma_p \beta_{ip} x^i.
\end{aligned} \tag{C.29}$$

Expression (C.29) will be equivalent to (4.25) provided that we choose the γ's so as to satisfy

$$\sum_{p=i}^{r} \gamma_p \beta_{ip} = c_i \tag{C.30}$$

for $i = 0, 1, ..., r$. To satisfy (C.30) for $i = r$, set $\gamma_r = c_r/\beta_{rr}$, which exists since $\beta_{rr} = \alpha_{rr}^2 = (2g^{-1})^2$. To satisfy (C.30) for $i = r-1$, set $\gamma_{r-1} = \beta_{r-1,r-1}^{-1}(c_{r-1} - \gamma_r \beta_{r-1,r})$. Continue in this fashion, setting $\gamma_i = \beta_{ii}^{-1}(c_i - \gamma_r\beta_{ir} - \gamma_{r-1}\beta_{i,r-1} - \cdots - \gamma_{i+1}\beta_{i,i+1})$ for $i = r-1, r-2, ..., 0$.

Proof of Lemma 4.10.

We see by integrating that

$$\begin{aligned}
\int_{z=y-g^{-1}}^{y+g^{-1}} \sin(\omega_p z)\ dz &= \omega_p^{-1} \int_{v=\omega_p(y-g^{-1})}^{\omega_p(y+g^{-1})} \sin(v)\ dv \qquad (C.31)\\
&= \omega_p^{-1} \left\{ \cos[\omega_p(y - g^{-1})] - \cos[\omega_p(y + g^{-1})] \right\} \\
&= \omega_p^{-1} \{ [\cos(\omega_p y)\cos(\omega_p g^{-1}) + \sin(\omega_p y)\sin(\omega_p g^{-1})] \\
&\quad - [\cos(\omega_p y)\cos(\omega_p g^{-1}) - \sin(\omega_p y)\sin(\omega_p g^{-1})] \} \\
&= (2/\omega_p)\sin(\omega_p/g)\sin(\omega_p y).
\end{aligned}$$

Therefore also

$$\begin{aligned}
\int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=y-g^{-1}}^{y+g^{-1}} \sin(\omega_p z)\ dz\ dy &= \int_{y=x-g^{-1}}^{x+g^{-1}} (2/\omega_p)\sin(\omega_p/g)\sin(\omega_p y)\ dy \qquad (C.32)\\
&= [(2/\omega_p)\sin(\omega_p/g)]^2 \sin(\omega_p x).
\end{aligned}$$

It follows from (C.32) that

$$\int_{y=x-g^{-1}}^{x+g^{-1}} \int_{z=y-g^{-1}}^{y+g^{-1}} \sum_{p=0}^{r} \gamma_p \sin(\omega_p z)\ dz\ dy = \sum_{p=0}^{r} \gamma_p [(2/\omega_p)\sin(\omega_p/g)]^2 \sin(\omega_p x),$$

which would succeed in demonstrating (4.27) provided we chose γ_p so that $\gamma_p[(2/\omega_p)\sin(\omega_p/g)]^2 = c_p$. For any c_p, such a γ_p exists provided that $\sin(\omega_p/g) \neq 0$.

Appendix D: Proofs for Section 5

Proof of Lemma 5.1.

Let $\mathbf{K}(\boldsymbol{\theta})$ be the Cholesky factor of the inverse of the matrix $\mathbf{W}(\mathbf{X};\boldsymbol{\theta})$ in (3.23),

$$\mathbf{K}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})' = [\mathbf{W}(\mathbf{X};\boldsymbol{\theta})]^{-1}.$$

Conditional on $\boldsymbol{\theta}$, expression (3.24) implies that $\tilde{\mathbf{y}} = \mathbf{K}(\boldsymbol{\theta})'\mathbf{y}$ is related to $\tilde{\mathbf{X}} = \mathbf{K}(\boldsymbol{\theta})'\mathbf{X}$ according to a classical Normal regression model. Hence the factorization of $f(\tilde{\mathbf{y}}, \boldsymbol{\psi}|\boldsymbol{\theta}, \tilde{\mathbf{X}})$ is well-known; (see for example equation [12.A.13] in Hamilton, 1994), and $f(\mathbf{y}, \boldsymbol{\psi}|\boldsymbol{\theta}, \mathbf{X})$ is simply the Jacobian $|\mathbf{K}(\boldsymbol{\theta})|$ times this expression. Lemma 5.1 then follows immediately by rewriting $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ in terms of $\mathbf{y}, \mathbf{X}$, and $\mathbf{W}(\mathbf{X};\boldsymbol{\theta})$.

Derivation of (5.14).

Observe that the vector $(\boldsymbol{\theta}^{(j)\prime}, \boldsymbol{\zeta}^{(j)\prime})'$ is generated from the density $I(\boldsymbol{\theta})f(\boldsymbol{\zeta}|\boldsymbol{\theta}, \mathbf{Y}_T)$. Dividing the numerator of (5.14) by N, it follows from the law of large numbers that

$$\begin{aligned} N^{-1}\sum_{j=1}^{N} \delta_{[\boldsymbol{\zeta}^{(j)}\in C]} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) \xrightarrow{p} & \int\int \quad \delta_{[\boldsymbol{\zeta}\in C]} w(\boldsymbol{\theta}, \mathbf{Y}_T) I(\boldsymbol{\theta}) f(\boldsymbol{\zeta}|\boldsymbol{\theta}, \mathbf{Y}_T)\, d\boldsymbol{\theta}\, d\boldsymbol{\zeta} \\ = & \int\int \quad \delta_{[\boldsymbol{\zeta}\in C]} f(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X}) f(\boldsymbol{\zeta}|\boldsymbol{\theta}, \mathbf{Y}_T)\, d\boldsymbol{\theta}\, d\boldsymbol{\zeta} \\ = & \int\int \quad \delta_{[\boldsymbol{\zeta}\in C]} f(\boldsymbol{\zeta}, \boldsymbol{\theta}, \mathbf{y}|\mathbf{X})\, d\boldsymbol{\theta}\, d\boldsymbol{\zeta}. \end{aligned} \tag{D.1}$$

Similarly, for the denominator of (5.14),

$$\begin{aligned} N^{-1}\sum_{j=1}^{N} w(\boldsymbol{\theta}^{(j)}, \mathbf{Y}_T) \xrightarrow{p} & \int w(\boldsymbol{\theta}, \mathbf{Y}_T) I(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ = & \int f(\boldsymbol{\theta}, \mathbf{y}|\mathbf{X})\, d\boldsymbol{\theta} \\ = & f(\mathbf{y}|\mathbf{X}). \end{aligned} \tag{D.2}$$

It follows from the ratio of (D.2) to (D.1) that as the number of Monte Carlo draws N increases,

$$\begin{aligned}
\widehat{\Pr}(\boldsymbol{\zeta} \in C|\mathbf{Y}_T) &\xrightarrow{p} \frac{\int\int \delta_{[\boldsymbol{\zeta}\in C]} f(\boldsymbol{\zeta}, \boldsymbol{\theta}, \mathbf{y}|\mathbf{X})\, d\boldsymbol{\theta}\, d\boldsymbol{\zeta}}{f(\mathbf{y}|\mathbf{X})} \\
&= \int\int \delta_{[\boldsymbol{\zeta}\in C]} f(\boldsymbol{\zeta}, \boldsymbol{\theta}|\mathbf{Y}_T)\, d\boldsymbol{\theta}\, d\boldsymbol{\zeta} \\
&= \Pr(\boldsymbol{\zeta} \in C|\mathbf{Y}_T)
\end{aligned}$$

as claimed.

Appendix E: Proofs for Section 6

Proof of Theorem 6.1(a).

The differential of (6.2) is found as in equation (6) in Magnus and Neudecker (1988, p. 315) to be

$$\begin{aligned} d \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta}) &= -(1/2)\mathrm{tr}(\boldsymbol{\Omega}_T^{-1} d\boldsymbol{\Omega}_T) + (1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \\ &\quad -(1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1} d(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right]. \end{aligned} \tag{E.1}$$

To find the derivative with respect to λ^2, set $d\boldsymbol{\Omega}_T = \mathbf{H}_T d\lambda^2$ and notice that $\partial\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'/\partial\lambda^2 = \mathbf{0}$:

$$\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\lambda^2} = -(1/2)\mathrm{tr}(\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T) + (1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T\boldsymbol{\Omega}_T^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]. \tag{E.2}$$

When evaluated at $\lambda^2 = 0$ and $\boldsymbol{\Omega}_T = \sigma^2\mathbf{I}_T$, expression (E.2) becomes

$$\left.\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\lambda^2}\right|_{\lambda^2=0} = -(2\sigma^2)^{-1}\mathrm{tr}(\mathbf{H}_T) + (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \tag{E.3}$$

from which (6.3) follows.

Proof of Theorem 6.1(b).

The expectation of (E.3) is

$$-(2\sigma^2)^{-1}\mathrm{tr}(\mathbf{H}_T) + (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T\sigma^2\mathbf{I}_T) = 0.$$

Recalling from Magnus and Neudecker (1988, p. 151, equation (1)) that $d\boldsymbol{\Omega}_T^{-1} = -\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}$, the differential of (E.2) is

$$d\left[\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\lambda^2}\right] = (1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T\right]$$

$$
\begin{aligned}
&-(1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T\boldsymbol{\Omega}_T^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] - \\
&(1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \\
&+(1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}\mathbf{H}_T\boldsymbol{\Omega}_T^{-1}d(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right]
\end{aligned}
$$

and

$$
\begin{aligned}
d\left[\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\lambda^2}\right]\Bigg|_{\lambda^2=0} &= (2\sigma^4)^{-1}\mathrm{tr}[(d\boldsymbol{\Omega}_T)\mathbf{H}_T] - (2\sigma^6)^{-1}\mathrm{tr}[(d\boldsymbol{\Omega}_T)\mathbf{H}_T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\
&\quad -(2\sigma^6)^{-1}\mathrm{tr}[\mathbf{H}_T(d\boldsymbol{\Omega}_T)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] + (2\sigma^4)^{-1}\mathrm{tr}[\mathbf{H}_T d(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')]
\end{aligned}
$$

from which

$$
\frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial(\lambda^2)^2}\Bigg|_{\lambda^2=0} = (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T^2) - \sigma^{-6}\mathrm{tr}(\mathbf{H}_T^2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \tag{E.4}
$$

$$
\frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\lambda^2\,\partial\sigma^2}\Bigg|_{\lambda^2=0} = (2\sigma^4)^{-1}\mathrm{tr}(\mathbf{H}_T) - \sigma^{-6}\mathrm{tr}(\mathbf{H}_T^2\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'). \tag{E.5}
$$

Similarly we have from (E.1) that

$$
\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\sigma^2} = -(1/2)\mathrm{tr}(\boldsymbol{\Omega}_T^{-1}) + (1/2)\mathrm{tr}(\boldsymbol{\Omega}_T^{-2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')
$$

and

$$
\begin{aligned}
d\left[\frac{\partial \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial\sigma^2}\right] &= (1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\right] \\
&\quad -(1/2)\mathrm{tr}\left\{\left[\boldsymbol{\Omega}_T^{-1}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-2} + \boldsymbol{\Omega}_T^{-2}(d\boldsymbol{\Omega}_T)\boldsymbol{\Omega}_T^{-1}\right]\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right\} \\
&\quad +(1/2)\mathrm{tr}\left[\boldsymbol{\Omega}_T^{-2}d\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]
\end{aligned}
$$

so

$$
\frac{\partial^2 \ln f(\mathbf{y}|\mathbf{X};\boldsymbol{\zeta})}{\partial(\sigma^2)^2}\Bigg|_{\lambda^2=0} = (2\sigma^4)^{-1}T - \sigma^{-6}\mathrm{tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'). \tag{E.6}
$$

Taking expectations of (E.4), (E.5), and (E.6) produces the upper left block of (6.4). The other terms in (6.4) are standard; see for example Magnus and Neudecker (1988, p. 320).

Proof of Theorem 6.1(c).

The upper left block of the inverse of (6.4) is

$$\frac{2\sigma^4}{T\text{tr}(\mathbf{H}_T^2) - [\text{tr}(\mathbf{H}_T)]^2} \begin{bmatrix} T & -\text{tr}(\mathbf{H}_T) \\ -\text{tr}(\mathbf{H}_T) & \text{tr}(\mathbf{H}_T^2) \end{bmatrix}. \tag{E.7}$$

The LM test is found by multiplying (6.3) times the square root of the (1,1) element of (E.7) and evaluating the result at $\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}, \sigma^2 = \hat{\sigma}_T^2$:

$$\aleph_T = (2\hat{\sigma}_T^4)^{-1} \left[\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \hat{\sigma}_T^2\text{tr}(\mathbf{H}_T)\right] \left\{\frac{2\hat{\sigma}_T^4}{\text{tr}(\mathbf{H}_T^2) - T^{-1}[\text{tr}(\mathbf{H}_T)]^2}\right\}^{1/2}$$

from which (6.5) follows.

Proof of Theorem 6.1(d).

Write the numerator of (6.5) as

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \hat{\sigma}_T^2\text{tr}(\mathbf{H}_T) = \hat{\boldsymbol{\varepsilon}}'[\mathbf{H}_T - T^{-1}\mathbf{I}_T\text{tr}(\mathbf{H}_T)]\hat{\boldsymbol{\varepsilon}}. \tag{E.8}$$

Recall (e.g. Hamilton, 1994, equation [8.1.11]) that $\hat{\boldsymbol{\varepsilon}} = [\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{\varepsilon}$, allowing (E.8) to be written

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \hat{\sigma}_T^2\text{tr}(\mathbf{H}_T) = \boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon} + z_T \tag{E.9}$$

where

$$z_T = \boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}'\mathbf{A}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

and

$$T^{-1/2}z_T \;=\; (T^{-1}\boldsymbol{\varepsilon}'\mathbf{X})(T^{-1}\mathbf{X}'\mathbf{X})^{-1}(T^{-1}\mathbf{X}'\mathbf{A}_T\mathbf{X})(T^{-1}\mathbf{X}'\mathbf{X})(T^{-1/2}\mathbf{X}'\boldsymbol{\varepsilon}) \tag{E.10}$$

$$-2(T^{-1}\boldsymbol{\varepsilon}'\mathbf{A}_T\mathbf{X})(T^{-1}\mathbf{X}'\mathbf{X})^{-1}(T^{-1/2}\mathbf{X}'\boldsymbol{\varepsilon})$$

$$\xrightarrow{p} \qquad \mathbf{0}'\mathbf{Q}^{-1}\mathbf{P}\mathbf{Q}^{-1}\mathbf{z} - 2\mathbf{0}'\mathbf{Q}^{-1}\mathbf{0}\mathbf{z}$$

where $\mathbf{z} \sim N(\mathbf{0},\sigma^2\mathbf{Q})$ and use has been made of conditons (i) through (iii). Hence

$$\aleph_T = \frac{\boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon} + z_T}{\hat{\sigma}_T^2\sqrt{2}\left\{\text{tr}(\mathbf{H}_T^2) - T^{-1}[\text{tr}(\mathbf{H}_T)]^2\right\}^{1/2}}. \tag{E.11}$$

Observe from Lemma 8.2 in White (1994, p. 170) that

$$\sigma^{-2}\boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon} = \sum_{t=1}^{T}\lambda_t Q_t \tag{E.12}$$

where $\{Q_t\}$ is an i.i.d. sequence of $\chi^2(1)$ variables and $\{\lambda_t\}$ are the eigenvalues of $\mathbf{A}_T$. Note that (E.12) is consistent with repeated eigenvalues ($\lambda_t = \lambda_s$ for some t and s) and zero eigenvalues.[1] Note further that expression (E.12) has mean zero,

$$E(\boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon}) = \sum_{t=1}^{T}\lambda_t = \text{tr}(\mathbf{A}_T) = 0,$$

and the tth element of the sum in (E.12) has variance $2\lambda_t^2$. Thus for

$$s_T^2 = T^{-1}\sum_{t=1}^{T}2\lambda_t^2, \tag{E.13}$$

we know from Theorem A.3.3 in White (1994, p. 356) that

$$(Ts_T^2)^{-1/2}\sigma^{-2}\boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon} \xrightarrow{L} N(0,1), \tag{E.14}$$

provided that

$$s_T^2 > \delta > 0 \qquad \text{for almost all } T. \tag{E.15}$$

[1] White writes this result slightly differently as $\sum_{i=1}^{c}\lambda_i Q_i$ where $\lambda_1, ..., \lambda_c$ represent c distinct eigenvalues, m_i is the number of times eigenvalue i is repreated, and $Q_i \sim \chi_i^2(m_i)$. Result (E.12) is equivalent since a $\varkappa_i^2(m_i)$ variable can be viewed as the sume of m_i independent $\varkappa^2(1)$ variables.

But notice that s_T^2 can be written

$$s_T^2 = 2T^{-1}\text{tr}(\mathbf{A}_T^2) \tag{E.16}$$

which exceeds δ for all T by condition (iv). Additional algegra reveals that

$$\begin{aligned} \text{tr}(\mathbf{A}_T^2) &= \text{tr}\left\{\left[\mathbf{H}_T - T^{-1}\mathbf{I}_T\text{tr}(\mathbf{H}_T)\right]^2\right\} \\ &= \text{tr}\left\{\mathbf{H}_T^2 - 2T^{-1}\mathbf{H}_T\text{tr}(\mathbf{H}_T) + T^{-2}\mathbf{I}_T\left[\text{tr}(\mathbf{H}_T)\right]^2\right\} \\ &= \text{tr}(\mathbf{H}_T^2) - T^{-1}[\text{tr}(\mathbf{H}_T)]^2. \end{aligned} \tag{E.17}$$

Substituting (E.16) and (E.17) into (E.14) establishes that

$$\frac{\boldsymbol{\varepsilon}'\mathbf{A}_T\boldsymbol{\varepsilon}}{\sigma^2\sqrt{2}\left\{\text{tr}(\mathbf{H}_T^2) - T^{-1}[\text{tr}(\mathbf{H}_T)]^2\right\}^{1/2}} \xrightarrow{L} N(0,1).$$

Using these results along with (E.10) and the familiar property that $\hat{\sigma}_T^2 \xrightarrow{p} \sigma^2$, it follows from (E.11) that $\aleph_T \xrightarrow{L} N(0,1)$, as claimed.

Proof of Lemma 5.2.

Note that the numerator of (6.7) can be written

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \tilde{\sigma}_T^2\text{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T) = \hat{\boldsymbol{\varepsilon}}'\left[\mathbf{H}_T - (T-k-1)^{-1}\mathbf{I}_T\text{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T)\right]\hat{\boldsymbol{\varepsilon}}. \tag{E.18}$$

Since $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}_T\boldsymbol{\varepsilon}$, expression (E.18) can be written

$$\hat{\boldsymbol{\varepsilon}}'\mathbf{H}_T\hat{\boldsymbol{\varepsilon}} - \tilde{\sigma}_T^2\text{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T) = \boldsymbol{\varepsilon}'\tilde{\mathbf{A}}_T\boldsymbol{\varepsilon} \tag{E.19}$$

for

$$\tilde{\mathbf{A}}_T = \mathbf{M}_T\mathbf{H}_T\mathbf{M}_T - (T-k-1)^{-1}\text{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T) \tag{E.20}$$

where use has been made of the fact that $\mathbf{M}_T$ is symmetric and idempotent. Substituting (E.19) and (E.20) into (6.7) establishes that

$$\tilde{\aleph}_T = \frac{\boldsymbol{\varepsilon}'\tilde{\mathbf{A}}_T\boldsymbol{\varepsilon}}{\left[2\mathrm{tr}\left(\tilde{\mathbf{A}}_T^2\right)\right]^{1/2}}. \tag{E.21}$$

But we know from Theorem 12 of Magnus and Neudecker (1988, p. 251) that

$$E(\boldsymbol{\varepsilon}'\tilde{\mathbf{A}}_T\boldsymbol{\varepsilon}) = \sigma_0^2\mathrm{tr}(\tilde{\mathbf{A}}_T) \tag{E.22}$$

$$\mathrm{Var}(\boldsymbol{\varepsilon}'\tilde{\mathbf{A}}_T\boldsymbol{\varepsilon}) = 2\sigma_0^4\mathrm{tr}\left(\tilde{\mathbf{A}}_T^2\right). \tag{E.23}$$

Further, from (E.20),

$$\mathrm{tr}(\tilde{\mathbf{A}}_T) = \mathrm{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T) - (T-k-1)^{-1}\mathrm{tr}(\mathbf{M}_T)\mathrm{tr}(\mathbf{M}_T\mathbf{H}_T\mathbf{M}_T)$$

which equals zero from the familiar result that $\mathrm{tr}(\mathbf{M}_T) = T-k-1$. Hence (E.21) has mean zero and variance σ_0^4.