ISSN 1440-771X
ISBN 0 7326 1044 3

MONASH UNIVERSITY

AUSTRALIA

Nonparametric seemingly unrelated regression

Michael Smith and Robert Kohn

Working Paper 7/98
June 1998

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Nonparametric Seemingly Unrelated Regression.

Michael Smith[a] and Robert Kohn[b]

[a] Department of Econometrics and Business Statistics, Monash University

[b] Australian Graduate School of Management, University of New South Wales

This Version February 1998

This papers presents a method for simultaneously estimating a system of nonparametric multiple regressions which may seem unrelated, but where the errors are potentially correlated between equations. We show that the prime advantage of estimating such a 'seemingly unrelated' system of nonparametric regressions is that substantially less observations can be required to obtain reliable function estimates than if each of the regression equations was estimated separately and the correlation ignored. This increase in efficiency is investigated empirically using both simulated and real data. The method suggested here develops a Bayesian hierarchical framework where the regression function is represented as a linear combination of a large number of basis terms, the number of which is typically greater than the sample size. All the regression coefficients, and the variance matrix of the errors, are estimated simultaneously using their posterior means. The computation is carried out using a Markov chain Monte Carlo sampling scheme that employs a 'focused sampling' step to combat the high dimensional representation of the function and a Metropolis-Hastings step to correctly account for the distribution of the covariance matrix. The methodology is also easily extended to other nonparametric multivariate regression models.

Key Words: Nonparametric Multivariate Regression, Bayesian hierarchical SUR model, Multivariate Subset Selection, Markov Chain Monte Carlo, Focused sampling.

1 Introduction

The aim of nonparametric regression is to estimate regression functions without assuming *a priori* knowledge of their functional forms. The price for this flexibility is that appreciably larger sample sizes are required to obtain reliable nonparametric estimators than for parametric estimators. In this paper we consider a system of regression equations that can seem unrelated, but actually are because their errors are correlated. Such a system of equations is called a set of 'seemingly unrelated' regressions, or a SUR model (Zellner, 1962). This paper provides a Bayesian framework for reliably estimating the regression functions in a nonparametric manner, even for moderate sample sizes, by taking advantage of the correlation structure in the errors. The most important consequence of this work is to show that if the errors are correlated, substantially better nonparametric estimators are obtained by taking advantage of this correlation structure compared to ignoring the correlation and estimating the equations one at a time.

Specifically, we consider the system of m regression equations

$$y^i = f^i(x^i) + e^i \quad \text{for } i = 1, 2, \ldots, m \tag{1.1}$$

Here, the superscript denotes that this is the ith of m possible regressions, y^i is the dependent variable, x^i is a vector of r^i independent variables and $f^1, \ldots, f^m$ are functions that require estimating in a nonparametric manner. As in the linear Gaussian SUR model, the regressions are related through the correlation structure of the Gaussian errors e^i. That is,

$$\mathbf{e} \sim N(0, \Sigma \otimes I_n) \tag{1.2}$$

where $\mathbf{e}' = (\mathbf{e}^{1\prime}, \mathbf{e}^{2\prime}, \ldots, \mathbf{e}^{m\prime})$, $\mathbf{e}^i$ is the vector of errors for the n observations of the ith regression and Σ is a positive definite $(m \times m)$ matrix that also requires estimation. This paper is concerned with providing a data-driven procedure for estimating the unknown functions f^i (for $i = 1, \ldots, m$) and covariance matrix Σ in this model.

Such systems of regressions are frequently used in econometric, financial and sociological modeling because taking into account the correlation structure in the errors results in more efficient estimates than ignoring the correlation and estimating the equations one at a time.

However, most of the literature on estimating a system of equations assumes that the f^i are linear functions. For recent examples, see Bartels, Fiebig and Plumb (1996), Min and Zellner (1993) and Mandy and Martins-Filho (1993). However, in practice the functional forms of the f^i in most regression applications are unknown *a priori*, so that an approach that estimates their form is preferable. We illustrate the need for a nonparametric approach and the gains in efficiency obtained by estimating a system of equations in section 4 by applying our methodology to two real data examples. The first concerns print advertisements in an Australian women's magazine and estimates the relationship between three measures of advertising exposure and the positioning of advertisements in the magazine. The second estimates the relationship between monthly returns and some key macroeconomic variables for five large mining companies listed on the Australian stock exchange. In both examples, significant nonlinear relationships are identified that would have been difficult to discern using a parametric SUR approach. In addition, the estimates are shown to differ substantially from those that arise from estimating each of the nonparametric regressions separately and ignoring the correlation between the equations.

Our approach for estimating the system of equations defined at (1.1) and (1.2) models each of the functions f^i as a linear combination of basis terms. We develop a Bayesian hierarchical model to explicitly parameterize the possibility that these terms may be superfluous and have corresponding coefficient values equal to exactly zero. A wide variety of bases can be used, including many with a desired structure, such as periodicity or additivity. The unknown regression functions are estimated by their posterior means which attach the proper posterior probability to each subset of the basis elements, providing a nonparametric estimate that is both flexible and smooth. We develop a Markov chain Monte Carlo (MCMC) sampling scheme to calculate the posterior means because direct enumeration is intractable. This sampling scheme is a generalization of the 'focused sampler' discussed in Wong, Hansen, Kohn and Smith (1997) and is shown to be reliable and fast. The performance of the new estimator is demonstrated empirically with a set of simulation experiments carefully designed to cover a range of potential regression curves, noise levels and several commonly employed bases. These highlight the improvement that can be made by exploiting correlation structure

in a system of regressions. We note that the solution to the nonparametric SUR model focused on here is easily extendible to other nonparametric multivariate (or vector) regression models, such as where the functions f^i are the same for all $i = 1, \ldots, m$.

Zellner (1962, 1963) provides the seminal analysis of a system of regressions when the unknown functions f^i are assumed linear in the coefficients. Srivastava and Giles (1987) summarize much of the literature dealing with this linear SUR model. However, recent advances in Markov chain Monte Carlo methodology enable Bayesian analyses of more complex variations of the SUR model. For example, Chib and Greenberg (1995a) develop sampling schemes that estimate a hierarchical linear SUR model with first order vector autoregressive or vector moving average errors and extend the analysis to a time varying parameter model. Markov chain Monte Carlo methods have also provided a solution to reliably estimating nonparametric regressions in a variety of hitherto difficult situations. For example, Smith and Kohn (1996) and Wong, et al. (1997) develop nonparametric regression estimators for regression models where a data transformation may be required and/or outliers may exist in the data.

The paper is organized as follows. Section 2 discusses how to model the unknown functions and why they are estimated using a hierarchical model. Section 3 discusses this Bayesian hierarchical model and develops a MCMC sampling scheme to enable its estimation. Section 4 uses the methodology to fit the Australian print advertising and mining returns datasets. Section 5 contains simulation examples which investigate the improvements that can be made using this estimation procedure over a series of separate nonparametric regressions, as well as comparing a variety of commonly used bases.

2 Basis representation of functions

Each regression function is modeled as a linear combination of basis functions, so that for a function f,

$$f(x) = \sum_{i=1}^{p} \beta_i b_i(x) \tag{2.1}$$

Here, $\mathcal{B} = \{b_1, \ldots, b_p\}$ is a basis of p functions, while the β_i's are regression parameters.

A number of authors have used such an approach in the single equation case using a variety of univariate and higher dimensional bases. For example, Friedman and Silverman (1989), Friedman (1991), Smith and Kohn (1996) and Denison, Mallick and Smith (1997) use regression splines, Luo and Wahba (1997) use several reproducing kernel bases and Donoho and Johnstone (1994) use wavelet bases to estimate non-smooth functions. In the empirical work in this paper we consider the following bases for a univariate independent variable x, with n observations $x_1, x_2, \ldots, x_n$.

1. **Cubic Regression Spline**: Here, $\mathcal{B}_1 = \{1, x, x^2, x^3, (x-x_1)_+^3, \ldots, (x-x_n)_+^3\}$, where $(\cdot)_+^3 = \max(0, \cdot)$ and $p = n+4$. A similar basis was used in Smith and Kohn (1996), but with a smaller number of basis functions. However, by employing the focused sampler discussed in section 3.2, bases with $p > n$ terms can be used.

2. **Quadratic Regression Spline**: Here, $\mathcal{B}_2 = \{1, x, x^2, (x-x_1)_+^2, \ldots, (x-x_n)_+^2\}$ and $p = n+3$.

3. **Quartic Reproducing Kernel**: Here, $\mathcal{B}_3 = \{b_1(x), \ldots, b_n(x), x, 1\}$, where

$$b_i(x) = \frac{1}{24}\left((|x-x_i| - \frac{1}{2})^4 - \frac{1}{2}(|x-x_i| - \frac{1}{2})^2 + \frac{7}{240}\right), \quad \text{for } i = 1, \ldots, n,$$

and $p = n+2$. This basis was discussed in Luo and Wahba (1997) and is defined over the unit interval, so that we simply scale the independent variable before calculation of $\mathcal{B}_3$, so that $\min(x) = 0$ and $\max(x) = 1$.

4. **Thin Plate Spline**: Here, $\mathcal{B}_4 = \{b_1(x), \ldots, b_n(x), x, 1\}$, where

$$b_i(x) = |x-x_i|^2 \log(|x-x_i|),$$

and $p = n+2$. This is an example of a radial basis function of the type discussed by Powell (1987) and was used in thin plate smoothing by Wahba (1990).

5. **Natural Cubic Spline**: This is a basis discussed in Wahba (1990), where $\mathcal{B}_5 = \{b_1(x), \ldots, b_n(x), x, 1\}$ and

$$b_i(x) = \begin{cases} \frac{1}{2}x^2(x_i - \frac{1}{3}x) & \text{if } x \leq x_i \\ \frac{1}{2}x_i^2(x - \frac{1}{3}x_i) & \text{if } x > x_i \end{cases} \quad \text{for } i = 1, \ldots, n.$$

The basis elements are assumed to be distinct, those that are not we simply remove. In the absence of replicated design points, the number of terms in each basis is $p > n$. All the above bases are known to be suited to approximating univariate functions that are both continuous and continuously differentiable.

In the case of multiple regressors in an equation a variety of other bases can be used, including tensor products of univariate bases (Friedman, 1991) and radial bases (Powell, 1987; Holmes and Mallick, 1997). In this paper we use additive combinations of the above univariate bases, so that for an r dimensional independent variable $\boldsymbol{x}$, the basis is

$$\mathcal{M} = \mathcal{B}^1 \cup \mathcal{B}^2 \cup \ldots \cup \mathcal{B}^r .$$

Here, $\mathcal{B}^i$ is a univariate basis for the ith element of $\boldsymbol{x}$ and $\mathcal{M}$ is the resulting multivariate basis. The number of elements in this basis is generally $p \approx rn$ (approximate because the number of replicated design points is unknown) and the model is made identifiable by retaining only a single intercept.

Given a choice of a particular basis for the approximation at (2.1), the ith regression at (1.1) can be written as the linear model

$$\mathbf{y}^i = X^i \boldsymbol{\beta}^i + \mathbf{e}^i . \tag{2.2}$$

Here, $\mathbf{y}^i$ is the vector of the n observations of the dependent variable, the design matrix $X^i = [\mathbf{b}_1 | \mathbf{b}_2 | \ldots | \mathbf{b}_{p^i}]$, $\mathbf{b}_j$ is a vector of the values of the basis function b_j evaluated at the n observations and $\boldsymbol{\beta}$ are the regression coefficients. The errors $\mathbf{e}^i$ are correlated with those from the other regressions, as specified in (1.2), and we denote the number of basis terms in the ith equation as p^i. It is inappropriate to estimate the regression coefficients using existing SUR methodology for three reasons. First, the columns of X^i are not generally linearly independent because usually $p^i > n$ and if there are several explanatory variables then $p^i \approx rn$. Second, even if a maximal linearly independent subset of columns was identified the resulting estimates of the regression coefficients would have high variance and the function estimate $\hat{f}^i$ would interpolate the data (rather than produce smooth estimates that account for the existence of noise in the regression). Third, it is difficult to identify one superior

linearly independent subset of basis terms over another . Therefore, we estimate the regression parameters using a Bayesian hierarchical SUR model (described below in section 3) that explicitly accounts for the possibility that many of these terms may be redundant. It is by estimating the regression parameters using this procedure, rather than simply by generalized least squares, that makes the function estimates nonparametric.

3 A Bayesian Hierarchical SUR Model

3.1 Modeling variable redundancy

Consider the ith regression of a linear SUR model given at equation (2.2), where the design matrix X^i is $(n \times p^i)$ and the coefficient vector β^i is of length p^i. To explicitly account for the notion that variables in this regression can be redundant, we introduce a vector of binary indicator variables $\gamma^i = (\gamma^i_1, \gamma^i_2, \ldots, \gamma^i_{p^i})'$. Here, γ^i_k corresponds to the kth element of the coefficient vector of the ith regression, say β^i_k, with $\gamma^i_k = 0$ if $\beta^i_k = 0$ and $\gamma^i_k = 1$ if $\beta^i_k \neq 0$. By dropping the redundant terms with zero coefficients, the ith regression can be rewritten, conditional on γ^i, as

$$\mathbf{y}^i = X^i_{\gamma^i} \beta^i_{\gamma^i} + \mathbf{e}^i \tag{3.1}$$

If $q^i_\gamma = \sum_{j=1}^{p^i} \gamma^i_j$, then the design matrix $X^i_{\gamma^i}$ is of size $(n \times q^i_\gamma)$ and $\beta^i_{\gamma^i}$ is a vector of q^i_γ elements.

By stacking together the linear models for the m regressions, the SUR model can also be written, conditional on $\gamma' = (\gamma^{1\prime}, \gamma^{2\prime}, \ldots, \gamma^{m\prime})$, so that

$$\mathbf{y} = X_\gamma \beta_\gamma + \mathbf{e} \tag{3.2}$$

Here, $\mathbf{y}' = (\mathbf{y}^{1\prime}, \mathbf{y}^{2\prime}, \ldots, \mathbf{y}^{m\prime})$, $X_\gamma = \mathrm{diag}(X^1_{\gamma^1}, X^2_{\gamma^2}, \ldots, X^m_{\gamma^m})$ and $\beta'_\gamma = (\beta^1_{\gamma^1}{}', \ldots, \beta^m_{\gamma^m}{}')$. If $q_\gamma = \sum_{i=1}^m q^i_\gamma$, then X_γ is an $(nm \times q_\gamma)$ matrix and β_γ a vector of q_γ elements. To complete this Bayesian hierarchical model, we introduce the following priors on the parameters.

(i) In a similar manner as O'Hagan (1995) we construct a conditional prior for β_γ by

setting

$$p(\beta_\gamma|\Sigma,\gamma) \propto p(\mathbf{y}|\beta_\gamma,\gamma,\Sigma)^{1/nm}$$

so that $\beta_\gamma|\Sigma,\gamma \sim N(\hat{\mu}_\gamma, nm(X'_\gamma A X_\gamma)^{-1})$, where $A = \Sigma^{-1}\otimes I_n$ and $\hat{\mu}_\gamma = (X'_\gamma A X_\gamma)^{-1} X'_\gamma A\mathbf{y}$. This data-based fractional prior contains much less information about β_γ than the likelihood.

(ii) The prior for Σ^{-1} is taken as independent of γ and is the commonly used non-informative prior discussed in Zellner (1971), where $p(\Sigma^{-1}|\gamma) \propto |\Sigma^{-1}|^{-(m+1)/2}$.

(iii) The γ_k^i are taken as *a priori* independent of one another with $p(\gamma_k^i = 1) = 1/2$ throughout this paper.

Note that the model here is a hierarchical SUR model as, conditional on γ, it is simply a linear SUR model; and that it is through the conditional prior for β_γ that γ is introduced into the model.

3.2 Markov chain Monte Carlo sampling

To estimate this model we use the following Markov chain Monte Carlo sampling scheme.

(1) Generate from $\beta_\gamma|\Sigma^{-1},\gamma,\mathbf{y}$

(2) Generate from $\Sigma^{-1}|\beta,\gamma,\mathbf{y} = \Sigma^{-1}|\beta_\gamma,\gamma,\mathbf{y}$

(3) For $i = 1,2,\ldots,m$

Choose $C_i \subset \{1,2,\ldots,p^i\}$ in the random manner discussed below.

(4) Repeat the following K times

For $i = 1,2,\ldots,m$

Generate from $\gamma_j^i|\Sigma^{-1},\gamma\backslash\gamma_j^i,\mathbf{y}$ for $j \in C_i$

In this sampling scheme β_γ is generated from a multivariate normal distribution and γ_j^i is generated from a binomial. Generation of the matrix Σ^{-1} directly from the posterior at step (2) is difficult because the fractional prior $\beta_\gamma|\Sigma^{-1},\gamma$ is centered at $\hat{\mu}_\gamma$, which is a

function of Σ. Consequently, we use a Metropolis-Hastings step where the proposal Wishart density is the posterior under a flat conditional prior for β_γ. This works well with between 60% and 90% of those iterates that are generated being accepted. Details of how to generate from the distributions at steps (1), (2) and (4) are given in the appendix. It is important to note that care has been taken to generate γ_j^i without conditioning on β_j^i at step (4), otherwise the sampling scheme would be reducible because γ_j^i is known exactly given β_j^i.

Step (3) is a 'focusing step' similar to that discussed in Wong et al. (1997) and is undertaken for each equation $i = 1, \ldots, m$. The idea is to identify a subset of the binary variables γ_j^i which are relatively more likely to be 'active' (that is, variables where $\gamma_j^i = 1$ and therefore the corresponding regression coefficients are non-zero) and focus most attention on these. This is important even in single equation nonparametric regression because the bases used can employ greater than n terms, most of the regression coefficients of which have a high probability of being zero. Focusing takes on a new importance in nonparametric SUR models because there are m times as many terms again as in the single equation model.

We use a 'focusing rule' to identify the variables to be generated at the jth iteration of the sampling scheme, which are indexed by the indexing set C_i for each of the i equations. The rule we use is to generate all the binary variables that were active last iteration, plus a randomly selected set of those that were inactive. Each previously inactive binary variable is selected to be generated with probability

$$\alpha = \max\left(20/(p^i - q_\gamma^i), q_\gamma^i/(p^i - q_\gamma^i)\right) .$$

This ensures that on average at least 20 previously inactive terms in the ith equation are generated, while more terms are generated for functions that require a lot of basis functions, so that $q_\gamma > 20$. This ensures that the sampler can move quickly and efficiently around the support of the posterior distribution.

Because the focus sets are selected in a random manner, the sampling scheme is irreducible and aperiodic, so that by Tierney (1994) it converges to its invariant distribution, which is the posterior distribution $\Sigma^{-1}, \gamma, \beta|\mathbf{y}$. It is both an order faster than a Gibbs sampling alternative that generates all the elements of γ one at a time, (that is, where $C_i = \{1, 2, \ldots, p^i\}$)

and possesses stronger convergence properties. The latter is because, at any iteration j, step (4) forms a Gibbs sub-chain of K iterations which converges to the conditional posterior distribution of the 'block' $\gamma_C|\Sigma^{-1}, \gamma\backslash\gamma_C, \mathbf{y}$, where γ_C is defined here to be all the binary variables to be generated.

Given an initial state for the Markov chain and a 'warmup period', after which the sampler is assumed to have converged to the joint posterior distribution, we can collect iterates $(\Sigma^{-1[1]}, \gamma^{[1]}, \beta^{[1]}), \ldots, (\Sigma^{-1[J]}, \gamma^{[J]}, \beta^{[J]})$ which form a Monte Carlo sample from the joint posterior distribution. It is this sample that we use for inference.

A sampler that generates solely from the parameter space of γ is not considered as it is difficult to generate from the posterior distribution $\gamma_j^i|\gamma\backslash\gamma_j^i, \mathbf{y}$. Similarly, samplers that generate from either the parameter space of (γ, Σ^{-1}) or (γ, β) are not considered because it appears difficult to generate from either the conditional posterior distribution $\Sigma^{-1}|\gamma, \mathbf{y}$, or $\gamma_j^i|\gamma\backslash\gamma_j^i, \beta_\gamma, \mathbf{y}$.

We have found this sampler to have strong empirical convergence properties– usually converging to a stable distribution (as witnessed by the marginal distributions of the parameters) in a handful of iterations. This appears to occur regardless of the initial starting state, which is best demonstrated by the fact that all of the very different examples in this paper had the same initial state of $\beta^{[0]} = \mathbf{0}$, $\Sigma^{-1[0]} = I_m$ and $\gamma^{[0]} = \mathbf{0}$. Any other arbitrary feasible state also appears to work fine. The overall reliability and efficiency of the scheme are demonstrated in section 5.

3.3 Estimation

Inference about the unknown functions and parameters is based on the Monte Carlo sample obtained from the sampling scheme. Here, we only consider posterior means, but higher posterior moments and diagnostic statistics (such as residuals) can be handled similarly.

The posterior mean of the regression parameters, $E[\beta|\mathbf{y}]$, is estimated using the mixture

estimate

$$\hat{\beta} = \frac{1}{J}\sum_{j=1}^{J} E\left[\beta|\gamma^{[j]}, \Sigma^{-1[j]}, \mathbf{y}\right] \tag{3.3}$$

Each of the conditional expectations in the sum is simple to calculate because $E[\beta_\gamma|\gamma, \Sigma^{-1}, \mathbf{y}] = \hat{\mu}_\gamma$, while elements of β that are not common to β_γ are set exactly to zero.

The posterior mean $E[\Sigma|\mathbf{y}]$ is estimated by the histogram estimate $\hat{\Sigma} = \left(\frac{1}{J}\sum_{j=1}^{J}\Sigma^{-1[j]}\right)^{-1}$. We do not use a mixture estimate because the distribution of $\Sigma^{-1}|\beta_\gamma, \gamma, \mathbf{y}$ is difficult to identify (which is also the reason a Metropolis-Hastings step is used at step (2) of the sampler).

The posterior means $E[f^i(\mathbf{z})|\mathbf{y}]$ of the functions at equation (1.1) at any point $\mathbf{z}$ in the domain of x^i is estimated using the mixture estimate

$$\hat{f}^i(\mathbf{z}) = \frac{1}{J}\sum_{j=1}^{J} E\left[f^i(\mathbf{z})|\gamma^{[j]}, \Sigma^{-1[j]}, \mathbf{y}\right] = v'\left(\frac{1}{J}\sum_{j=1}^{J} E\left[\beta^i|\gamma^{[j]}, \Sigma^{-1[j]}, \mathbf{y}\right]\right) = v'\hat{\beta}^i$$

Here, $v = (b_1(\mathbf{z}), \ldots, b_s(\mathbf{z}))'$ is a vector containing the basis function expansion of the function f^i evaluated at the point $\mathbf{z}$. The vector $\hat{\beta}^i$ is made up of the elements of $\hat{\beta}$ that correspond to β^i. If the function is univariate, so that x^i is a scalar, then $\hat{f}^i$ is an estimate of a curve, while for higher dimensions it is a surface. For additive nonparametric models the component function estimates can easily be calculated separately by identifying the basis terms and regression coefficient estimates that correspond to each function and forming the inner product of these.

4 Real Data Examples

4.1 Australian Print Advertising Data

In this section we demonstrate our procedure using $n = 457$ observations of data from six issues of an Australian monthly women's magazine. Each observation corresponds to an advertisement placed in the magazine and the following three advertisement exposure scores, which are recorded from an experimental audience, are used as measures of the various levels of effectiveness of the print advertisement.

y^1 (Noted Score): Proportion of respondents who claim to recognize the ad as having been seen by them in that issue.

y^2 (Associated Score): Proportion of the respondents who claim to have noticed the advertiser's brand or company name or logo.

y^3 (Read-Most Score): Proportion of respondents who claim to have read half or more of the copy.

These scores from y^1 to y^3 are thought to measure advertisement exposure at increasing levels of depth.

It long been thought that the positioning of an advertisement within an issue has an effect on its exposure to an audience (Hanssens and Weitz, 1980). To quantify this we constructed the variable P as

$$P = \frac{\text{page number}}{\text{number of pages in issue}}$$

to represent the position in the issue in which each advertisement appeared. Figures 1(a)-(c) provide scatter plots of P verses y^1, y^2 and y^3, respectively.

To estimate the effect the design variable P has on the exposure of a print advertisement, we considered the three nonparametric regressions

$$y^i = f^i(P) + e^i \qquad \text{for } i = 1, 2, 3$$

where the thin plate spline basis $\mathcal{B}_4$ is used to model f^i, for $i = 1, 2, 3$. Expected features in the functions f^i include high casual attention to advertisements placed in the front (and to a lesser extent back) of the magazine, while the pre-editorial slots (where P is about 0.7) are thought to attract more indepth attention.

The three regressions were estimated one at a time using the single equation analogy of the estimator introduced in this paper (where $\Sigma = I_3$) and the resulting function estimates plotted in figures 1(a)-(c) as dashed lines. However, the three scores y^1, y^2 and y^3 are highly positively correlated and it is likely that the assumption of independence in the errors is inappropriate. Therefore, we also estimated the equations as a nonparametric SUR (NSUR)

system. The estimate of the covariance and correlation matrix were

$$\hat{\Sigma} = 0.02 \begin{bmatrix} 1.050 & 1.024 & 0.586 \\ & 1.092 & 0.622 \\ & & 0.486 \end{bmatrix} \qquad \text{Estimated Correlation} = \begin{bmatrix} 1.000 & 0.956 & 0.819 \\ & 1.000 & 0.854 \\ & & 1.000 \end{bmatrix}$$

confirming the existence of high correlation, especially between the pair y^1 and y^2 and the pair y^2 and y^3. The function estimates $\hat{f}^i, \quad (i = 1, 2, 3)$ are also plotted in figures 1(a)-(c) as bold lines. They demonstrate that the front (and to a lesser extent) back of the magazine are areas in which advertisements achieve higher average exposure; though this is more prominent for the noted and associated scores, y^1, y^2, than for the read-most score y^3. The pre-editorial slots also result in increased exposure, with a particularly positive effect on indepth exposure, as measured by y^3. The function estimates differ substantially from those provided by single equation estimation and reveal that taking the correlation into account can seriously alter the function estimates. In addition, the relationships are distinctly nonlinear and would be hard to discern using parametric SUR estimation.

To help confirm that the NSUR estimates had correctly captured the apparent relationships between y^1, y^2, y^3 and P, we calculated Monte Carlo estimates of the posterior mean of the standardized uncorrelated residuals $\mathrm{r} = (R \otimes I_n)\mathrm{e}$, where $R'R = \Sigma^{-1}$. The estimate was calculated from the Monte Carlo sample as

$$E[\mathrm{r}|\mathrm{y}] \approx \hat{\mathrm{r}} = \frac{1}{J} \sum_{i=1}^{J} (R^{[j]} \otimes I_n)\mathrm{e}^{[j]},$$

where $R^{[j]}$ is a Cholesky factor, such that $R^{[j]\prime} R^{[j]} = \Sigma^{-1[j]}$, and $\mathrm{e}^{[j]} = \mathrm{y} - X\beta^{[j]}$. Note that as $\mathrm{r} \sim N(\mathbf{0}, I_{nm})$, it is expected that $\hat{\mathrm{r}}$ should have the approximately the same distribution. Figures 1(d)-(f) plot the standardized uncorrelated residuals $\hat{\mathrm{r}}^i$ corresponding to the three equations, where we have partitioned the vector $\hat{\mathrm{r}} = (\hat{\mathrm{r}}^1, \hat{\mathrm{r}}^2, \hat{\mathrm{r}}^3)$. They appear randomly distributed and seem to confirm that the functions f^1, f^2 and f^3 were estimated without bias.

4.2 Australian Mining Returns

To demonstrate the use of our methodology to systems of additive regressions we apply it to data concerning five large mining companies publicly listed on the Australian Stock Exchange: BHP, CRA, CMC, MIM and WMC. The data was collected for $n = 227$ consecutive months from December 1972 to November 1991. The dependent variable for the ith regression is the respective company's dividend adjusted return, which is defined to be $R_t^i = \ln(P_t^i + D_t^i) - \ln(P_{t-1}^i)$, where P_t^i is the stock price of company i at time t and D_t^i is equal to the dividend payment of company i over the period $(t-1, t]$. The independent variables are the macroeconomic variables given below.

- O_t: Change in the logarithm of the All Ordinaries (the major Australian stock index) at time t.
- X_t: Change in the logarithm of the real exchange rate (\$US/\$AUS) at time t.
- G_t: Change in the logarithm of the gold price at time t.

To investigate how the returns for each company related to these key macroeconomic variables, we posited the following nonparametric additive SUR model.

$$R_t^i = f_1^i(O_t) + f_2^i(X_t) + f_3^i(G_t) + e_t^i \quad \text{for } i = 1, \ldots, 5$$

Here, the regressions were labeled in the following order: $i = 1$ for BHP, $i = 2$ for CRA, $i = 3$ for CMC, $i = 4$ for MIM and $i = 5$ for WMC. No time dependency in the mean returns was considered as past returns are thought to have little, or no, information regarding future mean returns due to arbitrage arguments, though the errors are likely to be correlated across stocks.

We modeled the functions f_j^i using the cubic regression spline basis $\mathcal{B}_1$ and fit the model both as five separate nonparametric regressions and using the nonparametric SUR (or NSUR) estimator. The regression function estimates using the two approaches are given in figure 2 and differ substantially, demonstrating the difference that modeling potential correlation can

make in real data with moderate sample sizes. The NSUR estimated the variance of the errors as

$$\hat{\Sigma} = 0.01 \begin{bmatrix} 0.459 & 0.170 & 0.289 & 0.300 & 0.304 \\ & 0.902 & 0.479 & 0.432 & 0.440 \\ & & 0.964 & 0.691 & 0.632 \\ & & & 0.985 & 0.664 \\ & & & & 0.980 \end{bmatrix}$$

where all the stocks are positively correlated, even after the common effect of changes in the All Ordinaries (O_t) is removed. This is not surprising as all companies have heavy interests in Australian mining and/or base metal production. Of particular interest is the correlation between CMC, MIM and WMC, which are the companies that have their balance sheets almost exclusively focused on mining base metal ores during the period, whereas BHP and CRA are more diversified resource companies (Thomas, 1995).

The function estimates are given in figures 2(a)-(o); one panel for each of the fifteen function estimates $\hat{f}_j^i$, $i = 1, \ldots, 5$ and $j = 1, 2, 3$. A density estimate of the respective independent variable (O_t, X_t or G_t) has been included on top of each plot. Each of these plots has been produced over the domain of the middle 95% of the observations of the respective independent variable. This is because the independent variables have extreme outliers in the x-space (due to market shocks) and the resulting scale would distort the results.

—Figure 2 about here—

Figures 2(a)-(e) indicate that the returns of all five companies are highly related to changes in the All Ordinaries, which is reassuring as these companies form a major component of this index. BHP has an almost linear relationship with what would be a slope coefficient close to one, (figure 2(a)) which is not surprising as this company is the largest company listed with the Australian Stock Exchange and the most diversified of the five considered here. However, the relationships between the returns for the other four companies (especially WMC in figure 2(e)) and the All Ordinaries appear distinctly nonlinear. Here, company returns increase more with positive returns on the All Ordinaries index than they decrease with

negative returns on the index. This is because Australian mining and resources returns have proved fairly robust to downturns in general Australian returns during the period in which the data have been collected.

The relationship between these company returns and changes in the exchange rate (X_t) are minor and generally negative, (see figures 2(f)-(j)). This reflects the fact that all these companies export a large amount of ore and/or base metals and an increase in X_t makes their product more expensive. However, it should be be noted that these companies will also gain a short term increase in income on existing contracts already signed. Therefore, it is hard to say what effect individual changes in X_t will have on monthly returns R_t^i; something that appears to be reflected in the indeterminant nature of the estimated relationships found in figures 2(g) and (h).

Figures 2(k)-(o) plot the relationship between company returns and changes in the gold price (G_t). None of these companies are specifically gold miners (Thomas, 1995), but the relationship between the gold price and company returns increases from none for BHP to a significant nonlinear relationship for MIM and WMC (figures 2(n) and (o)). It useful to note that BHP and CRA were the largest and most diversified of the five companies during the period of our data, while CMC and WMC were the smallest and least diversified with an especially high focus on base metals. Therefore, it is possible that the gold price is capturing an effect that is peculiar to these undiversified base metal miners.

Overall, the estimates explain quite a large percentage of variation in the company returns for the companies. Many of the more interesting relationships appear distinctly nonlinear and would not be captured by simply fitting a linear in parameters SUR.

5 Simulation Experiments

The performance of the nonparametric SUR estimator is studied using simulated data. Yee and Wild (1996) use smoothing splines to estimate a system of equations in a nonparametric manner, but they do not have data-driven estimators for their smoothing parameters. In the example in section 5 of their paper they use values of the smoothing parameters based on

the independent variable, but not the dependent variable. Such an approach is not a satisfactory way of estimating the smoothing parameters because it does not take into account the curvature exhibited by the dependent variable. Nor is it fully automatic in that a value for the effective degrees of freedom has to be chosen by the user. For these reasons we do not include the Yee and Wild (1996) estimator in our simulations and instead compare our estimator with one that estimates each regression equation separately, ignoring any correlation between the regressions. In doing this we show that this can result in substantially improved estimates.

5.1 Example 1: Highly positively correlated univariate regressions

This simulated example highlights the case where the errors are highly correlated between regressions, with the true covariance matrix Σ being given below at (5.1). There are $m = 4$ univariate regressions, so that $r_1 = r_2 = r_3 = r_4 = 1$ and the standard deviation of the errors $(\text{var}(e^i))^{1/2} = 1$ is high compared to the range of the functions.

Four true functions were carefully chosen to represent a wide variety of possible relationships. These are $f^1(x) = \sin(8\pi x)$ (which is highly oscillatory), $f^2(x) = (\phi(x, 0.2, 0.25) + \phi(x, 0.6, 0.2))/4$, with $\phi(x, a, b)$ being a normal density of mean a and standard deviation b, (which requires a locally adaptive estimator as there are different degrees of smoothness on the left and right of the function), $f^3(x) = 1.5x$ (which was chosen as many relationships are often thought to be linear) and $f^4(x) = \cos(2\pi x)$, (which is a smooth nonlinear function). The independent variables for the four univariate regressions were $x^1 \sim U(0,1)$, $x^2 \sim U(0,1)$ and

$$\begin{pmatrix} x^3 \\ x^4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, 0.3 \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}\right)$$

We generated $n = 100$ data points from this true SUR model and applied the nonparametric SUR estimator to this data. To assess the resulting estimates of the four functions, we calculated the log mean squared difference between the function estimates and the true

functions. This measure of distance between the two is defined as

$$\log(MSD_i) = \log\left(\frac{1}{200}\sum_{k=1}^{200}\left(\hat{f}^i(z_k) - f^i(z_k)\right)^2\right)$$

where $\min(x^i) = z_1 < z_2 < \ldots < z_{200} = \max(x^i)$ is an evenly spaced grid over the domain of x^i. For the same data we also fit four single equation univariate nonparametric estimators corresponding to using the estimator proposed here with $m = 1$ on each of the four regressions. The log mean squared difference was also calculated for each of these four function estimates. We use the same bases for the SUR and single equation estimators, namely the reproducing kernel basis $\mathcal{B}_3$.

—Figure 3 about here.—

The entire process was repeated one hundred times. Figures 3(a)-(d) give boxplots of the one hundred resulting values of $\log(MSD_i)$ for each of the four functions ($i = 1, 2, 3, 4$) and for both the nonparametric SUR estimator (NSUR) and individual nonparametric estimators (NR). Figure 3 shows that taking into account the correlation between the errors has substantially and consistently improved the resulting estimates of all the regression functions.

To examine the qualitative improvement that occurs, we focus on the single data set corresponding to the 50th sorted value of $\sum_{i=1}^{4} MSD_i$ for the nonparametric SUR estimator. This data set can be regarded as providing a 'typical' example of the procedure and is plotted as four scatter plots in figures 3(e)-(h) and again in figures 3(i)-(l). The nonparametric SUR estimates of the four functions for this data appear in figures 3(e)-(h) and the estimates for the separate nonparametric regressions appear in figures 3(i)-(l). These figures show that the nonparametric SUR estimator significantly outperforms the separate nonparametric estimators which ignore the correlation between the separate regressions. The variance Σ of the errors and its estimate $\hat{\Sigma}$ for this data set is given below.

$$\Sigma = \begin{bmatrix} 1 & 0.96 & 0.64 & 0.93 \\ & 1 & 0.98 & 0.90 \\ & & 1 & 0.85 \\ & & & 1 \end{bmatrix} \quad \hat{\Sigma} = \begin{bmatrix} 1.571 & 1.269 & 0.944 & 1.261 \\ & 1.244 & 0.982 & 1.139 \\ & & 1.053 & 0.913 \\ & & & 1.199 \end{bmatrix} \tag{5.1}$$

It compares favorably to the 'best possible' estimate $\hat{\Sigma}_{\text{best}}$ that arises from the sample variance of the true errors themselves, which are known because this is a simulated example.

$$\hat{\Sigma}_{\text{best}} = \widehat{\text{var}}\left((e^1, e^2, e^3, e^4)'\right) = \begin{bmatrix} 1.568 & 1.234 & 0.921 & 1.256 \\ & 1.261 & 0.990 & 1.143 \\ & & 1.061 & 0.914 \\ & & & 1.200 \end{bmatrix}$$

5.2 Example 2: Different Bases

The choice of basis used to represent a function can make a large amount of difference in the empirical performance of any estimation methodology. The bases introduced in section 2 are those typically used to fit smooth functions and this example compares their ability to parsimoniously reproduce the function employed in the multivariate nonparametric systems examined in example 1. We applied the same nonparametric SUR estimator as in the first example, except that we used the five bases $\mathcal{B}_1, \ldots, \mathcal{B}_5$. Figure 4 provides the $\log(MSD_i)$ for the five bases and four functions. The performances are roughly comparable, which is because all of the bases are known to be suited to smooth function estimation. However, of the five alternatives, on average the quartic reproducing kernel basis is superior, which is why we focus on this basis throughout the paper.

—Figure 4 About Here.—

5.3 Example 3: Various noise levels and sample sizes

The first example investigated the properties of the procedure in the case where there was a fixed sample size ($n = 100$) and a fixed covariance matrix Σ. Although this particular combination was challenging (because of the high ratio of the standard deviation of the errors to the function ranges in each of the regressions), it important to see how the estimator performs with different sample sizes and noise levels.

To undertake this, we repeated the simulation experiment discussed in the first example, but where we considered all combinations of four sample sizes, $n = 100, 200, 400, 1600$, and four covariance matrices $0.25\Sigma, 0.5\Sigma, \Sigma, 2\Sigma$, where Σ is the same covariance matrix used in the first example and is given at (5.1). Notice that these are still highly correlated examples, just with different noise levels. To compare the nonparametric SUR estimator (NSUR) to the four separate nonparametric regressions (NR) we calculated the logarithm of the mean squared difference averaged over all four regression functions for both procedures. That is,

$$\log(AMSD) = \log\left(\frac{1}{4}\sum_{i=1}^{4} MSD_i\right)$$

Low values of this suggest that the average distance of the function estimates from the true functions is low (and therefore the performance of the estimator is good), while higher values suggest the function estimates are further away from the true function. For each combination of sample size and noise level, two boxplots (one each for the NSUR and NR estimators) of the values of $AMSD$ resulting from the 100 simulated data sets are included in figure 5.

—figure 5 about here—

Figure 5 reveals that regardless of sample size and noise level, the NSUR procedure consistently outperforms the NR procedure, where the correlation is ignored and separate regressions fit. The performance of the estimators converge as the sample size increases and the noise level decreases. For example, the performance between the two differ more when $\sigma = 2$ and $n = 100$ than when $\sigma = 0.25$ and $n = 1600$. In moderate sample size environments the benefits can be substantial. For example figure 5 reveals that, regardless of noise level, the NSUR estimator provides about the same level of performance (as measured by $AMSD$) with only $n = 100$ observations as simple NR estimation does with a sample size of between $n = 400$ and $n = 1600$. Although $AMSD$ is a distance measure averaged over the four regression functions, we have checked that the NSUR estimator also outperforms the separate NR estimates using the individual MSD_i criteria on all four individual functions.

To demonstrate that the NSUR estimator is practical to implement, we report the time required to fit models of each sample size for both it and the NR procedure. The computer

used was a standard DEC Alpha workstation running at 233 MHz and the code for both procedures was written in FORTRAN and compiled similarly. Although these timings are implementation dependent, they do indicate that this Markov chain Monte Carlo procedure is not overly computationally intensive.

——table 1 about here.——

5.4 Example 4: Unrelated regressions

The previous examples consider a highly related set of regressions and demonstrate the improvements in the regression function estimates that can occur when correlations between the regressions are modeled and estimated, rather than ignored. However, consider the case where it is uncertain whether, or not, there is correlation between the regressions. In this case, is there a risk of degrading the function estimates by modeling a correlation that does not exist?

To investigate this case, we repeated the simulation undertaken in example 1, except where the true regressions were fixed to be unrelated, with $\Sigma = I_4$. Figures 6(a)-(l) provide the equivalent output for this example as was produced in example 1. It can be seen from the boxplots in figures 6(a)-(d) that, in general, there is a slight deterioration in the $\log(MSD_i)$ for the NSUR estimator compared to the NR estimation procedure. This is expected as the regressions are actually not related and the NSUR procedure also estimates Σ. For the single median data set (which we take as a typical example in the same way as example 1) the estimate of Σ is

$$\hat{\Sigma} = \begin{bmatrix} 1.005 & -0.207 & -0.147 & 0.174 \\ & 0.851 & 0.043 & -0.065 \\ & & 0.829 & 0.079 \\ & & & 1.137 \end{bmatrix}$$

However, the loss in the efficiency of the function estimates is very small and in this median data set the function estimates from the NSUR estimator (figures 6(e)-(h)) are almost identical to those from the NR procedure (figures 6(i)-(l)). This suggests that if it is not known

whether a system of regressions is actually related, or not, using nonparametric SUR estimation can provide significant improvements if there really is correlation, while it is unlikely to result in a serious degradation of the function estimates if the regressions were not really related.

—figure 6 about here—

5.5 Implementation Details

The Markov chain in all these estimations ran with 1000 iterations for the warmup and a subsequent 500 iterations for the mixture estimation. The warmup period is conservative as the sampling scheme consistently appears to converge (as measured by the distributions of the iterates) within fifty iterations. In addition, we are using a conservative number of iterations for the mixture estimation as the estimates $\hat{f}^i$ appear to stabilize after around fifty to one hundred iterations.

Acknowledgements

The authors would like to thank Steve Marron, Gael Martin and Tom Smith for useful comments and Paul Kofman for supplying the Australian mining returns data. Both Michael Smith and Robert Kohn are grateful for the support of Australian Research Council grants.

Appendix 1 Generating from the conditional posterior distributions

A1.1 Generating from $\beta_\gamma|\Sigma^{-1},\gamma,\mathbf{y}$

This conditional distribution can be calculated exactly, as

$$p(\beta_\gamma|\Sigma^{-1},\gamma,\mathbf{y}) \propto p(\mathbf{y}|\beta_\gamma,\Sigma^{-1},\gamma)p(\beta_\gamma|\Sigma^{-1},\gamma)$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{nm+1}{nm}\right)(\beta_\gamma - \hat{\mu}_\gamma)' X_\gamma' A X_\gamma (\beta_\gamma - \hat{\mu}_\gamma)\right\},$$

so that $\beta_\gamma|\Sigma^{-1},\gamma,\mathbf{y} \sim N(\hat{\mu}_\gamma, \frac{nm}{nm+1}(X_\gamma' A X_\gamma)^{-1})$. Here, $\hat{\mu}_\gamma$ and A are defined in section 3.1.

A1.2 Generating from $\Sigma^{-1}|\beta_\gamma,\gamma,\mathbf{y}$

This conditional distribution is difficult to recognize as Σ is embedded in the conditional prior for β_γ. Therefore, to obtain an iterate we use a Metropolis-Hastings step; see Chib and Greenberg (1995b) for an introduction to this tool. The proposal density from which we generate a candidate iterate is given by

$$\begin{aligned} q(\Sigma^{-1}) &\propto p(\mathbf{y}|\beta_\gamma,\Sigma^{-1},\gamma)p(\Sigma^{-1}|\gamma) \\ &\propto |\Sigma^{-1}|^{(n-m-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\Omega\Sigma^{-1})\right\} \end{aligned}$$

which is a Wishart(Ω^{-1},n,m) density. Here, Ω is an $(m \times m)$ matrix with ijth element $\omega_{ij} = (\mathbf{y}^i - X^i_{\gamma^i}\beta^i_{\gamma^i})'(\mathbf{y}^j - X^j_{\gamma^j}\beta^j_{\gamma^j})$. A newly generated iterate Σ^{-1}_{new} is accepted over the old value Σ^{-1}_{old} with probability

$$\alpha = \min\left(\frac{p(\Sigma^{-1}_{\text{new}}|\beta_\gamma,\gamma,\mathbf{y})q(\Sigma^{-1}_{\text{old}})}{p(\Sigma^{-1}_{\text{old}}|\beta_\gamma,\gamma,\mathbf{y})q(\Sigma^{-1}_{\text{new}})},1\right) = \min\left(\frac{p(\beta_\gamma|\Sigma^{-1}_{\text{new}},\gamma)}{p(\beta_\gamma|\Sigma^{-1}_{\text{old}},\gamma)},1\right)$$

High acceptance rates of 60-90% are obtained because the proposal density $q(\cdot)$ is equal to the correct conditional density except for the factor $p(\beta_\gamma|\Sigma^{-1},\gamma)$.

A1.3 Generating from $\gamma^i_j|\Sigma^{-1},\gamma\backslash\gamma^i_j,\mathbf{y}$

This conditional density can be calculated exactly, with

$$\begin{aligned} p(\gamma^i_j|\Sigma^{-1},\gamma\backslash\gamma^i_j,\mathbf{y}) &\propto \int p(\mathbf{y}|\gamma,\Sigma^{-1},\beta_\gamma)p(\beta_\gamma|\gamma,\Sigma^{-1})d\beta_\gamma p(\gamma^i_j) \\ &\propto (nm+1)^{-q_\gamma/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}'A\mathbf{y} - \mathbf{y}'AX_\gamma(X_\gamma'AX_\gamma)^{-1}X_\gamma'A\mathbf{y}\right)\right\} \qquad \text{(A1.1)} \end{aligned}$$

In equation (A1.1) the regression coefficient is integrated out using $\beta_\gamma \sim N(\hat{\mu}_\gamma, \frac{nm}{nm+1}(X_\gamma'AX_\gamma)^{-1})$ and $p(\gamma^i_j) = 1/2$. The binary variable γ^i_j is generated by evaluating (A1.1) for $\gamma^i_j = 1$ and $\gamma^i_j = 0$ and then normalizing.

References

Bartels, R., Fiebig, D. and Plumb, M., (1996), 'Gas or electricity, which is cheaper?: an econometric approach with an application to Australian expendicture data', Dept. of Econometrics Working Paper, University of Sydney.

Chib, S. and Greenberg, E., (1995a), 'Hierarchical Analysis of SUR Models with Extensions to Correlated Serial Errors and Time Varying Parameter Models', *Journal of Econometrics*, 68, 339-360

Chib, S. and Greenberg, E., (1995b), 'Understanding the Metropolis-Hastings algorithm', *The American Statisitician*, 49, 327-335

De Boor, C., (1978), *A practical guide to splines*, New York: Springer-Verlag.

Denison, D., Mallick, B., and Smith, A., (1997), 'Automatic Bayesian Curve Fitting', preprint.

Donoho, D. and Johnstone, I., (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika*, vol. 81, 425-455

Friedman, J., (1991), 'Multivariate adaptive regression splines', *The Annals of Statistics* (with discussion), 19, 1-141

Friedman, J. and Silverman, B., (1989), 'Flexible parsimonious smoothing and additive modelling', *Technometrics*, 31, 3-39

Hanssens, D. and Weitz, B. (1980), 'The effectiveness of Indus trial Print Advertisements Across Product Categories,' *Journal of Marketing Research*, 17, 294-306

O'Hagan, A., (1995), 'Fractional Bayes factors for model comparison' (with discussion), *J. Royal Stat. Soc.*, Ser. B, 57, 99-138.

Holmes, C., and Mallick, B., (1997), 'Bayesian Radial Basis Functions of Unknown Dimension', preprint.

Luo, Z., and Wahba, G., (1997), 'Hybrid Adaptive Splines', *Journal of the American Statistical Association*, vol. 92, 107-116.

Mandy, D., and Martins-Filho, C., (1993), 'Seemingly unrelated regressions under additive heteroskedasticity: Theory and share equation applications', *Journal of Econometrics*, 58, 315-346

Min, C. and Zellner, A., (1993), 'Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates', *Journal of Econometrics*, 56, 89-118

Powell, M., (1987), 'Radial basis functions for multivariate interpolation: a review', in Mason, J. and Cox, M. (eds.), *Algorithms for approximation*

Smith, M., and Kohn, R., (1996), 'Nonparametric regression via Bayesian variable selection,' *Journal of Econometrics*, vol. 75, no. 2, 317-344

Smith, M., and Kohn R., (1997), 'A Bayesian Approach to Nonparametric Bivariate Regression', *Journal of the American Statistical Association*, vol. 92, no. 440, 1522-1535

Srivastava, V. and Giles, D., (1987), *Seemingly Unrelated Regression Equations Models*, New York: Marcel Dekker

Thomas, T., (1995), 'Bringing home the bacon', *Australian Business Review Weekly*, April 24, 68-89

Tierney, L., (1994), 'Markov chains for exploring posterior distributions', *The Annals of Statistics*, 22, 1701-1762

Wahba, G., (1990), *Spline models for observational data*, SIAM: Philadelphia

Wong, F., Hansen, M., Kohn, R. and Smith, M., (1997), 'Focused sampling and its application to nonparametric and robust regression', preprint.

Yee, T. and Wild, C., (1996), 'Vector generalised additive models', *Journal of the Royal Statistical Society*, Series B, 58, 481-493

Zellner, A., (1962), 'An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias,' *Journal of the American Statistical Association*, 57, 500-509

Zellner, A., (1963), 'Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results,' *Journal of the American Statistical Association*, 58, 977-992

Zellner, A., (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley

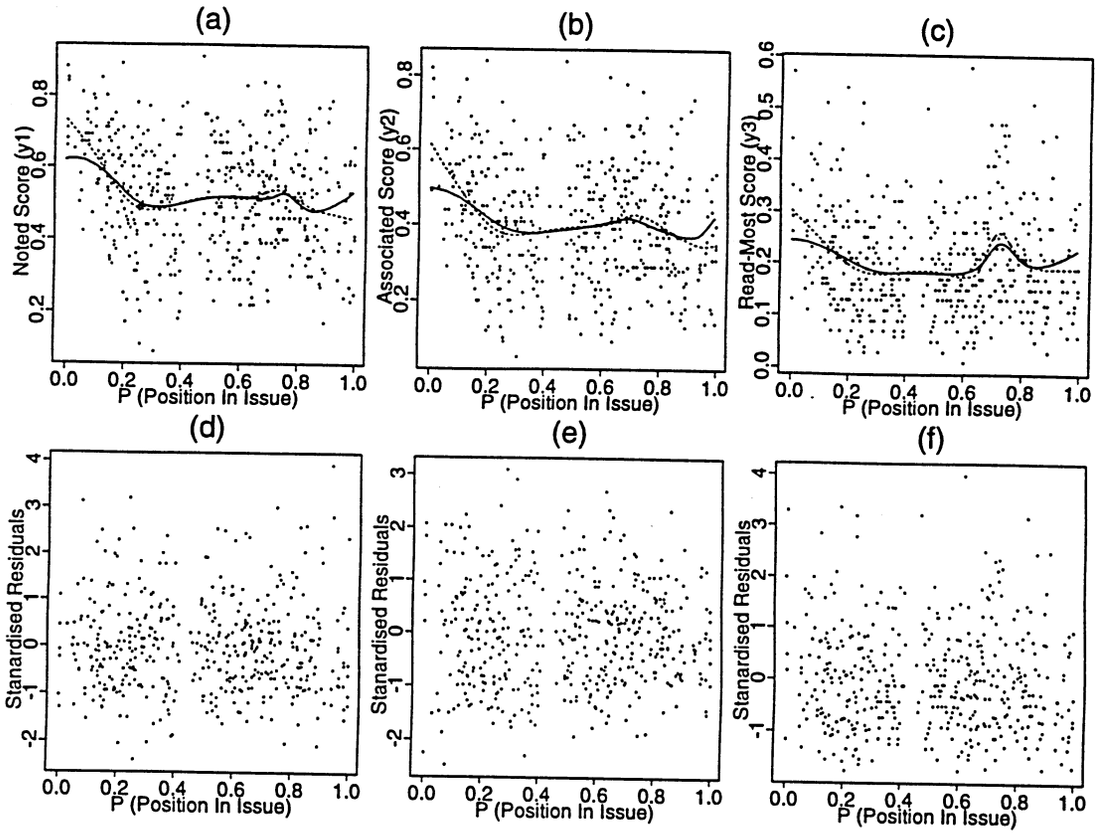

Figure 1: (a)-(c) Scatter plots are of P versus y^i; bold lines are nonparametric SUR estimates of $\hat{f}^i$, while the dashed lines are the single equation estimates. Panels (d)-(f) contain scatter plots of P versus the standardised uncorrelated residuals resulting from the nonparametric SUR fit.

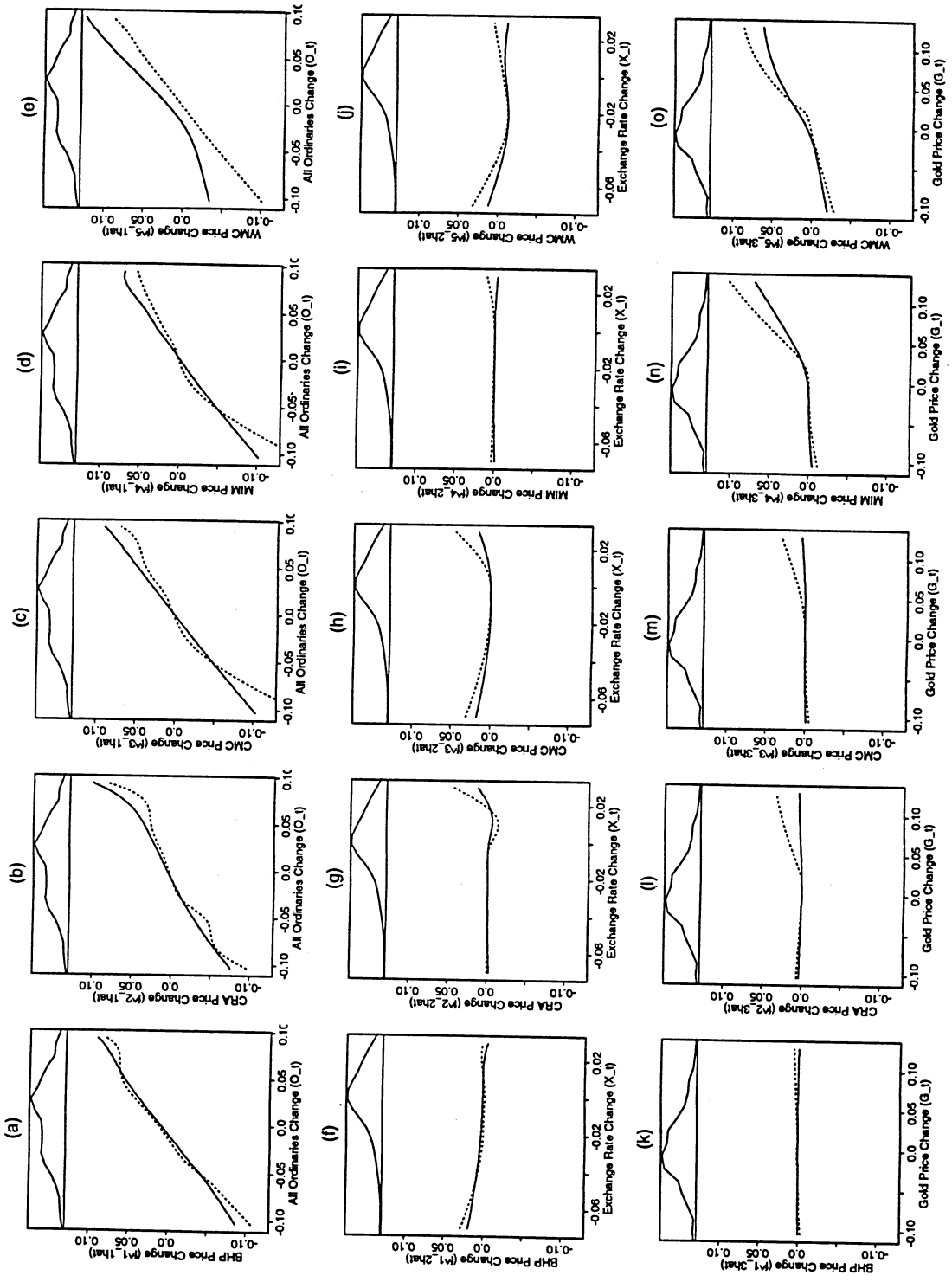

Figure 2: Bold lines are estimates of the regression functions for the NSUR estimator, while dotted lines are from the five seperate regressions. Panels (a)-(e) correspond to function estimates $\hat{f}_1^i$ for $i = 1, \ldots, 5$. Panels (f)-(j) correspond to function estimates $\hat{f}_2^i$ for $i = 1, \ldots, 5$. Panels (k)-(o) correspond to function estimates $\hat{f}_3^i$ for $i = 1, \ldots, 5$.

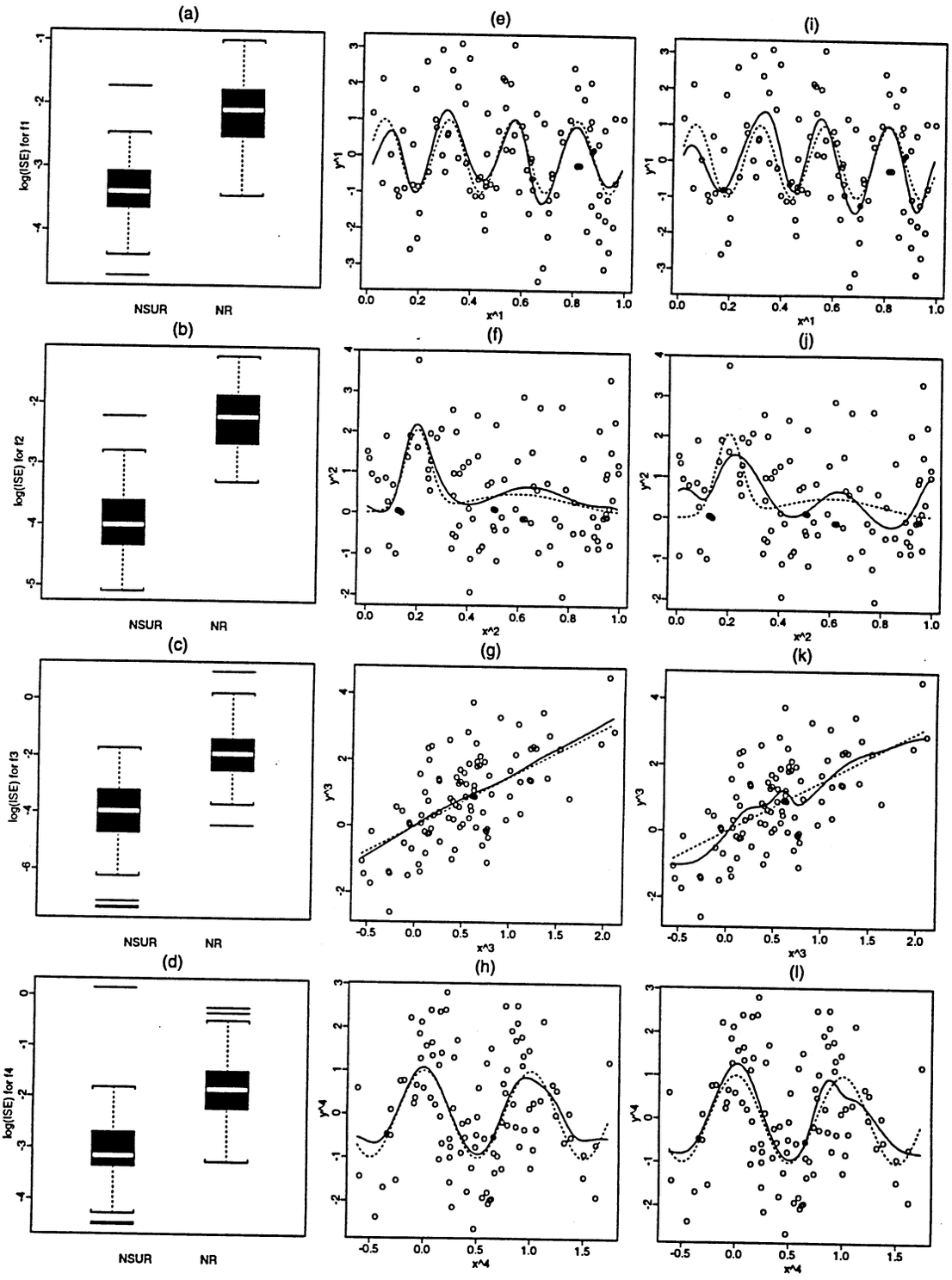

Figure 3: (a)-(d) Boxplots of the $\log(MSD_i)$ for $i = 1, 2, 3$ and 4, respectively. The left hand boxplot is for the NSUR estimator, while the right hand boxplot is for the NR estimation procedure. Panels (e)-(h) contain scatter plots of x^i against y^i, along with the function estimates $\hat{f}^i$ (bold line) and true functions f^i (dotted line) for $i = 1, 2, 3, 4$ that result from applying the NSUR estimator. Panels (i)-(l) plot the function estimates $\hat{f}^i$ (bold line) and true functions f^i (dotted line) for $i = 1, 2, 3, 4$ that result from applying the NR procedure to the same data.

sample size	NSUR estimator	NR procedure
$n = 100$	43s	9s
$n = 200$	58s	14s
$n = 400$	213s	49s
$n = 1600$	3850s	910s

Table 1: Average time (in seconds) taken to complete a fit to data generated from the model in example 2 for both the NSUR and NR estimation procedures and four sample sizes.

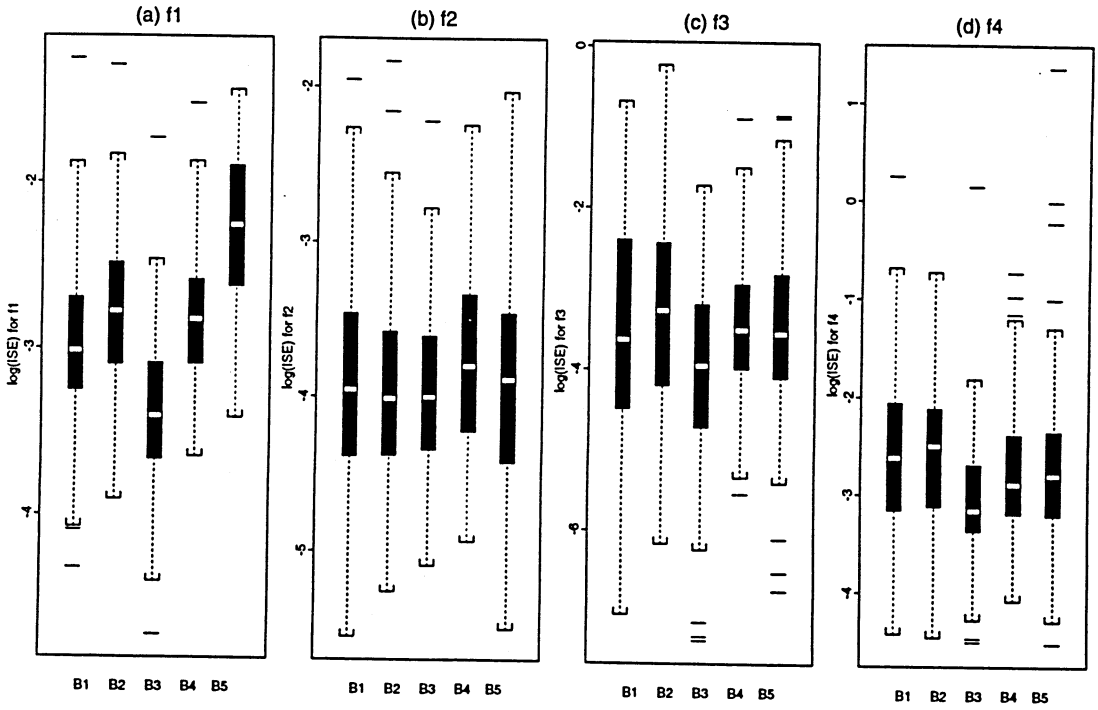

Figure 4: Comparision of the five bases $\mathcal{B}_1, \ldots, \mathcal{B}_5$. Each plot corresponds contains boxplots of the $\log(MSD_i)$ for the four functions/equations.

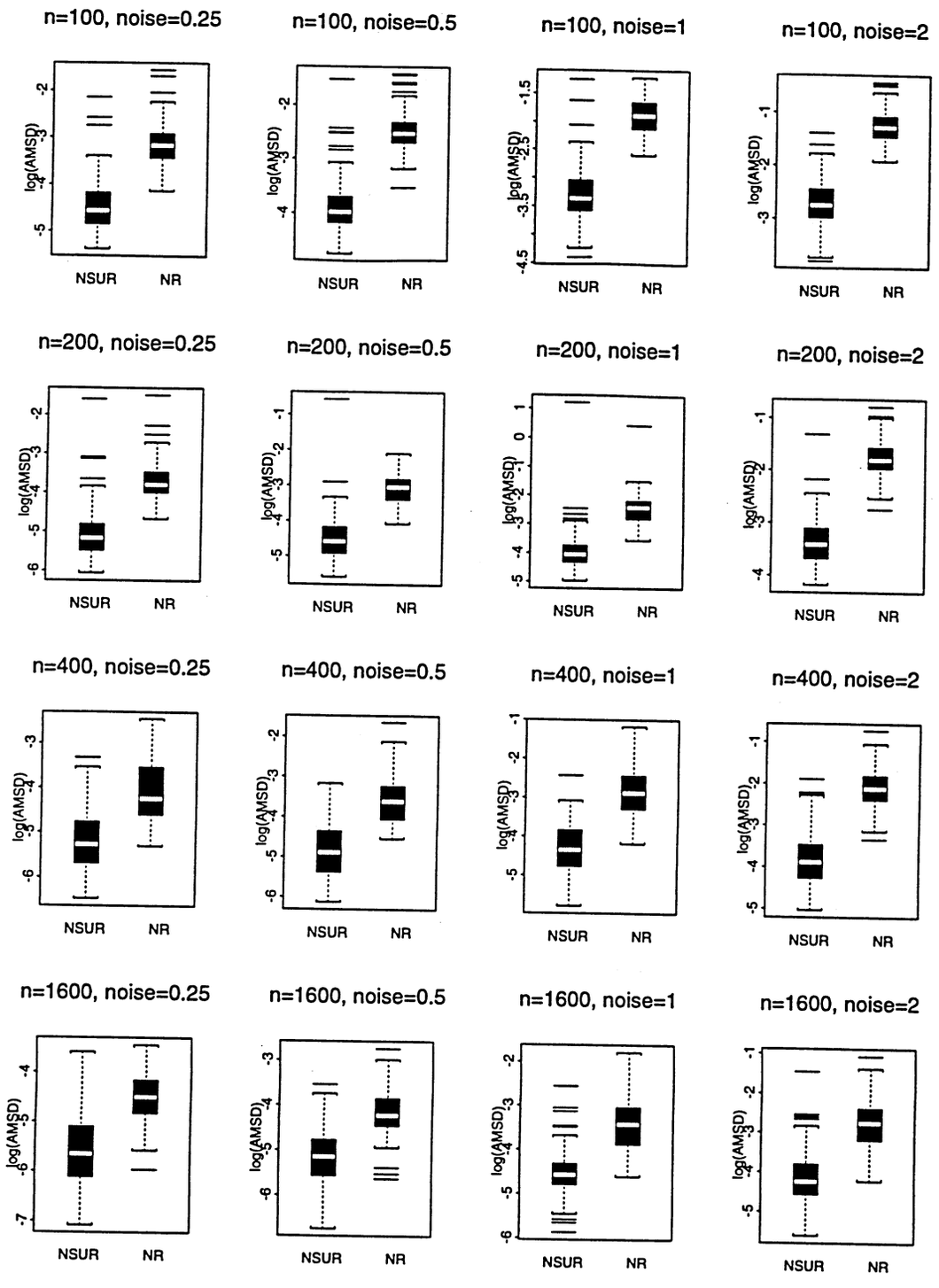

Figure 5: Results of the simulation experiment in example 3, with each row corresponding to a different sample size and column to a different noise level. In each panel, the left boxplot is of the $\log(AMSD)$ resulting from fitting the NSUR estimator to the 100 simulated data sets, while the right boxplot is of the $\log(AMSD)$ resulting from fitting the NR estimator to the same data.

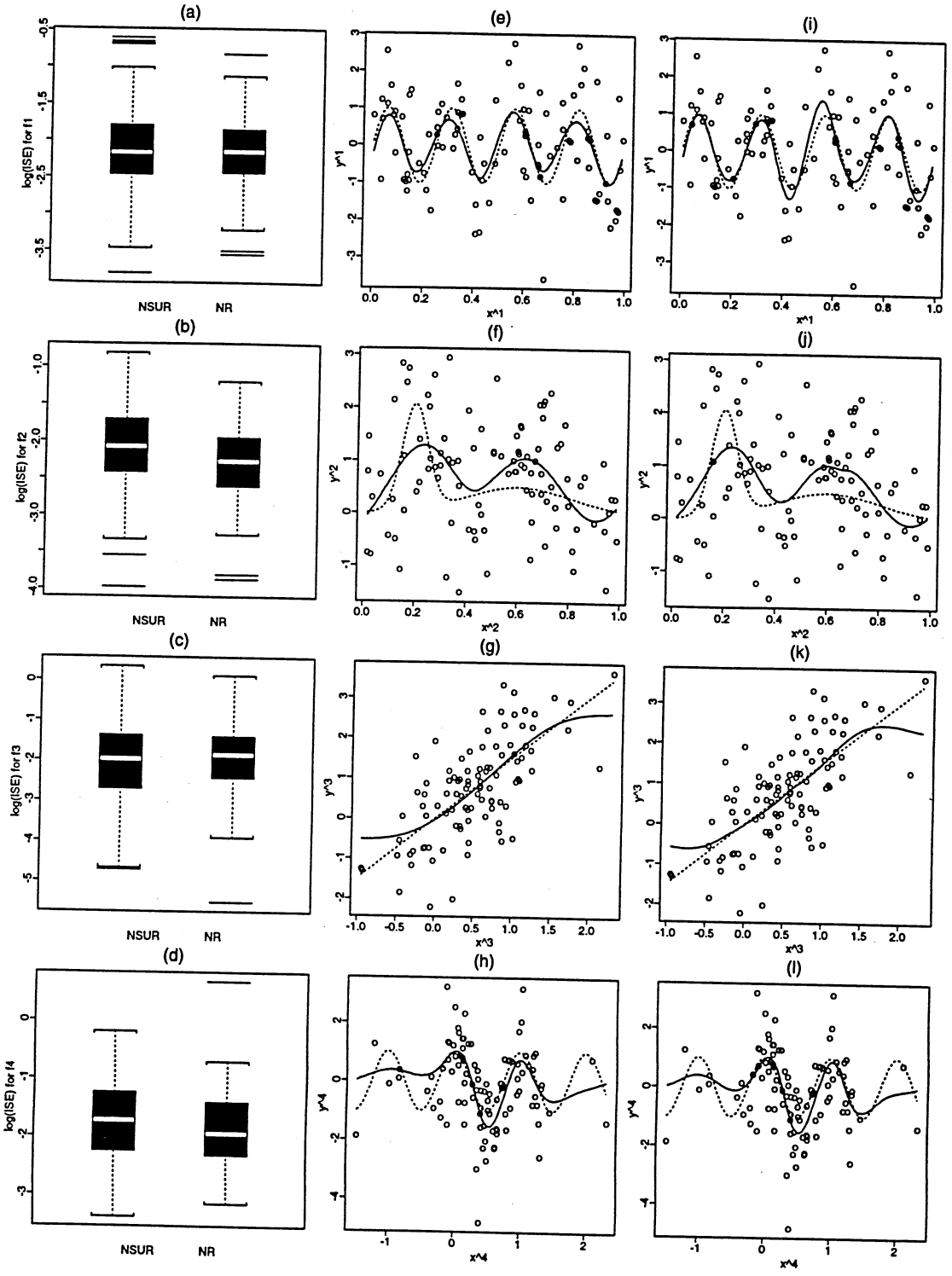

Figure 6: Same as figure 2, but for the data generated in example 4.