Model Selection in Nonparametric and Semiparametric Time Series Regression

Jiti Gao

Department of Statistics
School of Mathematics and Statistics
The University of Western Australia
Crawley WA 6009, Australia
Email: jiti@maths.uwa.edu.au
Http://www.maths.uwa.edu.au/˜jiti/kao45.pdf

Outline:

1. Motivation

2. Model Estimation and Selection

3. Example of Implementation

4. Discussion

1. Motivation

Example 1.0 (Real Data): Fisheries Western Australia (WA) manages commercial fishing in Western Australia. Simple Catch and Effort statistics are often used in regulating the amount of fish that can be caught and the number of boats that are licensed to catch them. The establishment of the relationship between the Catch (in kilograms) and Effort (the number of days the fishing vessels spent at sea) is very important both commerically and ecologically.

The monthly fishing data set from January 1976 to December 1999 is available from the Fisheries WA Catch and Effort Statistics (CAES) database. Existing studies suggest that the relationship between catch and effort is nonlinear while the dependence of the current catch on the past catch appears to be linear. This suggests using a partially linear model of form

$$C_t = \beta_1 C_{t-1} + \cdots + \beta_p C_{t-p} + \phi(E_{t-1}, E_{t-2}, \ldots, E_{t-q}) + \epsilon_t, \ t = r, \cdots, \tag{1.1}$$

where $r = \max(p, q)$, $\{\epsilon_t\}$ is a random error, and $\{C_t\}$ and $\{E_t\}$ represent the catch and the effort at time t.

Selection of subsets for $\{C_{t-i} : 1 \le i \le p\}$ and $\{E_{t-j} : 1 \le j \le q\}$ is important.

Example 1.1 (Nonparametric Regression):

$$Y_t = m(U_t, V_t) + e_t, \tag{1.2}$$

where U_t and V_t can be both time series;

- V_t–exogenous time series;

- $m(\cdot)$–unknown; and
- $E[e_t|U_t, V_t] = 0$ and $0 < E[e_t^2|U_t, V_t] < \infty$.

Model (1.2) includes some special cases:

- $U_t = (Y_{t-1}, \cdots, Y_{t-p})$; and
- $V_t = (V_{t1}, \cdots, V_{tq})^\tau$ is a vector of exogenous variables.

When $p + q \geq 3$, selection of p and q is important.

- Method I: In econometrics, work has been done on *testing for nonparametric significance*:

$$H_0 : \ E[Y_t|U_t, V_t] = E[Y_t|U_t] \text{ almost surely.} \tag{1.3}$$

- Method II: In statistics, concentration has been on selecting an optimum subset, Z_{t_c}, of (U_t, V_t) such that

$$E[Y_t|U_t, V_t] = E[Y_t|Z_{t_c}] \text{ almost surely.} \tag{1.4}$$

  For example, $Z_{t_c} = (U_{t1}, U_{t3}, U_{t5}, V_{t1}, V_{t2})^\tau$ for the case of $p = q = 5$.

- The main difference is that Method II may be able to treat all the lags equally without assuming that $\{V_t\}$ is less significant. As a result, Method II may involve more expensive computation than Method I.

Example 1.2 (Semiparametric Regression):

$$Y_t = U_t^\tau \beta + \phi(V_t) + e_t, \tag{1.5}$$

where U_t and V_t are as defined before;

- β and $\phi(\cdot)$–unknown; and
- $E[e_t|U_t, V_t] = 0$ and $0 < E[e_t^2|U_t, V_t] < \infty$.

In both theory and practice, selection of p and q is important:

- Parametric part: Work has been done on *parametric cross–validation (CV) selection*; and
- Nonparametric part: Work has also been done on *nonparametric CV1 selection.*

Consider model (1.5)

$$Y_t = \sum_{i=1}^{p} U_{ti}\beta_i + \phi(V_{t1}, \cdots, V_{tq}) + e_t. \tag{1.6}$$

Model (1.6) has some important special cases:

- When $U_{ti} = Y_{t-i}$ for $1 \le i \le p$, model (1.6) is called a *partially linear AR(p) model* of the form

$$Y_t = \sum_{i=1}^{p} Y_{t-i}\beta_i + \phi(V_{t1}, \cdots, V_{tq}) + e_t, \tag{1.7}$$

  where $(V_{t1}, \cdots, V_{tq})$ can be an exogenous time series.

- When $V_{tj} = Y_{t-j}$ for $1 \le j \le q$, model (1.6) is called a *partially nonlinear AR(q) model* of the form

$$Y_t = \sum_{i=1}^{p} U_{ti}\beta_i + \phi(Y_{t-1}, \cdots, Y_{t-q}) + e_t, \tag{1.8}$$

  where $(U_{t1}, \cdots, U_{tp})$ can also be an exogenous time series.

Question:

- How to select an optimum subset of $(U_{t1}, \cdots, U_{tp})$; and
- an optimum subset of $(V_{t1}, \cdots, V_{tq})$ *simultaneously* ?

2. Model Estimation and Selection

Consider model (1.5) only. Key notation includes:

- $A_p = \{1, \ldots, p\}$; $\mathcal{A} = \{\text{all nonempty subsets of } A_p\}$;
- $D_q = \{1, \ldots, q\}$; $\mathcal{D} = \{\text{all nonempty subsets of } D_q\}$;
- For $A \in \mathcal{A}$, $U_{tA} = \{U_{ti}, i \in A\}$; $\beta_A = \{\beta_i, i \in A\}$;
- For $D \in \mathcal{D}$, $V_{tD} = \{V_{ti}, i \in D\}$;
- $d_E = |E|$ denotes the cardinality of a set E;
-
$$\mathcal{A} = \{A : A \in \mathcal{A} : \ \beta_A \text{ contains all } \beta_i \neq 0 \text{ of } \beta\},$$
-
$$\mathcal{D} = \{D : D \in \mathcal{D} \ \text{ such that } \ E\left[Y_t|V_{tD}\right] = E[Y_t|V_t]\},$$
-
$$\mathcal{B} = \{(A, D) : A \in \mathcal{A} \ \text{ and } \ D \in \mathcal{D}\}.$$

  When $(A, D) \in \mathcal{B}$,

$$E[Y_t|U_{tA}, X_{tD}] = E[Y_t|U_t, X_t].$$

Key assumptions include

- U_t and V_t are both strictly stationary and α–mixing, but independent each other.

- Let $\mathcal{B}_0 = \{(A_0, D_0) \in \mathcal{B}$, such that $|A_0|+|D_0| = \min_{(A,D)\in\mathcal{B}}[|A|+|D|]\}$. Assume that (A_0, D_0) is the unique element of $\mathcal{B}_0$ and denoted by (A_*, D_*).

- Assume that there is a unique pair (β_*, ϕ_*) such that the true and compact version of model (1.5) is defined by

$$Y_t = U_{tA_*}^\tau \beta_* + \phi_*(V_{tD_*}) + e_t. \tag{2.1}$$

- Define $\theta_j(X_{tj}) = E[\phi_*(V_{tD_*})|V_{tj}]$ for $j \in D_q - D_*$. There exists an absolute constant $M_0 > 0$ such that

$$\min_{j\in D_q - D_*} \min_{\alpha,\beta} E\left[\theta_j(V_{tj}) - \alpha - \beta V_{tj}\right]^2 \geq M_0.$$

  This is imposed to exclude the case where $\phi(\cdot)$ is also a linear function of a subset of $\{V_{tj} : 1 \leq j \leq q\}$.

- $\eta_t(A) = U_{tA} - E\left[U_{tA}\right]$, $\eta(A) = (\eta_1(A), \ldots, \eta_T(A))^\tau$, $\eta_t = U_t - E[U_t]$, $\eta = (\eta_1, \ldots, \eta_T)^\tau$, $Q(A) = \eta(A)\left(\eta(A)^\tau\eta(A)\right)^+\eta(A)^\tau$, and

$$P_T(A) = \frac{1}{T}(\eta\beta)^\tau\left[I_T - Q(A)\right](\eta\beta).$$

  Assume that for each given $A \in \mathcal{A}$,

$$\liminf_{T\to\infty} P_T(A) > 0 \text{ in probability.}$$

As U_t and V_t are independent, there are $2^p + 2^q - 2$ possible pairs for (A, D). The selection of (A, D) is then carried out by using the data $\{(Y_t, U_t, V_t) : t = 1, 2, \ldots, T\}$ satisfying

$$Y_t = U_t^\tau\beta + \phi(V_t) + e_t. \tag{2.2}$$

We split the data set into two parts:

- $\{(Y_t, U_t, V_t) : t \in S\}$ and $\{(Y_t, U_t, V_t) : t \in S^c\}$,
- where S is a subset of $\{1, 2, \ldots, T\}$ containing T_v integers;
- S^c is its complement containing T_c integers, $T_v + T_c = T$.
- (2.2) is fitted using the construction data $\{(Y_t, U_t, V_t) : t \in S^c\}$;
- and the prediction error is assessed using the validation data $\{(Y_t, U_t, V_t) : t \in S\}$, treated as if they were future values.

Using the construction data we can estimate β_A by

$$\widehat{\beta}_c(A, D) = \left(U_c(A)^\tau U_c(A)\right)^+ U_c(A)^\tau Y_c(D), \tag{2.3}$$

where

$$U_c(A) = (U_{i_1,c}(A), \ldots, U_{i_{T_c},c}(A))^\tau, \quad Y_c(D) = (Y_{i_1,c}(D), \ldots, Y_{i_{T_c},c}(D))^\tau, \tag{2.4}$$

in which for $t \in S^c$ or $t \in S$,

$$\begin{aligned} U_{t,c}(A) &= U_{tA} - \hat{\phi}^c_{2t}(A), \quad \hat{\phi}^c_{2t}(A) = \frac{1}{T_c}\sum_{s\in S^c} U_{sA}, \\ Y_{t,c}(D) &= Y_t - \hat{\phi}^c_{1t}(D), \quad \hat{\phi}^c_{1t}(D) = \sum_{s\in S^c} W_D(t,s)Y_s, \end{aligned}$$

with

$$W_D(t,s) = \frac{K_D((V_{tD} - V_{sD})/h)}{\sum_{l=1}^{T} K_D((V_{tD} - V_{lD})/h)},$$

in which T is the number of observations, K_D is a multivariate kernel function, and h is a bandwidth parameter satisfying

$$h \in H_{TD} = \left[a_D T^{-\frac{1}{4+|D|}-c_D}, b_D T^{-\frac{1}{4+|D|}+c_D} \right],$$

in which the constants a_D, b_D and c_D satisfy $0 < a_D < b_D < \infty$ and $0 < c_D < \frac{1}{2(4+|D|)}$.

For $t \in S$, let $\widehat{Y}^c_t(A,D) = U_{t,c}(A,D)^\tau \widehat{\beta}_c(A,D)$. The average squared prediction error is then defined by

$$\mathrm{CV}(A,D;h) = \frac{1}{T_v}\sum_{t\in S}\left(Y_{t,c}(D) - \widehat{Y}^c_t(A,D)\right)^2. \tag{2.5}$$

The $\mathrm{CV}(A,D;h)$ function is called the semiparametric leave–T_v–out CV function, abbreviated as semiparametric CVT_v function. Randomly draw a collection $\mathcal{R}$ of n subsets of $\{1,2,\ldots,T\}$ that have size T_v and select a model by minimizing

$$\mathrm{MCCV}(A,D;h) = \frac{1}{n}\sum_{S\in\mathcal{R}} \mathrm{CV}(A,D;h) = \frac{1}{nT_v}\sum_{S\in\mathcal{R}}\sum_{t\in S}\left(Y_{t,c}(D) - \widehat{Y}^c_t(A,D)\right)^2. \tag{2.6}$$

This is called the semiparametric Monte Carlo CV(T_v) function, abbreviated as semiparametric MCCV(T_v) function.

Since the semiparametric leave–one–out CV function (i.e. $T_v = 1$) is generally inconsistent in the selection of parametric subsets, we suggest using the semiparametric leave–T_v–out CV function in this paper.

Let

$$(\widehat{A}, \widehat{D}, \widehat{h}) = \arg \min_{\{A\in\mathcal{A},\ D\in\mathcal{D},\ h\in H^c_{TD}\}} \mathrm{MCCV}(A,D;h),$$

where $H^c_{TD} = \left[a_D T_c^{-\frac{1}{4+|D|}-c_D}, b_D T_c^{-\frac{1}{4+|D|}+c_D} \right]$, in which the constants a_D, b_D and c_D satisfy $0 < a_D < b_D < \infty$ and $0 < c_D < \frac{1}{2(4+|D|)}$.

We now state the following main result of this paper.

THEOREM 2.1. *Under suitable assumptions, including*

$$\lim_{T\to\infty} \frac{T_v}{T} = 1,$$

$$\lim_{T\to\infty} \frac{T_c}{T} = 0,$$

$$\lim_{T\to\infty} \frac{T^2}{T_c^2 n} = 0.$$

(i)

$$\lim_{T\to\infty} P\left(\widehat{A} = A_*,\ \widehat{D} = D_*\right) = 1. \tag{2.7}$$

(ii) *As* $T \to \infty$,

$$\frac{\widehat{h}}{h_*} \to_p 1, \tag{2.8}$$

where $h_* = c_* T_c^{-\frac{1}{4+|D_*|}}$ *and* c_* *is a positive constant.*

3. Example of Implementation

EXAMPLE 3.1. Consider a nonlinear time series model of the form

$$\begin{aligned} Y_t &= 0.47U_{t-1} - 0.45U_{t-2} + \frac{0.5V_{t-1} - 0.23V_{t-2}}{1 + V_{t-1}^2 + V_{t-2}^2} + e_t, \\ U_t &= 0.55U_{t-1} - 0.12U_{t-2} + \delta_t, \\ V_t &= 0.3\sin(2\pi V_{t-1}) + 0.2\cos(2\pi V_{t-2}) + \epsilon_t,\ t = 3, \ldots, T, \end{aligned} \tag{3.1}$$

in which

- δ_t, ϵ_t and e_t are i.i.d. over $\mathrm{U}(-1,1)$, $\mathrm{U}\left(-\frac{1}{2},\frac{1}{2}\right)$ and $N(0,1)$, respectively;
- U_1, U_2, V_1, V_2 are i.i.d. uniform distribution $\mathrm{U}(-1,1)$;
- U_s and V_t are mutually independent for all $s, t \geq 3$; and
- $\{(\eta_t, \epsilon_t, e_t) : t \geq 3\}$ are independent of both (U_1, U_2) and (V_1, V_2).

In this example, we consider the case where

- V_t, V_{t-1} and V_{t-2} as the candidates of nonparametric regressors; and
- U_{t-1} and U_{t-2} as the candidates of parametric regressors.

In the detailed implementation, we choose

- $n = T = 72$, 152, or 302;
- $h \in H_{TD}^c = \left[0.1 \cdot T_c^{-\frac{2}{9}}, 3 \cdot T_c^{-\frac{1}{9}}\right]$ with $T_c = [T^{3/4}]$; and
- $K(u_1, \ldots, u_j) = \prod_{i=1}^{j} k(u_i)$ for $1 \leq j \leq 3$, where $k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$.
- 1000 replications used in producing the following probabilities.

As U_t and V_t are assumed to be independent, there are only $2^2 + 2^3 - 2 = 10$ possible models.

Table 3.1. The semiparametric MCCV(T_v) function based probabilities

Significant semiparametric	Probability		
regressor	$T = 72$	$T = 152$	$T = 302$
$\{U_{t-1}, U_{t-2}, V_{t-1}, V_{t-2}\}$	0.723	0.891	0.984
$\{U_{t-1}, U_{t-2}, V_t, V_{t-1}, V_{t-2}\}$	0.124	0.064	0.011
$\{U_{t-1}, U_{t-2}, V_{t-1}, V_t\}$	0.065	0.018	0.003
$\{U_{t-1}, U_{t-2}, V_{t-2}, V_t\}$	0.063	0.018	0.002
$\{U_{t-1}, V_{t-1}, V_{t-2}\}$	0.012	0.005	0.000
$\{U_{t-2}, V_{t-1}, V_{t-2}\}$	0.013	0.004	0.000

Remark 3.1: Table 3.1 shows that

- the MCCV(T_v) function can be implemented in practice; and
- both $\widehat{A}$ and $\widehat{D}$ are reasonably good estimators of A_* and D_* even when the sample size T is as modest as 72.
- Comparison between the semiparametric MCCV(T_v) function with the conventional nonparametric CV1 selection function.
- For the same example, consider the case where U_{t-1}, U_{t-2}, V_t, V_{t-1} and V_{t-2} are selected as the candidates of nonparametric regressors.
- There are $2^5 - 1 = 31$ possible nonparametric regressors, since we treat each parametric component as a nonparametric regressor. In addition, the independence between U_t and V_t has not been used.
- To ensure that the empirical comparison between the semiparametric MCCV(T_v) model selection criterion and the conventional nonparametric CV1 model selection criterion can be done in a reasonable way, we choose the same functions as used before.

Table 3.2. The nonparametric CV1 based probabilities

Significant nonparametric	Probability		
regressor	$T = 72$	$T = 152$	$T = 302$
$\{U_{t-2}, V_{t-1}, V_{t-2}\}$	0.473	0.472	0.477
$\{U_{t-1}, V_{t-1}, V_{t-2}\}$	0.464	0.470	0.476
$\{U_{t-1}, U_{t-2}, V_{t-1}, V_{t-2}\}$	0.016	0.018	0.016
$\{U_{t-1}, U_{t-2}, V_t, V_{t-2}\}$	0.017	0.017	0.016
$\{U_{t-1}, U_{t-2}, V_t, V_{t-1}\}$	0.015	0.019	0.014
$\{V_{t-1}, V_{t-2}\}$	0.013	0.004	0.001
$\{V_t, V_{t-1}, V_{t-2}\}$	0.002	0.000	0.000

Remark 3.2: Table 3.2 shows that

- unlike the semiparametric MCCV(T_v) function, the conventional nonparametric CV1 function can not identify the true set of regressors $(U_{t-1}, U_{t-2}, V_{t-1}, V_{t-2}\}$.

- This is a reflection of the fact that the semiparametric MCCV(T_v) selection takes into account the existence of both the the parametric and nonparametric regressors while the nonparametric CV1 neglects the existence of the parametric component but treats each parametric regressor as a nonparametric regressor.

Conclusion

- This talk only concentrates on the case where U_t and V_t are independent.
- Under the independence, we need only to consider selecting $2^p + 2^q - 2$ candidates.
- The methodology applies to the case where U_t and V_t are not necessarily independent.
- For this case, there are possible

$$(2^p - 1) \times (2^q - 1) = 2^{p+q} - 1 - (2^p + 2^q - 2)$$

  candidates when using the semiparametric selection function.
- When using the conventional nonparametric CV1 function, one needs to consider $(2^{p+q} - 1)$ candidates.

Appendix

LEMMA A.1. If $A \in \mathcal{A}$ and $D \in \mathcal{D}$, then there exists $R_T \geq 0$ such that

$$\text{MCCV}(A, D; h) = \frac{1}{T_v n} \sum_{S \in \mathcal{R}} \sum_{t \in S} e_t^2 + \frac{d_A}{T_c} \sigma_0^2 + N_T(D, h) + R_T + o_p(1), \tag{A.1}$$

where R_T is independent of (A, D), $\sigma_0^2 = E[e_t^2]$, and

$$N_T(D, h) = c_1(D) \frac{1}{T_c h^{|D|}} + c_2(D) h^4 + o_p\left(\frac{1}{T_c h^{|D|}}\right) + o_p(h^4) \tag{A.2}$$

for $D \in \mathcal{D}$ and $h \in H_{TD}^c$.

PROOF OF THEOREM 2.1. It can be shown that for each given D, there exists $\bar{h}_D = c_D T_c^{-\frac{1}{4+|D|}}$ such that

$$N_T(D) = N_T(D, \bar{h}_D) = \min_{h \in H_{TD}^c} N_T(D, h) = C_D T_c^{-\frac{4}{4+|D|}} + o_p\left(T_c^{-\frac{4}{4+|D|}}\right), \tag{A.3}$$

in which c_D and C_D are positive constants possibly depending on D.

Let $\text{MCCV}(A, D) = \min_{h \in H_{TD}^c} \text{MCCV}(A, D; h)$. In view of the structure of (A.1) and (A.2), it is known that for each given (A, D),

$$\begin{aligned} \text{MCCV}(A, D) \quad &- \quad \text{MCCV}(A_*, D_*) = \frac{(d_A - d_{A_*})}{T_c} \sigma_0^2 \\ &+ \quad N_T(D) - N_T(D_*) + o_p\left(\frac{1}{T_c}\right). \end{aligned}$$

This implies

$$\lim_{T \to \infty} P(\widehat{A} = A_*, \ \widehat{D} = D_*) = 1.$$

Let $\widehat{h} = \bar{h}_{\widehat{D}}$, $c_* = c_{D_*}$ and $h_* = \bar{h}_{D_*} = c_* T_c^{-\frac{1}{4+D_*}}$. Then the proof of $\frac{\widehat{h}}{h_*} \to_p 1$ follows immediately.