<u>Recursive Least Squares Algorithm.</u> 4/23/00.

Scalar ARMAX Model.

$$Y_k + a_1 Y_{k-1} + \dots + a_n Y_{k-n} = b_1 U_{k-1} + \dots + b_m U_{k-m} + e_k$$

$$K = 0, 1, 2, \dots$$

$$Y_k = 0, \quad U_k = 0, \quad K < 0.$$

$e_0, e_1, \dots$ zero-mean independent.

$$\theta = [a_1 \dots a_n ; b_1 \dots ; b_m].$$

Let $\hat{\theta}_N$: estimate based on Y^N, u^{N-1}, obtained

by minimizing

$$V_N(\theta) = \frac{1}{N} \sum_{K=1}^{N} \varepsilon_K^2(\theta)$$

where $\varepsilon_K(\theta)$ are the prediction errors.

$$\varepsilon_k(\theta) = Y_k + a_1 Y_{k-1} + \dots + a_n Y_{k-n} - b_1 U_{k-1} \dots - b_m U_{k-m}$$

$$\hat{\theta}_N = \left(X_N^T X_N \right)^{-1} X_N^T Y_N$$

$$X_N = col\left[x_1^T, \dots, x_N^T \right] \quad Y_N = [y_1 \dots y_N]^T$$

$$x_K = \left[-Y_{k-1}, -Y_{k-2} \dots, -Y_{k-n}, U_{k-1} \dots U_{k-m} \right]^T$$

At time $N+1$, y_{N+1} and U_N (and hence X_N)

becomes known:

2.

We get
$$\hat{\theta}_{N+1} = \left(X_{N+1}^T X_{N+1}\right)^{-1} X_{N+1}^T Y_{N+1}.$$

let $P_N = \left(X_N X_N^T\right)^{-1}.$

We can compute
$$\hat{\theta}_{N+1} = F_n\left(\hat{\theta}_N, Y_{N+1}, u_N, P_N\right)$$

Matrix Inversion Lemma.

Let $A = n \times n$ $b, c : n \times 1$ and assume that A and $A + bc^T$ are non-singular and $1 + c^T A^{-1} b \neq 0$. Then

$$\left(A + bc^T\right)^{-1} = A^{-1} - \left(1 + c^T A^{-1} b\right)^{-1} A^{-1} bc^T A^{-1}.$$

Proof: Check
$$\left(A^{-1} - \left(1 + c^T A^{-1} b\right)^{-1} A^{-1} bc^T A^{-1}\right)\left(A + bc^T\right) = I.$$

$$= I + A^{-1} bc^T - \left(1 + c^T A^{-1} b\right)^{-1} A^{-1} bc^T - \left(1 + c^T A^{-1} b\right)^{-1} A^{-1} bc^T A^{-1} bc^T$$

$$= I + A^{-1} bc^T - \left(1 + c^T A^{-1} b\right)^{-1}\left[\text{~~~~~~~} 1 + A^{-1} bc^T\right] A^{-1} bc^T$$

$$= I.$$

If we do identify $A = X_N X_N^T$ and $b = c = X_{N+1}$,

we get $P_{N+1} = \left(X_{N+1}^T X_{N+1}\right)^{-1} = \left(X_N^T X_N + x_{N+1} x_{N+1}^T\right)^{-1}$

$$= P_N - \left(1 + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1} x_{N+1}^T P_N$$

$$= \left[I - \left(1 + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1} x_{N+1}^T\right] P_N.$$

Recursion for $\hat{\theta}_N$.

$$\hat{\theta}_{N+1} = P_{N+1} X_{N+1}^T y_{N+1}.$$

$$= P_{N+1} \left(X_N^T y_N + x_{N+1} y_{N+1}\right).$$

$$= \left[I - \left(1 + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1} x_{N+1}^T\right] P_N \left(X_N^T y_N + x_{N+1} y_{N+1}\right)$$

$$= \cancel{P_N \left(X_N^T y_N + x_{N+1} y_{N+1}\right)} + \text{Algebra}.$$

$$= \cancel{P_N X_N^T y_N}$$

$$\hat{\theta}_{N+1} = \hat{\theta}_N + \left(1 + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1} \cdot \left[y_{N+1} - x_{N+1}^T \hat{\theta}_N\right]$$

$$= \hat{\theta}_N + K_{N+1} \, \varepsilon_{N+1}\left(\hat{\theta}_N\right) \longleftarrow \text{Prediction Error}$$

Recursive Generalized Least Squares Algorithm.

$$A(q) Y_k = B(q) U_k + \xi_k. \qquad k = 0, 1, 2 \cdots$$

$$C(q) \xi_k = e_k \qquad \text{zero-initial conditions}$$

$$(e_k : \text{sequence of independent zero-mean r.v's}).$$

$$A(q) = 1 + a_1 q + \cdots + a_n q^n.$$

$$B(q) = b_1 q + \cdots + b_n q^n.$$

$$C(q) = 1 + c_1 q + \cdots + c_n q^n.$$

Define

$$\psi = (a_1 \cdots a_n ; b_1 \cdots b_n)^T \qquad \gamma = (c_1, \cdots, c_n)^T.$$

<u>Predictor.</u> $\varepsilon_k(\psi, \gamma) = A(q) \overline{Y_k} - B(q) \overline{U_k} \qquad k = 1, 2 \cdots$

$$\overline{Y_k} = C(q) Y_k \quad \text{and} \quad \overline{U_k} = C(q) U_k$$

If γ or ψ we get a least-squares problem.

Algorithm:

$\hat{\psi}_0, \hat{\gamma}_0$, P_0, Q_0 pos. definite given

$$\hat{\psi}_{N+1} = \hat{\psi}_N + K_{N+1}\left(y_{N+1}(\hat{\gamma}_N) - x_{N+1}^T \hat{\psi}_N\right).$$

$$K_{N+1} = \left(I + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1}$$

$$P_{N+1} = \left[I - \left(I + x_{N+1}^T P_N x_{N+1}\right)^{-1} P_N x_{N+1} x_{N+1}^T\right] P_N$$

where $y_k(\hat{\gamma}_N) = C_N(q) y_k$ $K = N-n+1, \ldots N+1$.

$$u_k(\hat{\gamma}_N) = C_N(q) u_k$$

$$x_{N+1} = \left[-y_N(\hat{\gamma}_N) \cdots -y_{N-n+1}(\hat{\gamma}_N), u_N(\hat{\gamma}_N) \cdots u_{N-n+1}(\hat{\gamma}_N)\right]^T$$

$C_N(q) = $ polynomial with coefficient entries of $\hat{\gamma}_N$

$$\hat{\gamma}_{N+1} = \hat{\gamma}_N + L_{N+1}\left(\eta_{N+1} \hat{\psi}_{N+1}\right) - \xi_{N+1}^T \hat{\gamma}_N$$

$$L_{N+1} = \left(I + \xi_{N+1}^T Q_N \xi_{N+1}\right)^{-1} Q_N \xi_{N+1}.$$

$$Q_{N+1} = \left[I - \left(I + \xi_{N+1}^T Q_N \xi_{N+1}\right)^{-1} Q_N \xi_{N+1} \xi_{N+1}^T\right] Q_N$$

$$\eta_k(\hat{\psi}_{N+1}) = A_{N+1}(q) Y_k - B_{N+1}(q) u_k$$

$$k = N-n+1, \cdots, N+1.$$

$$\xi_{N+1} = \left[-\xi_N(\hat{\psi}_{N+1}), \cdots, -\xi_{N-n+1}(\hat{\psi}_{N+1}) \right]^T.$$

and $A_{N+1}(q)$ and $B_{N+1}(q)$ are polynomials
with coefficients $\hat{\psi}_{N+1}$.

<u>Remark:</u> Suppose ψ_N and γ_N based on data
$\left(Y^N, u^{N-1} \right)$ are available. One iteration
of generalized least squares - algorithm applied
with ψ_N and γ_N as initial values, gives estimates
$\overline{\psi}, \overline{\gamma}$ which minimize.

$$\psi \longmapsto \underbrace{J_{N+1}(\psi, \gamma_N)}_{\text{Least-squares}} \quad \text{and} \quad \gamma \longmapsto J_{N+1}(\overline{\psi}, \gamma).$$

We have to solve two Ricatti eqns
These equations determined at each step by the
most recent estimates $\gamma_N, \overline{\psi}$ of the parameters
γ and ψ must be solved over the interval $[1, N+1]$.

Take as of estimates ψ_{N+1}, r_{N+1} estimates of $\overline{\psi}, \overline{r}$ calculated from solutions to two approximating Riccati equations, determined at each time step j not by the most recent parameter estimates r_N, ψ_{N+1} but by the estimates r_j, ψ_{j+1} available at time j. Hence to calculate ψ_{N+1}, r_{N+1} we need advance solns. to approximating Riccati eqns. by only 1-step. since solutions at time N available from calculation of ψ_N and r_N.

Order Determination.

Scalar ARMAX.

$$Y_K = b_1 U_{k-1} + \cdots + b_q U_{k-q} + e_K \quad , \quad k = 1, \cdots p.$$

Initial data: $U_0 \cdots \quad U_{-q+1}$

$$e_K \simeq N(0, \sigma^2)$$

$$\text{Let } Y = \begin{pmatrix} Y_1 & \cdots & Y_p \end{pmatrix}^T$$

$$\theta^* = \begin{pmatrix} b_1 & \cdots & b_q \end{pmatrix}^T$$

$$e = \begin{pmatrix} e_1 & \cdots & e_p \end{pmatrix}^T$$

$$Y = X\theta^* + e$$

$$X = \begin{pmatrix} u_0 & \cdots & u_{1-q} \\ u_1 & \cdots & u_{2-q} \\ \vdots & & \\ u_{p-1} & \cdots & u_{p-q} \end{pmatrix} \qquad \underline{p \times q \text{ matrix}} \quad p > q.$$

$$\text{lin. indep. columns}$$

$q =$ Model order.

Let $\quad 0 \leq d \leq q$

$$\theta^* = \begin{pmatrix} \theta_1^* \, , & \theta_{q-d}^* & \theta_{q-d+1}^* \cdots & \theta_{q-1}^* & \theta_q^* \end{pmatrix}$$

Problem:

Determine on the basis of observations when the Hypothesis $\theta_q^* = \theta_{q-1}^* = \cdots = \theta_{q-d+1}^* = 0$

should be rejected.

χ^2 and F distributions.

$k_1, k_2 > 0$. A r.v. X is said to have a $\chi^2(k_1)$ distribution if.

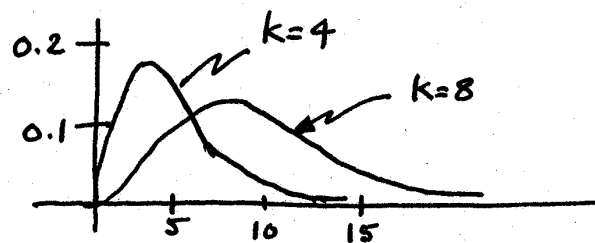$$X = X_1 + X_2 + \cdots + X_{k_1} \quad \text{each } X_i \sim N(0,1)$$

A r.v. X is said to have $F(k_1, k_2)$ distribution if it can be written as

$$X = \frac{Y_1/k_1}{Y_2/k_2} \quad \text{where } Y_1 \text{ and } Y_2 \text{ are}$$

indep. with $\chi_\theta^2(k_1)$ and $\chi^2(k_2)$ distributions.

$\chi^2(0) =$ distribution of r.v (degenerate) which is zero a.s.

For large k_2 if X has $F(k_1, k_2)$ distribution.
then $k_1 X$ has $\chi^2(k_1)$ distribution approximately.

Let $\hat{\theta}$ = l.s. estimate of θ

$\hat{\hat{\theta}}$ = l.s. $\cdot\cdot$ θ under hypothesis above

$$\hat{\hat{\theta}} = (\theta_0, 0 \cdots , 0) \quad \theta_0 = (X_o^T X_o)^{-1} X_o^T \gamma$$

$X_o = p \times q - d$ matrix obtained by removing
      last d-columns.

Let $S(\theta) = \varepsilon^T(\theta)\varepsilon(\theta) \quad \varepsilon(\theta) = \gamma - X\theta , \theta \in \mathbb{R}^2$

Statistic

$S(\hat{\theta}) - S(\hat{\hat{\theta}})$: statistic.

If $S(\hat{\theta}) - S(\hat{\hat{\theta}})$ is large : reject hypothesis.

Proposition.

$\hat{\hat{\theta}}$, $S(\hat{\theta})$, $S(\hat{\hat{\theta}}) - S(\hat{\theta})$ are independent.

$$\frac{1}{\sigma^2} S(\hat{\theta}) \sim \chi^2(p-q) \, , \, \frac{1}{\sigma^2}\left(S(\hat{\hat{\theta}}) - S(\hat{\theta})\right) \sim \chi^2(d).$$

$\left(\text{Note } d = 0 \Rightarrow \hat{\theta} \text{ and } S(\hat{\theta}) \text{ are independent and} \right.$

$$\left. \frac{1}{\sigma^2} S(\hat{\theta}) \sim \chi^2(p-q). \right.$$

Proof:

$0 < d < q$.

X_0 = matrix with d-columns removed from X.

$$Y - X\hat{\theta} \perp \mathcal{R}(x).$$
$$Y - X\hat{\hat{\theta}} \perp \mathcal{R}(X_0).$$
$$\Rightarrow X(\hat{\theta} - \hat{\hat{\theta}}) \perp \mathcal{R}(X_0).$$

$\Rightarrow$ ~~$X(\hat{\theta} - \hat{\hat{\theta}})$~~ $S(\hat{\hat{\theta}}) - S(\hat{\theta}) = \|X(\hat{\theta} - \hat{\hat{\theta}})\|^2$

Let $b_1 \ldots b_p$ o. basis for $\mathbb{R}^p$ s.t.

(i) $b_1 \ldots b_{q-d}$ span $\mathcal{R}(X_0)$.

(ii) $b_{q-d+1} \ldots b_q$ ~~span $\mathcal{R}(X_0)$~~ are orthogonal to the range of X_0

and $b_1 \ldots b_q$ span $\mathcal{R}(X)$.

(iii) $b_{q+1} \ldots b_p$ are orthogonal to the range of X.

Known σ^2

Statistic $\frac{1}{\sigma^2}\left(S(\hat{\hat{\theta}}) - S(\hat{\theta})\right)$ has $\chi^2(d)$ distribution

Let K_α be the upper α-percentile of $\chi^2(d)$-distribution

ie. event $X > K_\alpha$ has probability α

$$\Rightarrow \quad S(\hat{\hat{\theta}}) - S(\hat{\theta}) > \sigma^2 K_\alpha \text{ will occur}$$

when hypothesis is true is α

σ^2 unknown.

$$\frac{S(\hat{\hat{\theta}}) - S(\hat{\theta})}{d} \Bigg/ \frac{S(\hat{\theta})}{p-q} \sim F(d, p-q).$$

if hypothesis is true.

Procedure for Choice of Model.

S_n = min. of least squares criterion over
vectors of parameters of dim. n, $n = 1, 2, \ldots$

Let p = # of data points be large

and q/p small q = model order.

Let $g\left(\frac{n}{p}\right) = S_n \quad n = 1, 2, \cdots p.$

g continuously differentiable. $g: [0, 1] \rightarrow \mathbb{R}.$

S_n : uniform discretization of g

$$\left. S_{\hat{n}-1} - S_{\hat{n}} \middle/ \left(\frac{1}{p}\right) \right. \sim -\frac{dg}{dx} \quad \text{as} \quad \hat{x} = \frac{\hat{n}}{p}.$$

(*) can be expressed as

$$-\left.\frac{\frac{dg}{dx}}{g(x)} - K\right|_{x = \frac{\hat{n}}{p}} = 0.$$

or $\left.\frac{d}{dx}\left[\log g(x) + Kx\right]\right|_{x = \frac{\hat{n}}{p}} = 0.$

$\Rightarrow \quad \text{Min} \log g(x) + Kx$

$\Rightarrow \quad \hat{n}$ minimizes $\log\left(S_n\right) + \frac{K}{p} \cdot n.$

Since p is fixed : Minimize

$$A(n) = \log\left\{\frac{S_n}{p}\right\} + \tilde{K} n \qquad \tilde{K} = \frac{K}{p}.$$

$\Rightarrow \quad F\left(1, p-q\right)$. closely approximates $\chi^2(1)$.
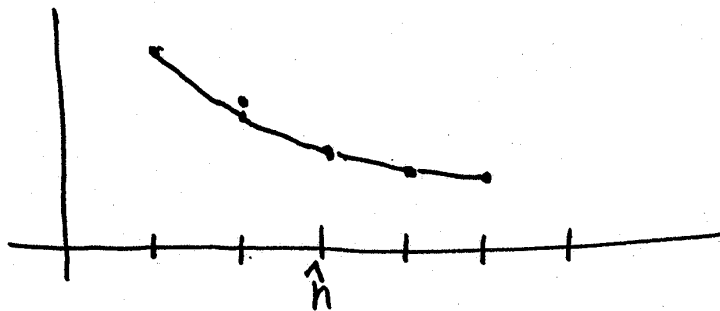
$\Rightarrow \quad$ If n is a possible model order, from Prob.

$$\frac{\left(S_n - S_{n+1}\right)/1}{\dfrac{S_{n+1}}{p}} \quad \text{has approx } \chi^2(1) \text{ distribution.}$$

0.05 percentile for $\chi^2(1)$ is approx. 4.

If $\quad S_n - S_{n+1} > K \dfrac{S_{n+1}}{p} \qquad K = 4$.

grounds for rejecting $\overset{n \text{ as}}{\underset{\wedge}{\text{model}}}$ model order at 5% risk level.

Choose estimate $\hat{n}$ to satisfy

$$S_{n-1} - S_{\hat{n}} > K \dfrac{S_{\hat{n}}}{p}$$

$$S_{\hat{n}} - S_{\hat{n}+1} \leq K \dfrac{S_{\hat{n}+1}}{p} \qquad \quad K \text{ pre-set level.} \qquad (\ast)$$

Can be done since columns of X are independent.

$$\text{Let } v = B^T e \qquad B = (b_1 \vdots \cdots \vdots b_p).$$

$$B^T = B^{-1} \quad (\text{orthonormality}).$$

$$\text{Cov}(v) = \sigma^2 I \implies v \sim N(0, \sigma^2 I).$$

Write

$$e = Y - X\theta^* = (Y - X\hat{\theta}) + X(\hat{\theta} - \hat{\hat{\theta}}) + X(\hat{\hat{\theta}} - \theta^*).$$

$$v = B^{-1}(Y - X\hat{\theta}) + B^{-1}X(\hat{\theta} - \hat{\hat{\theta}}) + B^{-1}X(\hat{\hat{\theta}} - \theta^*).$$

$$X \longmapsto B^{-1}X \quad \text{transform std. basis into}$$

coordinates w.r.t bases $b_1 \cdots b_p$.

Now $(Y - X\hat{\theta}) \perp b_1 \cdots b_q$ $\left(\text{spans } R(X)\right)$

$$\implies B^{-1}(Y - X\hat{\theta}) \in \left\{ \xi \in R^p \mid \xi_1 = \cdots = \xi_q = 0 \right\}.$$

Now $X(\hat{\theta} - \hat{\hat{\theta}}) \in R(X)$

$$\perp R(X_0)$$

$$\implies B^{-1}X(\hat{\theta} - \hat{\hat{\theta}}) \in \left\{ \xi \in R^p \mid \xi_1 = \cdots = \xi_q \right\}$$

<u>Linear algebra shows</u>

$$B^T(Y - X\hat{\theta}) = (0, \ldots, 0, v_{q+i} \ldots, v_p)^T$$

$$B^T(X(\hat{\theta} - \hat{\hat{\theta}})) = (0 \ldots, 0, v_{q-d+1} \ldots, v_q, 0, \ldots 0)^T$$

$$B^T(X(\hat{\theta} - \theta^*)) = (v_1 \ldots v_{q-d}, 0, \ldots 0)^T$$

$$\Rightarrow \quad S(\hat{\hat{\theta}}) - S(\hat{\theta}) = \sum_{i=q-d+1}^{q} v_i^2$$

$$S(\hat{\theta}) = \sum_{i=q+1}^{p} v_i^2$$

$$\Rightarrow \quad \hat{\hat{\theta}} = (X^T X)^{-1} X^T B (v_1 \ldots v_q, 0 \ldots 0)^T$$

$$\left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\} \Rightarrow \begin{array}{l} \hat{\theta}, S(\hat{\theta}) \\[2mm] S(\hat{\hat{\theta}}) - S(\hat{\theta}) \\[2mm] \text{are indep.} \end{array}$$

Now v_i are independent with distribution $N(0, \sigma^2)$

$$\Rightarrow \quad \frac{1}{\sigma^2}\left(S(\hat{\hat{\theta}}) - S(\hat{\theta})\right) \sim \chi^2(d)$$

and $\frac{1}{\sigma^2} S(\hat{\theta}) \sim \chi^2(p-q)$.

Modify the proof when $d = 0$ or q. �

For ARMA model.

$$L_N(\ell, n, \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{K=1}^{N} \varepsilon_K^2(\theta)\right)$$

n, ℓ orders of $A(q)$ and $C(q)$.

θ : parameter vector.

Let $\hat{\sigma}_N^2 = \frac{1}{N} S_N(\hat{\theta}_N)$

Criteria.

$$H_c(\ell, n, \theta, \sigma^2) = \log \hat{\sigma}_N^2 + (\ell + n) \frac{c(N)}{N}.$$

$c(N)$ chosen for consistency.

Choice.

$$c(N) = (\log N)^{1+\delta}$$

upper bds $\bar{n}(N) = \bar{\ell}(N) = (\log N)^{\beta}.$

System Identification as Approximation   4/19/00

1. Parametric Methods.

Choice of Model Structures ⟶

Parametrization.   ⟶   Model Class

i) Model Class contains true system : Consistency

   Analysis

ii) True system not contained in Model class :

   best approximation in Model Class.

Proposition 1

Suppose that the system which generates the data is stable and the model provides uniformly stable predictors. Suppose that the identification criterion is

$$V_N(\theta) = \frac{1}{N} \sum_{K=1}^{N} \varepsilon_K^T(\theta) \, \varepsilon_K(\theta) \qquad N = 1, 2, \cdots$$

where ε_K is the prediction error.

and the limit

$$W(\theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{K=1}^{N} E \left\| \hat{Y}_K - \hat{Y}_K(\theta) \right\|^2 \qquad (4)$$

exists for $\forall \theta \in D$, where $\hat{Y}_K$ is the prediction based on the true system and $\hat{Y}_K(\theta)$ is the prediction based on the model class. Then

$$\theta_N \longrightarrow \left\{ \theta \; \middle| \; W(\theta) = \underset{\lambda \in D}{\text{Min}} \; W(\lambda) \right\} \text{ a.s.}$$

(D is a compact set).

Proof. One can show the following:

θ_N converges, almost surely, into the set

$$\mathcal{D}_I = \left\{ \theta \; \middle| \; \underset{\lambda \in D}{\max} \left(\underset{N \to \infty}{\lim} \inf \frac{1}{N} \sum_{K=1}^{N} \left(E \, \varepsilon_K^T(\theta) \varepsilon_K(\theta) - E \, \varepsilon_K^T(\lambda) \varepsilon_K(\lambda) \right) \right) = 0 \right\}.$$

But for any $\theta \in D$

$$\varepsilon_K(\theta) = \hat{Y}_K - \hat{Y}_K(\theta) + \nu_K$$

where $\nu_K = Y_K - \hat{Y}_K$ (innovations).

$$E\left[\varepsilon_K^T(\theta)\varepsilon_K(\theta)\right] = E\left\|\hat{Y}_k - \hat{Y}_k(\theta)\right\|^2 + E\left\|\nu_k\right\|^2$$

$$\text{(2)}$$

from the property of the Innovations.

From (1) and (2), we can write

$$\lim_{N\to\infty} \frac{1}{N}\sum_{K=1}^{N} E\left\|\hat{Y}_k - \hat{Y}_k(\theta)\right\|^2 \leq \lim_{N\to\infty} \frac{1}{N}\sum_{K=1}^{N} E\left\|\hat{Y}_k - \hat{Y}_k(\lambda)\right\|^2$$

$$\forall \lambda \in D$$

and hence the Proposition. ▨

Remark: i) Existence of limit in (1) is reasonable when the system is LTI and driven by a disturbance which is stationary and when the input is in the form of a linear time-varying feedback (defined through parameters which have limits as $t \to \infty$) together with stationary additive disturbance.

ii) The proposition states that the identification

procedure selects a model $M(\theta^*)$ which is a best approximation for a particular input and not for all inputs. Hence the choice of inputs in the identification experiment is an important design issue.

The question of Bias.

Consider the dynamical system
$$Y_K = a\, Y_{K-1} + W_K \quad , \quad K \in \mathbb{Z}$$
where the disturbances W_K are generated by
$$W_K = V_K + c\, V_{K-1} \quad K \in \mathbb{Z}.$$
$a, c \in \mathbb{R}$, $|a| < 1$. V_K: zero-mean, uncorrelated, uniformly bounded fourth order moments
$$Var(V_K) = R_{vv}(0) > 0 \quad (\text{indep. of } K).$$
Least-squares estimate $\hat{a}_N$ of a given data for $K = 1, \cdots N$, disregarding correlation
$$\hat{a}_N = \left(\frac{1}{N} \sum_{K=1}^{N} y_{K-1}^2 \right)^{-1} \left(\frac{1}{N} \sum_{K=1}^{N} y_K y_{K-1} \right.$$

Let $d_k = Y_{k-1}^2$ or $Y_k Y_{k-1}$ $k = 1, 2 \cdots$.

We can show

$$\text{Cov}\left(d_t, d_{t+s}\right) \leq \text{const.} \cdot \lambda^s \quad t, s \geq 0.$$
$$\lambda \in (0, 1)$$

Hence by the Ergodic Th. for non-stationary processes

$$\frac{1}{N} \sum_{k=1}^{N} Y_{k-1}^2 \longrightarrow R_{yy}(0) \text{ a.s.}$$
$$= E\left(Y_{k-1}^2\right)$$

and $\frac{1}{N} \sum_{k=1}^{N} Y_k Y_{k-1} \longrightarrow R_{yy}(1)$ a.s.

$$= E\left(Y_k Y_{k-1}\right)$$

Hence

$$\hat{a}_N \longrightarrow \frac{R_{yy}(1)}{R_{yy}(0)} \quad \text{as } N \to \infty \text{ a.s.}$$

(Asymptotic estimate)

Now from the eqns.

$$Y_k Y_{k-1} = a\left(Y_{k-1}\right)^2 + \left(V_k + c V_{k-1}\right) Y_{k-1}$$

and $Y_k V_k = a Y_{k-1} V_k + \left(V_k + c V_{k-1}\right) V_k$.

Taking expectations and using the properties of V_k being uncorrelated with V_{k-1}

and Y_{k-1}, we get

$$R_{yy}(1) = a R_{yy}(0) + c R_{Yv}(0).$$

and $R_{Yv}(0) = R_{rv}(0)$

Hence $\hat{a}_N \longrightarrow a + c \dfrac{R_{vv}(0)}{R_{YY}(1)}$ as $N \longrightarrow \infty$ a.s.

Hence there is an asymptotic bias
$$c \frac{R_{vv}(0)}{R_{YY}(1)}$$

Implications for Control.

(a) Identify $M(\theta^*)$ and then apply control based on identified model (e.g. Minimum Variance Control)

(b) At time k, obtain $M(\theta_k)$ on the basis of Y^k, u^{k-1} and then obtain minimum variance control based on $M(\theta_k)$ to obtain Y^{k+1}, u^k etc.

First approach gives biased estimates and poor control. Second approach gives biased estimates.

Issues.

1) **Consistency** : this is really well-posedness.
   If the true system $M(\theta^*)$ is an element
   of the model set is it true that
   the estimate $\hat{\theta}_N \longrightarrow \theta^*$ a.s.

2) **Asymptotic variance** : $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \left(\hat{\theta}_N - \theta\right)$

$$\longrightarrow N(0, G).$$

Non-parametric Identification

$$\hat{Y}(t|\theta) = g\left(Z^{t-1}, \theta\right).$$

Represent:

$$g\left(\underset{\uparrow}{Z^{t-1}}, \theta\right) = g\left(\varphi(t), \theta\right) \qquad \varphi : \text{ of fixed dimension sufficient statistic.}$$

growing

$$\varphi(t) = \varphi\left(Z^{t-1}\right).$$

or $\varphi(t) = \varphi\left(Z^{t-1}, \theta\right)$: parametrized.

Chose: Regression vector

$$\varphi(t) \quad \text{from} \quad Z^{t-1} \quad (\text{past inputs \& outputs}).$$

Chose non-linear mapping

$$g(\varphi, \theta)$$

– Let g map $\mathbb{R}^d \longrightarrow \mathbb{R}^1$ (scalar output).

for each θ

Write: $g(\varphi, \theta) = \sum_{K=1}^{n} \alpha_K g_K(\varphi)$ $\theta = (\alpha_1 \cdots \alpha_n)$.

g_K : basis functions. $\Big\}$ "Gabor" fns.

$$g_K(\varphi) = \kappa\left(\beta_K(\varphi - \gamma_K)\right)$$

$\uparrow$ dilation $\uparrow$ translation

$$\kappa : \mathbb{R}^1 \longrightarrow \mathbb{R}^{1.}$$

Ex: $g_K(\varphi) = \kappa\left(\|\varphi - \gamma_K\|_{\beta_K}\right)$ $\underline{\text{Radial Basis}}$

Value of fn. depends only on distance from center.

$\|\cdot\|_{\beta_K}$ norm on regression φ.

Ridge basis

$$g_k(\varphi) = \zeta\left(\beta_k^T \varphi + r_k\right) : \varphi \in \mathbb{R}^d.$$

function depends only on φ's distance from hyperplane

Ridge fn. constant $\forall \varphi$ on $\left\{\varphi \mid \beta_k^T \varphi = \text{const.}\right\}$

Neural Networks.

$$\varphi_k^{(2)}(t) = g_k(\varphi) = \zeta\left(\varphi(t), \beta_k, r_k\right).$$

$$\varphi^{(2)}(t) = \left[\varphi_1^{(2)}(t), \dots, \varphi_n^{(2)}(t)\right]$$

$$g(\varphi, \theta) = \sum_{\ell} \alpha_\ell^{(2)} \zeta\left(\varphi^{(2)}, \beta_\ell^{(2)}, r_\ell^{(2)}\right)$$

↑ New regressors.

In general:

$$F : \mathcal{U} \longrightarrow \mathcal{Y} \quad \text{causal}$$

$$\pi_t F(\pi_t u) = \pi_t y \qquad \text{Approximate causal}$$
$$\text{Input output map.}$$