MODEL STRUCTURE SELECTION AND MODEL VALIDATION

ADVANCED TOPICS

- **F-Test and Consistency of Information Criteria**

- **Methods for Estimating Model Quality:**
  - **Deterministic:** **Set Membership**
    Estimation in l_1
    $\mathcal{H}_\infty$ Identification
  - **Stochastic:** **Stochastic Embedding**
    Model Error Modeling

F-TEST

Assume that $S \in \mathcal{M}_1 \subset \mathcal{M}_2$, where $\theta_1 \in \mathbb{R}^{n_1}$ and $\theta_2 \in \mathbb{R}^{n_2}$, and let V_N^i be the minimum of $V_N(\theta)$ in $\mathcal{M}_i$. Then, it can be shown that [Söderström & Stoica, Appendix A11.2]

$$N\frac{V_N^1 - V_N^2}{V_N^2} \xrightarrow[N\to\infty]{d} \chi^2(n_2 - n_1)$$

This gives rise to the *F-test* for model order selection (of significance α):

$$\boxed{\text{Choose } \mathcal{M}_1 \text{ over } \mathcal{M}_2 \text{ iff } N\frac{V_N^1 - V_N^2}{V_N^2} \le \chi_\alpha^2(n_2 - n_1)}$$

or equivalently, iff

$$V_N^1 \le V_N^2\left[1 + \frac{1}{N}\chi_\alpha^2(n_2 - n_1)\right]$$

EQUIVALENCE BETWEEN F-TEST AND INFORMATION CRITERIA

Consider a model selection criterion of the form

$$W_N = \underbrace{-2l(\hat{\theta}_N)}_{N \ln V_N} + \gamma(N, n)$$

This criterion selects $\mathcal{M}_1$ over $\mathcal{M}_2$ iff

$$V_N^1 \leq V_N^2 \exp\left[\frac{\gamma(N, n_2) - \gamma(N, n_1)}{N}\right]$$

Therefore, we can interpret this criterion as the F-test with significance level such that

$$\chi_\alpha^2(n_2 - n_1) = N\left(\exp\left[\frac{\gamma(N, n_2) - \gamma(N, n_1)}{N}\right] - 1\right)$$

E.g. for:

AIC: $\chi_\alpha^2(n_2 - n_1) \approx 2(n_2 - n_1)$

BIC: $\chi_\alpha^2(n_2 - n_1) \approx (n_2 - n_1)\ln N$

CONSISTENCY OF INFORMATION CRITERIA

Risk of Overfitting:

If $S \in \mathcal{M}_1 \subset \mathcal{M}_2$, the probability of choosing $\mathcal{M}_2$ (with the F-test) is

$$P\left\{V_N^1 > V_N^2\left[1+\frac{1}{N}\chi_\alpha^2(n_2-n_1)\right]\right\} = P\left\{N\frac{V_N^1-V_N^2}{V_N^2} > \chi_\alpha^2(n_2-n_1)\right\} = \alpha$$

Thus, to avoid risk of overfitting as $N \to \infty$ we require $\alpha \to 0$, or equivalently

$$\chi_\alpha^2(n_2-n_1) \xrightarrow[N\to\infty]{} \infty$$

CONSISTENCY OF INFORMATION CRITERIA (CONT.)

Risk of Underfitting:

If $S \notin \mathcal{M}_1 \subset \mathcal{M}_2$, the probability of choosing $\mathcal{M}_1$ (with the F-test) is

$$P\left\{N\frac{V_N^1 - V_N^2}{V_N^2} < \chi_\alpha^2(n_2 - n_1)\right\} = P\left\{\frac{V_N^1 - V_N^2}{V_N^2} < \frac{\chi_\alpha^2(n_2 - n_1)}{N}\right\}$$

In this case, $V_N^1 - V_N^2 \not\to 0$ as $N \to \infty$, so a sufficient condition for this probability to tend to 0 is that

$$\frac{\chi_\alpha^2(n_2 - n_1)}{N} \xrightarrow[N\to\infty]{} 0$$

Consistency of Information Criteria (cont.)

Consistency of AIC and BIC:

From the previous conditions, we have that

For AIC: Risk of underfitting $\to 0$
Risk of overfitting $\not\to 0$ Hence, AIC is *inconsistent*

For BIC: Risk of underfitting $\to 0$
Risk of overfitting $\to 0$ Hence, BIC is *consistent*

ESTIMATION OF MODEL QUALITY

We know how to estimate the variance of a model $\hat{G}$, but...

Question: How can we estimate the bias of a model?

Some approaches:

- Deterministic: Set Membership
  Estimation in l_1
  $\mathcal{H}_\infty$ Identification

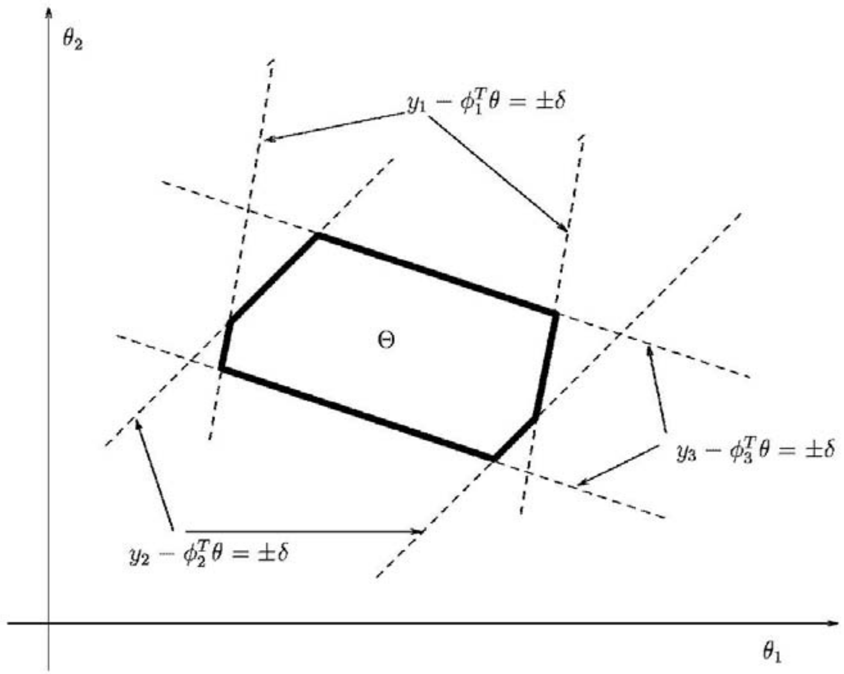
- Stochastic: Stochastic Embedding
  Model Error Modeling

Estimation of Model Quality (cont.)

Deterministic Approaches: *Set membership/worst-case estimation*

Model: $y_t = G(q;\theta)u_t + v_t$

Only assumption on $\{v_t\}$: $|v_t| \le \delta, \quad t = 1, \dots, N$



The bias error is included in $\{v_t\}$, since there are no assumptions on its whiteness or independence of $\{u_t\}$

Problems: - It needs an upper bound for δ
- It is inconsistent if δ is not *tight*

Estimation of Model Quality (cont.)

Deterministic Approaches (cont.): *Estimation in l_1*

Model: $$y_t = (g * u)_t + v_t = \sum_{k=0}^{\infty} g_k u_{t-k} + v_t$$

Assumptions: $$g \in \mathcal{G} = \{g : |g_k| \le M\rho^{-k}, k \in \mathbb{N}\}$$
$$|v_t| \le \delta, \quad t = 1, \ldots, N$$

Goal: $$g^{opt} = \arg\min_{\hat{g} \in \mathcal{G}} \sup_{\substack{g \in \mathcal{G} \\ \|v\|_\infty \le \delta}} \|g - \hat{g}\|_1$$

This approach is similar to set membership, but its motivation comes from $\mathcal{H}_\infty$ identification, since the l_1-norm is an upper bound for the $\mathcal{H}_\infty$ norm of a system

ESTIMATION OF MODEL QUALITY (CONT.)

Deterministic Approaches (cont.): $\mathcal{H}_\infty$ *identification*

Model: $f_k = G(e^{j2\pi k/n}) + v_k = \sum_{m=0}^{d-1} g_m e^{j2\pi mk/n} + v_k, \quad k = 1,\dots,n$ (freq-domain data)

Assumptions: $G \in \mathcal{H}_\infty(D_\rho)$ where $D_\rho := \{z \in \mathbb{C} : \|z\| > \rho\}$, for some $0 < \rho < 1$

$|v_k| \leq \varepsilon, \quad k = 1,\dots,n$

Goal: Find $\hat{G} \in \mathcal{H}_\infty(D_\rho)$ which is consistent in the sense that

$$\lim_{\substack{\varepsilon \to 0 \\ n \to \infty}} \sup_{\substack{G_0 \in \mathcal{H}_\infty(D_\rho) \\ \|v\|_\infty \leq \varepsilon}} \left\| G_0 - \hat{G} \right\|_\infty = 0$$

where G_0 is the system that actually generated the data $\{f_k\}$

ESTIMATION OF MODEL QUALITY (CONT.)

Stochastic Approaches: *Stochastic Embedding*

Model:

$$y_t = \underbrace{G(q;\theta)}_{\text{nominal model}} u_t + \underbrace{G_\Delta(q;\lambda)}_{\text{undermodeling}} u_t + v_t$$

$$G_\Delta(q;\lambda) = \sum_{k=0}^{L-1} \eta_k(\lambda) q^{-k}$$

Assumptions: η_k's are independent *Gaussian*, with $E\{\eta_k\} = 0$, $E\{\eta_k^2\} = M\rho^{-k}$
$\{v_k\}$ is zero mean Gaussian white noise with variance λ

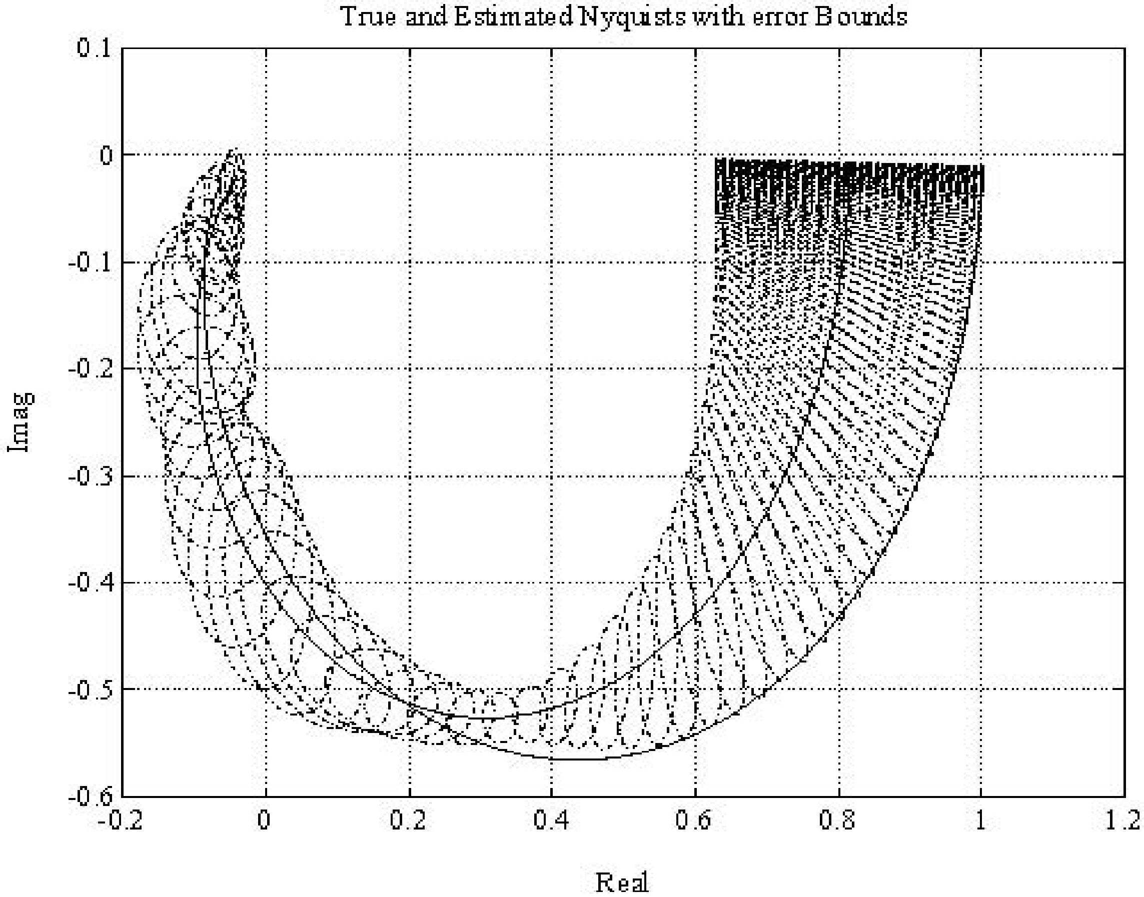
Idea: This is a *quasi-Bayesian* approach (θ is deterministic and $\{\eta_k\}$ are random)
θ can be estimated via ML (e.g., by maximizing $P\{Y^N;\theta\}$)

If M, ρ and λ are unknown, they can also be estimated via ML (by maximizing $P\{Y^N;\theta, M, \rho, \lambda\}$)

Problem: This approach does not have a sound physical interpretation, and it doesn't give hard bounds on undermodeling

ESTIMATION OF MODEL QUALITY (CONT.)

Stochastic Approaches (cont.): *Stochastic Embedding (cont.)*



Estimation of Model Quality (cont.)

Stochastic Approaches (cont.): *Model Error Modeling*

Nominal Model: $y_t = G(q;\theta)u_t + \varepsilon_t$

Idea: Estimate a nominal model $\hat{G}_N$. If there is undermodeling, $\{\varepsilon_t\}$ is not white noise independent of $\{u_t\}$, so we can fit a model to $\{\varepsilon_t\}$!

Model Error Model: $\varepsilon_t = G_\Delta(q)u_t + v_t$

Typically G_Δ is a large model, so we can construct confidence regions for it, which can be translated as uncertainty regions for $\hat{G}_N$ (which also account for its bias!)

Model validation: $\hat{G}_N$ is validated if 0 belongs to the confidence region of G_Δ

Problem: This approach is highly sensitive to the order of G_Δ

ESTIMATION OF MODEL QUALITY (CONT.)

Stochastic Approaches (cont.): *Model Error Modeling (cont.)*