MODEL QUALITY EVALUATION

- **Bias / Variance Tradeoff**
- **Stein's Paradox and Biased Estimators**
- **Confidence Intervals/Regions**
- **Variance Error Quantification**
- **Geometric Approach to Variance Analysis**

Bias / Variance Tradeoff

Def: The *mean square error* (MSE) of an estimator $\hat{\theta}_N$ of a parameter θ is

$$MSE(\hat{\theta}_N) := E\left\{\left\|\hat{\theta}_N - \theta\right\|^2\right\}$$

$$= \underbrace{E\left\{\left\|\hat{\theta}_N - E\{\hat{\theta}_N\}\right\|^2\right\}}_{\text{tr}\{\text{cov}\,\hat{\theta}_N\}} + \underbrace{\left\|E\{\hat{\theta}_N\} - \theta_0\right\|^2}_{\left\|\text{bias}\,\hat{\theta}_N\right\|^2}$$

Here the *(parametric) bias on* $\hat{\theta}_N$ is $E\{\hat{\theta}_N\} - \theta_0$

BIAS / VARIANCE TRADEOFF (CONT.)

In terms of G, we have, for N large enough,

$$MSE(\hat{G}_N(e^{j\omega})) = E\left\{\left\|\hat{G}_N(e^{j\omega}) - G_0(e^{j\omega})\right\|^2\right\}$$
$$\approx E\left\{\left\|\hat{G}_N(e^{j\omega}) - G_*(e^{j\omega})\right\|^2\right\} + \left\|G_*(e^{j\omega}) - G_0(e^{j\omega})\right\|^2$$

In system identification it is common to define $G_*(e^{j\omega}) - G_0(e^{j\omega})$ as the *(asymptotic) bias* of $\hat{G}_N(e^{j\omega})$. Because of the convergence of PEM under mild conditions, this bias is due exclusively to undermodelling

Bias/Variance Tradeoff:

By increasing the model set, we can in general reduce the bias of G and H. However, the variance of G and H will increase (recall $\operatorname{var}\hat{G}_N \approx (n/N)(\Phi_v/\Phi_u)$, which increases with n)

STEIN'S PARADOX AND BIASED ESTIMATORS

The C-R bound establishes a lower bound for the MSE of unbiased estimators
Is it possible to obtain better results with biased estimators?

Stein's Paradox:

Let $Y \sim N(\theta, \sigma^2 I)$, where σ^2 is known, and $Y, \theta \in \mathbb{R}^n$. The MVU estimator of θ is $\hat{\theta}_{MVU} = Y$. James and Stein (1961) proposed

$$\hat{\theta}_{JS} = \left(1 - \frac{(n-2)\sigma^2}{\|Y\|^2} \right) Y$$

and showed that for $n > 2$, $MSE(\hat{\theta}_{JS}) < MSE(\hat{\theta}_{MVU})$ for every θ!

The idea of James and Stein was to scale the $\hat{\theta}_{MVU}$ $\Rightarrow$ *Shrinkage Estimators*

STEIN'S PARADOX AND BIASED ESTIMATORS (CONT.)

The shrinkage idea can be extended to general estimators: (Kay and Eldar, 2008)
If $\hat{\theta}_u$ is an unbiased estimator of θ (scalar), take

$$\hat{\theta}_b = (1+m)\hat{\theta}_u$$

Then: $$MSE(\hat{\theta}_b) = (1+m^2)\operatorname{var}\hat{\theta}_u + m^2\theta^2$$

which is minimized at: $$m = -\frac{1}{1+\theta^2 / \operatorname{var}\{\hat{\theta}_u\}}$$

If $\theta^2 / \operatorname{var}\{\hat{\theta}_u\}$ is constant, this can be easily obtained. Otherwise, if $\theta \in \Theta$ consider

$$m^* = \arg\min_{m\in\mathbb{R}} \max_{\theta\in\Theta} [MSE(\hat{\theta}_b) - MSE(\hat{\theta}_u)]$$

CONFIDENCE INTERVALS/REGIONS

Def: A *confidence interval of a parameter* θ_0 is an interval (θ_1, θ_2), where $\theta_i = g_i(y)$. It has a *confidence coefficient of* $100\alpha\%$ if $P\{\theta_1 < \theta_0 < \theta_2\} = \alpha$. $1-\alpha$ is called the *confidence level* of (θ_1, θ_2)

θ_1 and θ_2 are not unique for a given α, so we prefer $E\{|\theta_2 - \theta_1|\}$ to be minimum

These concepts can be generalized to multi-dimensional *confidence regions*

Asymptotic Regions

If $\hat{\theta} \in \mathbb{R}^p$ is asymptotically normal, then for N large enough,

$$P\{(\hat{\theta} - \theta_0)^T P_\theta^{-1} (\hat{\theta} - \theta_0) < \chi_\alpha^2(p)\} \approx \alpha$$

where $\chi_\alpha^2(p)$ is the α-percentile of the $\chi^2(p)$ distribution

Then, an confidence ellipsoid for θ_0 of level $1-\alpha$ is $\{\theta_0 : (\hat{\theta} - \theta_0)^T P_\theta^{-1} (\hat{\theta} - \theta_0) < \chi_\alpha^2(p)\}$

VARIANCE ERROR QUANTIFICATION

Covariance Estimators:

$S \in \mathcal{M}$: The (normalized by N) covariance matrix of $\hat{\theta}_N$ can be estimated as:

$$\hat{P}_N := \hat{\lambda}_N \left[\frac{1}{N} \sum_{t=1}^{N} \psi_t(\hat{\theta}_N) \psi_t^T(\hat{\theta}_N) \right]^{-1}$$

$$\hat{\lambda}_N := \frac{1}{N} \sum_{t=1}^{N} \varepsilon_t^2(\hat{\theta}_N)$$

$S \notin \mathcal{M}$: "Sandwich" estimator (White, 1982)

$$\hat{P}_N := [V_N''(\hat{\theta}_N)]^{-1} \left[\sum_{t=1}^{N-1} V_t'(\hat{\theta}_N) V_t'^T(\hat{\theta}_N) \right] [V_N''(\hat{\theta}_N)]^{-1}$$

See also (Hjalmarsson and Ljung, 1992)

Variance Error Quantification (cont.)

Confidence Regions for θ:

Asymptotic confidence ellipsoid: $U_\theta := \{\theta : N(\hat{\theta}_N - \theta)^T \hat{P}_N^{-1}(\hat{\theta}_N - \theta) < \chi_\alpha^2(p)\}$

U_θ contains θ_0 with confidence α (assuming $S \in \mathcal{M}$)

Confidence Regions for G and H:

$$\begin{bmatrix} \operatorname{Re}\hat{G}_N(e^{j\omega}) \\ \operatorname{Im}\hat{G}_N(e^{j\omega}) \end{bmatrix} \approx \begin{bmatrix} \operatorname{Re}G_0(e^{j\omega}) \\ \operatorname{Im}G_0(e^{j\omega}) \end{bmatrix} + \Gamma(e^{j\omega})[\hat{\theta}_N - \theta_0], \quad \Gamma(e^{j\omega}) = \begin{bmatrix} \partial \operatorname{Re}G_\theta(e^{j\omega}) \big/ \partial\theta^T \\ \partial \operatorname{Im}G_\theta(e^{j\omega}) \big/ \partial\theta^T \end{bmatrix}_{\theta_0}$$

Confidence ellipsoid:

$$U_G(e^{j\omega}) := \left\{ G : N \begin{bmatrix} \operatorname{Re}\hat{G}_N - \operatorname{Re}G \\ \operatorname{Im}\hat{G}_N - \operatorname{Im}G \end{bmatrix}^T [\Gamma \hat{P}_N \Gamma^T]^{-1} \begin{bmatrix} \operatorname{Re}\hat{G}_N - \operatorname{Re}G \\ \operatorname{Im}\hat{G}_N - \operatorname{Im}G \end{bmatrix} < \chi_\alpha^2(p) \right\}$$

GEOMETRIC APPROACH TO VARIANCE ANALYSIS

Consider SISO LTI models with G_ρ and H_η in open loop

Idea: The (per sample) information matrix for ρ is a *Gramian*

$$P_\rho^{-1} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma(e^{j\omega})\Gamma^H(e^{j\omega})\frac{\Phi_u(\omega)}{\Phi_v(\omega)}d\omega = \left\langle \Gamma, \Gamma \right\rangle_{\Phi_u/\Phi_v}$$

Hence, if $J : D_{\mathcal{M}} \to \mathbb{R}$ is a function of θ (e.g. G_ρ),

$$\begin{aligned}\operatorname{var} \hat{J}_N &\approx \frac{1}{N}\frac{\partial J}{\partial \theta^T}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma(e^{j\omega})\Gamma^H(e^{j\omega})\frac{\Phi_u(\omega)}{\Phi_v(\omega)}d\omega\right]^{-1}\frac{\partial J}{\partial \theta} \\ &= \frac{1}{N}\frac{\partial J}{\partial \theta^T}\left\langle \Gamma, \Gamma \right\rangle_{\Phi_u/\Phi_v}^{-1}\frac{\partial J}{\partial \theta}\end{aligned}$$

Geometric Approach to Variance Analysis (cont.)

This can be further simplified if there is a function γ such

$$\frac{\partial J}{\partial \theta} = \langle \Gamma, \gamma \rangle_{\Phi_u/\Phi_v}$$

because in this case we have

$$\operatorname{var} \hat{J}_N \approx \frac{1}{N} \langle \gamma, \Gamma \rangle_{\Phi_u/\Phi_v} \langle \Gamma, \Gamma \rangle_{\Phi_u/\Phi_v}^{-1} \langle \Gamma, \gamma \rangle_{\Phi_u/\Phi_v} = \frac{1}{N} \left\| \mathrm{Proj}_{\Gamma}\, \gamma \right\|_{\Phi_u/\Phi_v}^{2}$$

This expression gives a geometric interpretation of $\operatorname{var} \hat{J}_N$, which decomposes the variance error into:

1. Γ: information about model structure
2. Φ_u / Φ_v: experimental conditions
2. γ: quantity of interest (γ can be considered as the *Fréchet derivative* of J w.r.t. G_θ, "$\gamma(e^{j\omega}) = \partial J / \partial G_\theta(e^{j\omega})$")

Geometric Approach to Variance Analysis (cont.)

Example: *Adding parameters increases the variance* (Parsimony Principle)

Let $S \in \mathcal{M}_1 \subset \mathcal{M}_2$, i.e. $\theta_2 = [\ \theta_1^T \quad \theta_\Delta^T\ ]^T$, so that $\mathcal{M}_2([\ \theta_1^T \quad 0\ ]^T) = \mathcal{M}_1(\theta_1)$. Then

$$\text{rowspace}\{\Gamma_2\} = \text{rowspace}\{\Gamma_1\} \oplus \mathcal{X}$$

so

$$\begin{aligned} \text{var}_{\mathcal{M}_2}\{\hat{J}_N\} &\approx \frac{1}{N}\left\|\text{Proj}_{\Gamma_2}\gamma\right\|^2_{\Phi_u/\Phi_v} \\ &= \frac{1}{N}\left\|\text{Proj}_{\Gamma_1}\gamma\right\|^2_{\Phi_u/\Phi_v} + \left\|\text{Proj}_{\mathcal{X}}\gamma\right\|^2_{\Phi_u/\Phi_v} \\ &\geq \text{var}_{\mathcal{M}_1}\{\hat{J}_N\} \end{aligned}$$

with equality iff $\gamma \in \text{rowspace}\{\Gamma_1\}$