Model Class Selection

- **Given**: Data D from system and set M of candidate model classes

$$M = \{M_1, M_2, ..., M_J\}$$

where each model class M_j defines a set of possible predictive models for system:

$$\{p(Y_n | U_n, \theta_j) : \theta_j \in \Theta_j \subset R^{N_j}\}$$

& a probability model $p(\theta_j | M_j)$ over this set

- **Find**: Most plausible model class
- **Goal:** Selection of level of model complexity

Model Class Selection

- **Most Plausible Model Class Based on Data** D

  Prior info

  Maximize:

  $$P(M_j \mid D, M) \text{ over all j}$$

- **Higher level of robustness:** Can include predictions of all model classes (model class averaging):

$$p(Y_n \mid U_n, D, M) = \sum_{j=1}^{J} p(Y_n \mid U_n, D, M_j) P(M_j \mid D, M)$$

Model Class Selection

- **Evaluation of Model Class Probability**

Bayes Theorem:

Evidence

Prior

$$P(M_j \mid D, M) = \frac{p(D \mid M_j) P(M_j \mid M)}{p(D \mid M)}$$

where denominator is chosen to normalize

$P(M_j \mid D, M)$ over $j = 1, \ldots, J$

Evaluation of Evidence

- **Total Probability Theorem gives evidence:**

$$p(D|M_j) = \int_{\mathbf{\Theta}_j} p(D|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j$$

- **Can use asymptotic expansion about MPV**

$$p(D|M_j) \approx (2\pi)^{\frac{N_j}{2}} \frac{p(D|\hat{\boldsymbol{\theta}}_j, M_j)\, p(\hat{\boldsymbol{\theta}}_j|M_j)}{\sqrt{\det \mathbf{H}_j(\hat{\boldsymbol{\theta}}_j)}}$$

$$\mathbf{H}_j(\boldsymbol{\theta}_j) = -\nabla\nabla \ln p(D|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)$$

Model Class Selection using Evidence

- Assume all model classes equally plausible *a priori*, then plausibility of each model class M_j is ranked by its **log evidence**:

$$\ln p(D|M_j) \approx \ln p(D|\hat{\boldsymbol{\theta}}_j, M_j) + \left[\ln p(\hat{\boldsymbol{\theta}}_j | M_j) - \tfrac{1}{2}\ln\det \mathbf{H}_j(\hat{\boldsymbol{\theta}}_j) + \tfrac{N_j}{2}\ln(2\pi)\right]$$

= **log likelihood + log Ockham factor**

= **Data fit + Bias against parameterization**

- Gives a quantitative Principle of Parsimony

Bias Against Parameterization

- **Log Ockham Factor β_j for M_j:**

For a large number *N* of data points in *D*,

$$\beta_j \approx -\sum_{i=1}^{N_j} \ln \frac{\rho_{j,i}}{\sigma_{j,i}} - \frac{1}{2}\sum_{j=1}^{N_j}\left(\frac{\hat{\theta}_{j,i} - \bar{\theta}_{j,i}}{\rho_{j,i}}\right)^2$$

where $\rho^2{}_{j,i}, \sigma^2{}_{j,i}$ are the prior and principal posterior variances for θ_j and $\bar{\theta}_{j,i}, \hat{\theta}_{j,i}$ are the prior and posterior most probable values of $\theta_{j,i}$

$$\Rightarrow \beta_j = -\tfrac{1}{2} N_j \ln N + O(1) \quad \text{(for large } N\text{)}$$

So Log Ockham factor decreases with number of model parameters

Interpretation using information theory

- From asymptotics for large amount of data *N* and globally identifiable model classes (Beck and Yuen 2004):

Log evidence = [Data fit of optimal model] – [Information gain about θ_j in *D*]

Recently generalized this result to any model class

Comparison with AIC and BIC

- **Bayesian model class selection criterion**

  Maximize $\ln P\left(M_j \mid D, M_j\right)$ w.r.t. M_j, or equivalently (from asymptotic result):

  log evidence = log likelihood + log Ockham factor

  i.e. $\ln p\left(D \mid M_j\right) = \ln p\left(D \mid \hat{\theta}_j, M_j\right) + \beta_j$

- **Akaike (1974)**

  Maximize: **AIC** = $\ln p\left(D \mid \hat{\theta}_j, M_j\right) - N_j$

- **Akaike (1976), Schwarz (1978)**

  Maximize: **BIC** = $\ln p\left(D \mid \hat{\theta}_j, M_j\right) - \frac{N_j}{2} \ln N$

  (agrees with above criterion for large *N* except for terms of *O*(1))

Evaluation of Likelihood

- Likelihood function $p(D \mid \theta_j, M_j)$ is based on **prediction-error model**:

  Predicted response

  = (Stochastic) response of model θ_j

  + Prediction error

- In examples, prediction error η modeled as zero-mean Gaussian discrete white noise with covariance matrix $\sigma_\eta^2\, \mathrm{I}$ (i.e. maximum information entropy PDF)

Evaluation of Likelihood

- **Details for dynamical models with input-output measurements:**

  e.g. Beck & Katafygiotis: “Updating models and their uncertainties. I: Bayesian statistical framework”, *J. Engng Mech*., April 1998.

- **Details for output-only measurements:**

  e.g. Yuen and Beck: “Updating properties of nonlinear dynamical systems with uncertain input”, *J. Engng Mech*., Jan. 2003.

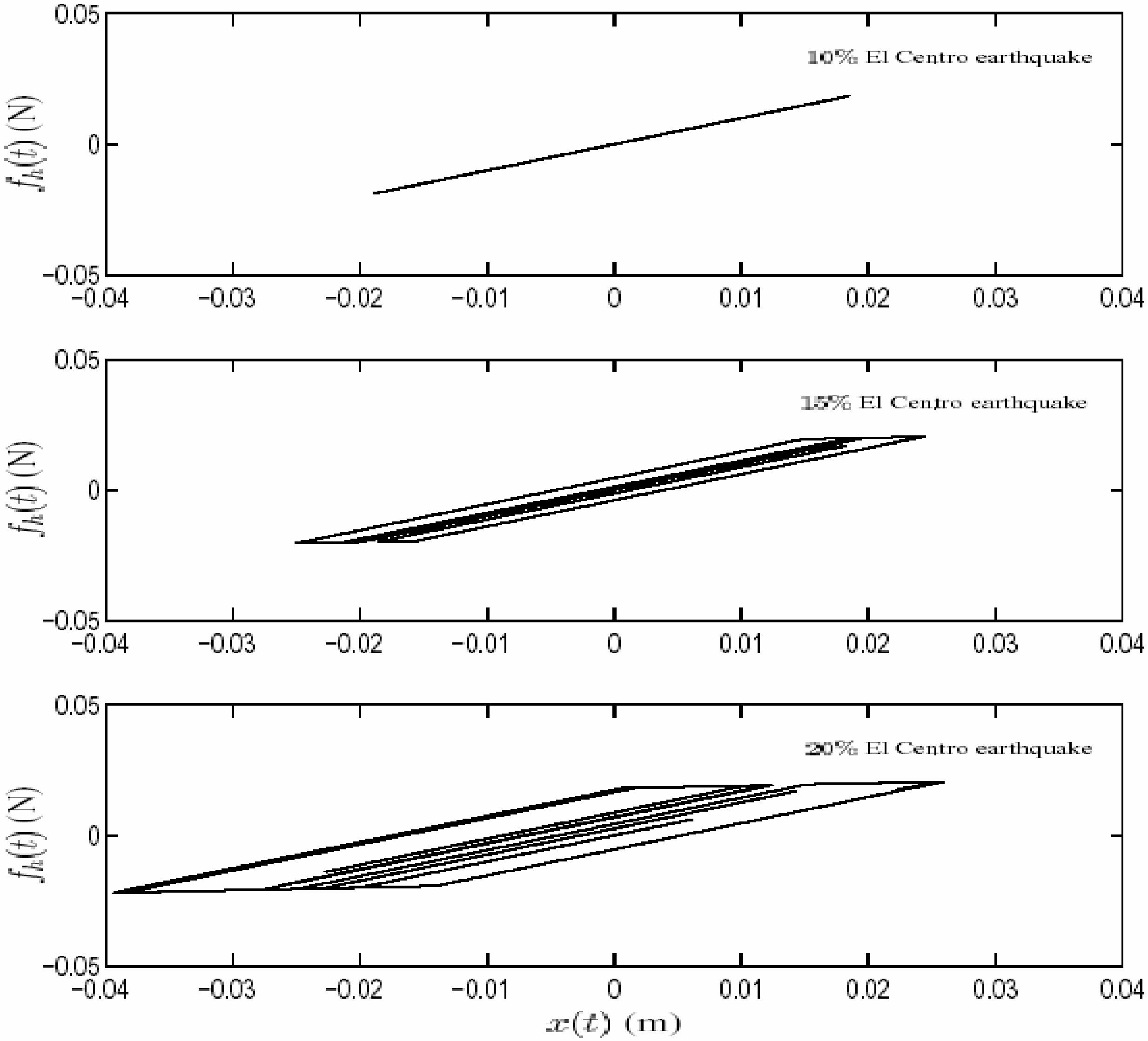
Example 1: SDOF Hysteretic Oscillator

$$m\ddot{x} + c\dot{x} + f_s(x; k_1, k_2, x_y) = f(t)$$

- f_s = bilinear hysteretic restoring force
- f = scaled 1940 El Centro earthquake record
- Simulated noise (5% of rms simulated displacement)
- Prediction error η modeled as zero-mean Gaussian discrete white noise with variance σ_η^2
  i.e. predicted displacement at time step n,

$$\hat{x}(n) = x(n) + \eta(n)$$

Hysteretic force-displacement behavior



Example 1: Choice of Model Classes

- **Model Class 1** ($M_{\mathbf{1}}$ - 3 parameters)
  Linear oscillators with damping coefficient c>0, stiffness $k_1 > 0$ and prediction-error variance σ_η^2
- **Model Class 2** ($M_{\mathbf{2}}$ - 3 parameters)
  Elasto-plastic oscillators (i.e. $k_2 = 0$) with stiffness $k_1 > 0$, yield displacement x_y and prediction-error variance σ_η^2
- Independent uniform prior distributions on all parameters

Example 1: Conclusions

- Class of linear models (M_1) much more probable than elasto-plastic models (M_2) for lower level excitation, but other way around for higher levels
- Illustrates an important point: there is no exact class of models for a real system and the most probable class may depend on the excitation level.

Example 2: Modal Model for 10-Story Linear Shear Building

- Examine most plausible number of modes based on measured accelerations at the roof during base excitation
- Excitation not measured; modeled as stationary Gaussian white noise with uncertain spectral intensity
- Other model parameters: Modal frequencies, modal damping ratios and prediction-error variance

Example 2: Most Probable Frequencies

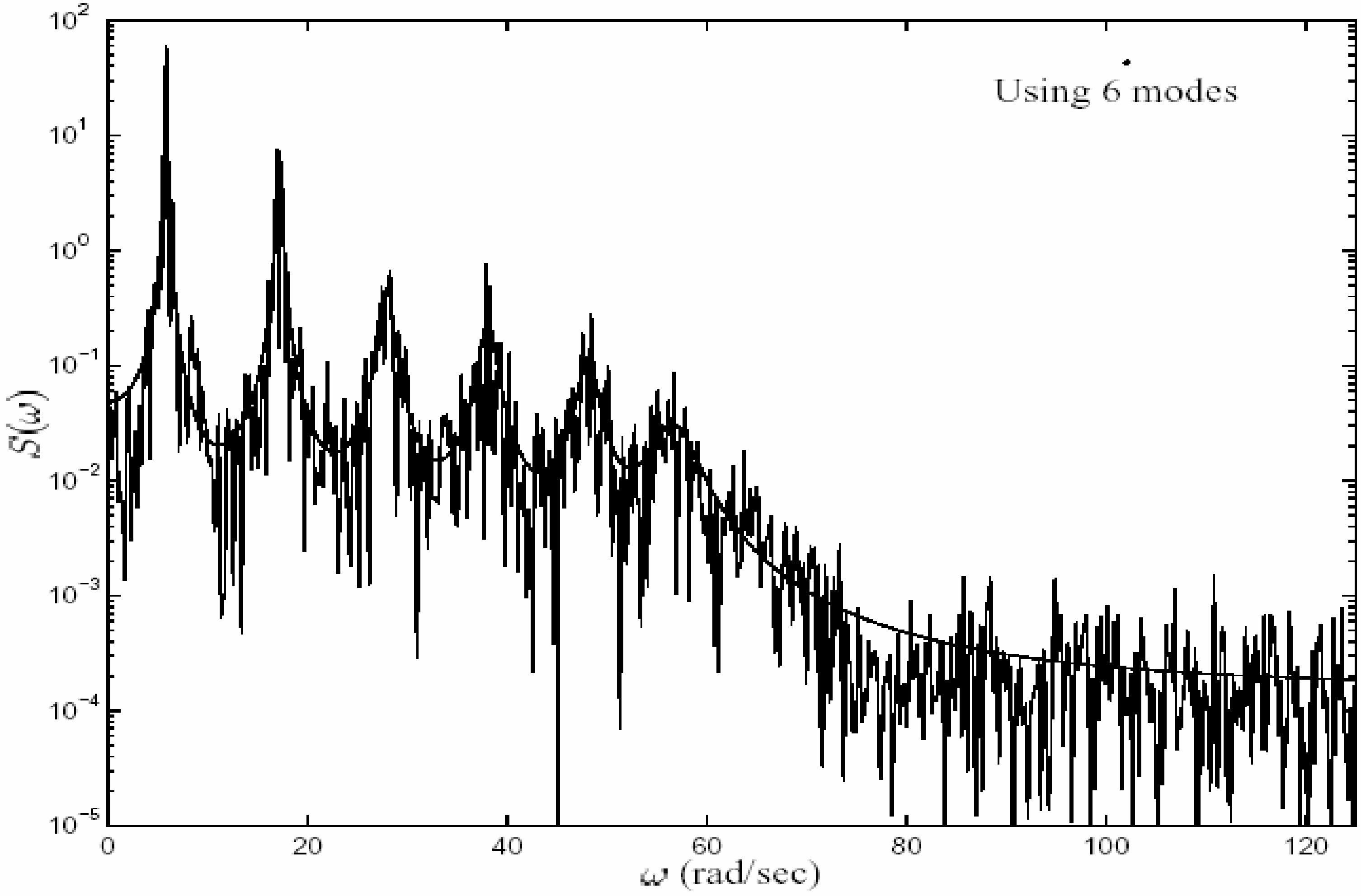
Number of modes	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
Exact	5.789	17.24	28.30	38.73	48.30	56.78	64.00	69.79
1	6.946	—	—	—	—	—	—	—
2	5.799	20.68	—	—	—	—	—	—
3	5.814	17.16	33.96	—	—	—	—	—
4	5.842	17.18	27.94	43.82	—	—	—	—
5	5.848	17.19	27.97	38.06	50.58	—	—	—
6	5.849	17.19	27.97	38.09	48.10	56.72	—	—
7	5.849	17.19	27.97	38.09	48.13	56.34	64.18	—
8	5.849	17.19	27.97	38.09	48.13	56.34	64.18	69.41

Example 2: Evidence for Model Classes

Number of modes	Log likelihood	Log Ockham factor	Log evidence
1	1894	-43.7	1850
2	2251	-56.4	2195
3	2511	-68.9	2442
4	2619	-69.2	2550
5	2682	-75.9	2606
6	2714	-91.2	2623 (& BIC)
7	2723	-109	2614
8	2723	-121	2602 (AIC)

Probability of model class with 6 modes completely dominates, e.g. next class has probability 0.0002

Example 2: Frequency Response Fit for Most Probable 6-mode Model



Concluding Remarks

- The Bayesian probabilistic approach for model class selection is generally applicable; illustrated here for linear & non-linear dynamical systems with input-output or output-only dynamic data
- The most plausible class of models is the one with the maximum probability (or evidence) based on the data
- Rather than taking most probable, can use all classes by model class averaging (Total Prob.)