Model Class Selection

- **Given**: Data $D$ from system and set $M$ of candidate model classes
  
  \[ M = \{M_1, M_2, \ldots, M_J\} \]

  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
Most Plausible Model Class Based on Data $D$

Maximize:

$$P(M_j | D, M)$$ over all $j$

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

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

Evaluation of Model Class Probability

Bayes Theorem:

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

where denominator is chosen to normalize

\[
P(M_j \mid D, M) \quad \text{over} \quad j = 1, \ldots, J
\]
Total Probability Theorem gives evidence:

\[ p(D \mid M_j) = \int_{\Theta_j} p(D \mid \theta_j, M_j) p(\theta_j \mid M_j) d\theta_j \]

Can use asymptotic expansion about MPV

\[ p(D \mid M_j) \approx (2\pi)^{\frac{N_j}{2}} \frac{p(D \mid \hat{\theta}_j, M_j) p(\hat{\theta}_j \mid M_j)}{\sqrt{\det H_j(\hat{\theta}_j)}} \]

\[ H_j(\theta_j) = -\nabla\nabla \ln p(D \mid \theta_j, M_j) p(\theta_j \mid M_j) \]
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:

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

= log likelihood + log Ockham factor

= Data fit + Bias against parameterization

- Gives a quantitative Principle of Parsimony
Bias Against Parameterization

- **Log Ockham Factor** $\beta_j$ for $M_j$:

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

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

  where $\rho_{j,i}^2, \sigma_{j,i}^2$ are the prior and principal posterior variances for $\theta_{j,i}$ and $\overline{\theta}_{j,i}, \hat{\theta}_{j,i}$ are the prior and posterior most probable values of $\theta_{j,i}$

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

  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):

\[
\text{Log evidence} = [\text{Data fit of optimal model}] - [\text{Information gain about } \theta_j \text{ in } D]
\]

Recently generalized this result to any model class
Comparison with AIC and BIC

- Bayesian model class selection criterion
  Maximize $\ln P(M_j \mid D, M_j)$ w.r.t. $M_j$, or equivalently (from asymptotic result):
  $\log$ evidence $= \log$ likelihood $+ \log$ Ockham factor
  i.e. $\ln p(D \mid M_j) = \ln p(D \mid \hat{\theta}_j, M_j) + \beta_j$

- Akaike (1974)
  Maximize: $\text{AIC} = \ln p(D \mid \hat{\theta}_j, M_j) - N_j$

- Akaike (1976), Schwarz (1978)
  Maximize: $\text{BIC} = \ln p(D \mid \hat{\theta}_j, M_j) - \frac{N_j}{2} \ln N$
  (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 \( \theta_j \) + Prediction error

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

- Details for dynamical models with input-output measurements:

- Details for output-only measurements:
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 \( \eta \) modeled as zero-mean Gaussian discrete white noise with variance \( \sigma_\eta^2 \)
  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** (\(M_1\) - 3 parameters)
  Linear oscillators with damping coefficient \(c > 0\), stiffness \(k_1 > 0\) and prediction-error variance \(\sigma_\eta^2\)

- **Model Class 2** (\(M_2\) - 3 parameters)
  Elasto-plastic oscillators (i.e. \(k_2 = 0\)) with stiffness \(k_1 > 0\), yield displacement \(x_y\) and prediction-error variance \(\sigma_\eta^2\)

- 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

<table>
<thead>
<tr>
<th>Number of modes</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>$\omega_3$</th>
<th>$\omega_4$</th>
<th>$\omega_5$</th>
<th>$\omega_6$</th>
<th>$\omega_7$</th>
<th>$\omega_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>5.789</td>
<td>17.24</td>
<td>28.30</td>
<td>38.73</td>
<td>48.30</td>
<td>56.78</td>
<td>64.00</td>
<td>69.79</td>
</tr>
<tr>
<td>1</td>
<td>6.946</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.799</td>
<td>20.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.814</td>
<td>17.16</td>
<td>33.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5.842</td>
<td>17.18</td>
<td>27.94</td>
<td>43.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.848</td>
<td>17.19</td>
<td>27.97</td>
<td>38.06</td>
<td>50.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5.849</td>
<td>17.19</td>
<td>27.97</td>
<td>38.09</td>
<td>48.10</td>
<td>56.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5.849</td>
<td>17.19</td>
<td>27.97</td>
<td>38.09</td>
<td>48.13</td>
<td>56.34</td>
<td>64.18</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.849</td>
<td>17.19</td>
<td>27.97</td>
<td>38.09</td>
<td>48.13</td>
<td>56.34</td>
<td>64.18</td>
<td>69.41</td>
</tr>
</tbody>
</table>
### Example 2: Evidence for Model Classes

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

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.)