AIC and cross-validation (cont.)

- Let $\hat{\theta}_\alpha$ and $\hat{\theta}_\alpha^{[-1]}$ be the MLE's of $\theta_\alpha$ for $D$ and $D^{[-1]}$.
- Under regularity conditions, $CV(\alpha)$ is asymptotically log $L(\hat{\theta}_\alpha) - p_\alpha$ with $p_\alpha = \text{dim}(\Theta_\alpha)$.
- Asymptotic equivalence of choice of model by CV and AIC
- For two models $\alpha_1$ and $\alpha_2$, both containing a true model with $\Theta_\alpha_1 \subset \Theta_\alpha_2$,
  
  $$2\left\{ \log L(\hat{\theta}_{\alpha_2}) - \log L(\hat{\theta}_{\alpha_1}) \right\}$$
  
  is asymptotically $\chi^2_{df}$ with $df = p_{\alpha_2} - p_{\alpha_1}$.
- $CV(\alpha_2) - CV(\alpha_1)$ is asymptotically $\chi^2_{df}/2 - df$
- The simpler model would be favored by $CV(\alpha)$.

Comparison with AIC

$$BIC = -2 \log L(\hat{\theta}_k) + (\log n)k$$

- The factor 2 in AIC and $C_p$ is replaced by log $n$.
- Sample size dependent penalty
- When $n > e^2 \approx 7.4$ (log $n > 2$), BIC penalizes complex models more heavily.
- A large sample version of Bayes procedures

Bayesian Information Criterion

Schwarz, G. (1978), *Estimating the dimension of a model*

- Data: $D = \{z_i = (x_i, y_i) \mid i = 1, \ldots, n\}$
- Models $M_k$ are indexed by $k$:
  
  $$\{f(z|\theta_k, M_k), \theta_k \in \Theta_k \mid k = 1, \ldots, p\}$$
  
  where $\Theta_k$ is a $k$-dimensional space.
- A Bayesian solution to selection of models of different dimensions
- Choose $M_k$ minimizing
  
  $$BIC = -2 \log L(\hat{\theta}_k) + (\log n)k$$
  
  where $\hat{\theta}_k$: MLE of $\theta_k$ under $M_k$ and $L(\theta_k)$: likelihood of $\theta_k$
- Also known as SIC (Schwarz Information Criterion)
  
  Select $M_k$ maximizing
  
  $$SIC = \log L(\hat{\theta}_k) - \frac{k}{2}(\log n)$$

Bayesian approach to model selection

- Candidate models $\{M_k, k = 1, \ldots, p\}$
- Select the model that is a posteriori most probable, the one with the maximum posterior probability $P(M_k|D)$.
- A prior distribution over models $P(M_k)$
- A prior distribution $P(\theta_k|M_k)$ for the parameters $\theta_k$
- A probability model for the data given $M_k$ and $\theta_k$
  
  $$L(\theta_k) = P(D|\theta_k, M_k) = \prod_{i=1}^n f(z_i|\theta_k, M_k)$$
Posterior probability of a model

- For a given model $\mathcal{M}_k$,
  
  \[
  P(\mathcal{M}_k|\mathcal{D}) \propto P(\mathcal{M}_k, \mathcal{D}) = P(\mathcal{M}_k)P(\mathcal{D}|\mathcal{M}_k)
  \]

- (Marginal) likelihood of $\mathcal{M}_k$
  
  \[
  P(\mathcal{D}|\mathcal{M}_k) = \int P(\mathcal{D}, \theta_k|\mathcal{M}_k)d\theta_k
  = \int P(\mathcal{D} | \theta_k, \mathcal{M}_k)P(\theta_k|\mathcal{M}_k)d\theta_k
  \]

Model comparison

- For two models $\mathcal{M}_k$ and $\mathcal{M}_\ell$, consider the posterior odds
  
  \[
  \frac{P(\mathcal{M}_k|\mathcal{D})}{P(\mathcal{M}_\ell|\mathcal{D})} = \frac{P(\mathcal{M}_k)P(\mathcal{D}|\mathcal{M}_k)}{P(\mathcal{M}_\ell)P(\mathcal{D}|\mathcal{M}_\ell)}
  \]

- If the odds $> 1$, then choose $\mathcal{M}_k$, otherwise choose $\mathcal{M}_\ell$.

- When the prior over models is uniform, the posterior odds are determined by the Bayes factor
  
  \[
  \frac{P(\mathcal{D}|\mathcal{M}_k)}{P(\mathcal{D}|\mathcal{M}_\ell)}.
  \]

- The Bayes factor provides a scale of evidence in favor of one model versus another.
  
  e.g. $BF(k, \ell) > 10$: strong evidence for $\mathcal{M}_k$
  
  $3 < BF(k, \ell) < 10$: moderate evidence for $\mathcal{M}_k$

Marginal likelihood of a model

- Recall
  
  \[
  P(\mathcal{D}|\mathcal{M}_k) = \int P(\mathcal{D}|\theta_k, \mathcal{M}_k)P(\theta_k|\mathcal{M}_k)d\theta_k
  \]

- Analytically approximate $P(\mathcal{D}|\mathcal{M}_k)$

  - Suppose that $P(\theta_k|\mathcal{M}_k)d\theta_k$ is Lebesgue measure on $\Theta_k$
  
  - $P(\mathcal{D}|\theta_k, \mathcal{M}_k) = \exp(\log P(\mathcal{D}|\theta_k, \mathcal{M}_k))$

Laplace approximation

- Let $\hat{\theta}_k$ be the MLE of $\theta_k$ under $\mathcal{M}_k$
  
  - An analytical approximation to integral by fitting a multivariate normal density in $\theta_k$ at $\hat{\theta}_k$

  \[
  \log L(\theta_k) \approx \log L(\hat{\theta}_k) + \frac{\partial \log L(\theta_k)}{\partial \theta_k}^\top (\hat{\theta}_k - \theta_k)
  + \frac{1}{2} (\hat{\theta}_k - \theta_k)^\top \frac{\partial^2 \log L(\theta_k)}{\partial \theta_k^2} (\hat{\theta}_k - \theta_k)
  \]

- $\log P(\mathcal{D}|\theta_k, \mathcal{M}_k) \approx \log P(\mathcal{D}|\hat{\theta}_k, \mathcal{M}_k) - \frac{\lambda}{2} \| \hat{\theta}_k - \theta_k \|^2$
  
  for some positive constant $\lambda$
For $A_n = \log L(\hat{\theta}_k) = \log P(D|\hat{\theta}_k, \mathcal{M}_k)$, 

$$P(D|\mathcal{M}_k) \approx \int \exp(A_n - \frac{n\lambda}{2}\|\hat{\theta}_k - \theta_k\|^2) d\theta_k \approx \exp(A_n)(\sqrt{2\pi/n\lambda})^k$$

Then 

$$\log P(D|\mathcal{M}_k) \approx A_n - \frac{k}{2}\log(n\lambda/2\pi)$$ 

$$= A_n - \frac{k}{2}\log n - \frac{k}{2}\log(\lambda/2\pi)$$ 

$$= A_n - \frac{k}{2}\log n + O(1)$$ 

$$= \log L(\hat{\theta}_k) - \frac{k}{2}\log n + O(1)$$

**BIC as a large sample criterion**

For large sample $n$, the model with the maximum $P(\mathcal{M}_k|D)$ maximizes $\log L(\hat{\theta}_k) - \frac{k}{2}\log n$ or equivalently minimizes 

$$BIC = -2\log L(\hat{\theta}_k) + (\log n)k$$

Leading terms of the asymptotic expansion of $P(\mathcal{M}_k|D)$ do not depend on the prior distribution over models. 

**BIC is asymptotically valid beyond the Bayesian context.**

**Remarks on BIC**

- Both AIC and BIC provide a mathematical formulation for the parsimony principle in modeling.
- BIC leans towards smaller models.
- For large samples, AIC and BIC differ markedly.
- BIC is known to be asymptotically consistent in various situations. When a given class of models includes the true model, BIC selects the correct model almost surely as $n \to \infty$.
- For finite samples, BIC often selects models that are too simple.