A Tutorial on Model Selection

Enes Makalic and Daniel F. Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

Work in Progress, 2011

Outline

Outline

Problem Description (1)

- We have
  - Data points $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$, $y_i \in \mathbb{R}$, $(i = 1, 2, \ldots, n)$
  - Probabilistic source p^*

$$\mathbf{y} \sim p^*$$

- Task: learn a good approximation to p^* using data $\mathbf{y}$

Problem Description (1)

- We have
  - Data points $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$, $y_i \in \mathbb{R}$, $(i = 1, 2, \ldots, n)$
  - Probabilistic source p^*

$$\mathbf{y} \sim p^*$$

- Task: learn a good approximation to p^* using data $\mathbf{y}$

Problem Description (2)

- Determine a suitable model
  - Model structure
  - Model parameters
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - Distribution? (e.g., Weibull, Log-Normal, Gamma)
  - Parameters? (e.g., scale and shape parameters)
- Task: learn a good approximation to p^* using data $\mathbf{y}$
  - Not tractable, in general!

Problem Description (2)

- Determine a suitable model
  - Model structure
  - Model parameters
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - Distribution? (e.g., Weibull, Log-Normal, Gamma)
  - Parameters? (e.g., scale and shape parameters)
- Task: learn a good approximation to p^* using data $\mathbf{y}$
  - Not tractable, in general!

Problem Description (2)

- Determine a suitable model
  - Model structure
  - Model parameters
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - Distribution? (e.g., Weibull, Log-Normal, Gamma)
  - Parameters? (e.g., scale and shape parameters)
- Task: learn a good approximation to p^* using data $\mathbf{y}$
  - Not tractable, in general!

Problem Description (2)

- Determine a suitable model
  - Model structure
  - Model parameters
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - Distribution? (e.g., Weibull, Log-Normal, Gamma)
  - Parameters? (e.g., scale and shape parameters)
- Task: learn a good approximation to p^* using data $\mathbf{y}$
  - Not tractable, in general!

Problem Description (3)

- Assumptions regarding unknown source p^*
  - Can be approximated by a distribution from $p(\mathbf{y}|\boldsymbol{\theta};\gamma)$
    - Model $\gamma \in \Gamma$
    - Model structure $\Gamma \subset \mathbb{N}$
    - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k_{|\gamma|}})' \in \Theta_\gamma \subset \mathbb{R}^{k_{|\gamma|}}$
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - $\Gamma = \{$Weibull, Log-normal, Gamma$\}$
  - If $\gamma = \{$Log-normal$\}$, $\boldsymbol{\theta} = (\mu, \sigma^2)'$, or
  - If $\gamma = \{$Weibull$\}$, $\boldsymbol{\theta} = (k, \lambda)'$

Problem Description (3)

- Assumptions regarding unknown source p^*
  - Can be approximated by a distribution from $p(\mathbf{y}|\boldsymbol{\theta};\gamma)$
    - Model $\gamma \in \Gamma$
    - Model structure $\Gamma \subset \mathbb{N}$
    - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k_{|\gamma|}})' \in \Theta_\gamma \subset \mathbb{R}^{k_{|\gamma|}}$
- Example: $\mathbf{y} = (10.26, 7.95, 4.59, 7.00, 10.17, 11.78)'$
  - $\Gamma = \{$Weibull, Log-normal, Gamma$\}$
  - If $\gamma = \{$Log-normal$\}$, $\boldsymbol{\theta} = (\mu, \sigma^2)'$, or
  - If $\gamma = \{$Weibull$\}$, $\boldsymbol{\theta} = (k, \lambda)'$

Problem Description (4)

- Parameter estimation: method of maximum likelihood
- Choose $\boldsymbol{\theta}$ such that probability of observed $\mathbf{y}$ is maximised

$$\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_\gamma} p(\mathbf{y}|\boldsymbol{\theta};\gamma)$$

- Many attractive statistical properties
  - May not be used for model selection!
- Talk will concentrate on inference of model structure $\gamma \in \Gamma$
  - Running example: linear regression model

Linear Regression Model (1)

- Linear regression model for explaining data $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ is the full design matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ are the unknown parameter coefficients
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ are i.i.d. Gaussian variates
- Only a subset of covariates $\mathbf{X}$ is associated with $\mathbf{y}$
- Task: determine which covariates, if any, are associated with $\mathbf{y}$

Linear Regression Model (1)

- Linear regression model for explaining data $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ is the full design matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ are the unknown parameter coefficients
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ are i.i.d. Gaussian variates
- Only a subset of covariates $\mathbf{X}$ is associated with $\mathbf{y}$
- Task: determine which covariates, if any, are associated with $\mathbf{y}$

Linear Regression Model (2)

- Let $\gamma \subset \{1, 2, \ldots, q\}$ denote which covariates are in design submatrix $\mathbf{X}_\gamma$
- Linear model indexed by $\gamma \in \Gamma$

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - Set of all candidate subsets Γ
  - $\mathbf{X} = (\mathbf{x}_{\gamma_1}, \mathbf{x}_{\gamma_2}, \ldots, \mathbf{x}_{\gamma_{|\gamma|}})$ is the design sub-matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{|\gamma|})'$ is the unknown parameter vector
  - Total number of unknown parameters is $k = |\gamma| + 1$
- Example
  - $q = 10$, $\gamma = \{2, 3, 6, 10\}$, $\mathbf{X}_\gamma = (\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_{10})$
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' \in \boldsymbol{\Theta}_\gamma$
  - $k = 5$

Linear Regression Model (2)

- Let $\gamma \subset \{1, 2, \ldots, q\}$ denote which covariates are in design submatrix $\mathbf{X}_\gamma$
- Linear model indexed by $\gamma \in \Gamma$

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - Set of all candidate subsets Γ
  - $\mathbf{X} = (\mathbf{x}_{\gamma_1}, \mathbf{x}_{\gamma_2}, \ldots, \mathbf{x}_{\gamma_{|\gamma|}})$ is the design sub-matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{|\gamma|})'$ is the unknown parameter vector
  - Total number of unknown parameters is $k = |\gamma| + 1$
- Example
  - $q = 10,\ \gamma = \{2, 3, 6, 10\},\ \mathbf{X}_\gamma = (\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_{10})$
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' \in \boldsymbol{\Theta}_\gamma$
  - $k = 5$

Linear Regression Model (3)

- The set Γ may be of nested or non-nested structure
- Nested structure
  - Polynomial regression with $(q = 3)$ covariates
  - Constant term $\mathbf{x}_1$, linear term $\mathbf{x}_2$ and quadratic term $\mathbf{x}_3$

$$\Gamma = \{\emptyset, \{1\}, \{1,2\}, \{1,2,3\}\}$$

- Non-nested structure
  - All-subsets regression problem

$$\Gamma = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

Linear Regression Model (4)

- Maximum likelihood estimates

$$\begin{aligned}\hat{\beta}(\mathbf{y};\gamma) &= (\mathbf{X}'_\gamma \mathbf{X}_\gamma)^{-1}\mathbf{X}'_\gamma \mathbf{y} \\ \hat{\tau}(\mathbf{y};\gamma) &= \frac{1}{n}(y - \mathbf{X}_\gamma \hat{\beta}(\mathbf{y};\gamma))'(y - \mathbf{X}_\gamma \hat{\beta}(\mathbf{y};\gamma))\end{aligned}$$

- Negative log-likelihood evaluated at maximum likelihood estimates

$$-\log p(\mathbf{y}|\mathbf{X}_\gamma, \hat{\beta}, \hat{\tau};\gamma) = \frac{n}{2}\log 2\pi + \frac{n}{2}\log \hat{\tau}(\mathbf{y};\gamma) + \frac{n}{2}$$

Outline

Introduction (1)

- How close is *fitted* model $p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y});\gamma)$ to unknown $p^*(\mathbf{y})$?
- Require measure of distance between distributions
  - Kullback-Leibler (KL) divergence

$$\begin{aligned}\Delta_n(p^*, p_{\boldsymbol{\theta}_\gamma}) &= \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)} dx \\ &= \mathrm{E}_{\mathbf{x}\sim p^*}\left[\log \frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)}\right] \\ &= \mathrm{E}_{\mathbf{x}\sim p^*}\left[\log p^*(\mathbf{x})\right] + \mathrm{E}_{\mathbf{x}\sim p^*}\left[-\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right]\end{aligned}$$

Introduction (1)

- How close is *fitted* model $p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y});\gamma)$ to unknown $p^*(\mathbf{y})$?
- Require measure of distance between distributions
    - Kullback-Leibler (KL) divergence

$$\begin{aligned}
\Delta_n(p^*, p_{\boldsymbol{\theta}_\gamma}) &= \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)} d\mathbf{x} \\
&= \mathrm{E}_{\mathbf{x}\sim p^*}\left[\log \frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)}\right] \\
&= \mathrm{E}_{\mathbf{x}\sim p^*}\left[\log p^*(\mathbf{x})\right] + \mathrm{E}_{\mathbf{x}\sim p^*}\left[-\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right]
\end{aligned}$$

Introduction (2)

- Examples
  - KL divergence between $X_1 \sim \mathrm{N}(\mu_1, \tau_1)$ and $X_2 \sim \mathrm{N}(\mu_2, \tau_2)$

$$\Delta_1(X_1, X_2) = \frac{(\mu_1 - \mu_2)^2}{2\tau_2} + \frac{1}{2}\left(\frac{\tau_1}{\tau_2} - 1 - \log\frac{\tau_1}{\tau_2}\right)$$

  - KL divergence between $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$

$$\Delta_1(X_1, X_2) = \frac{\lambda_2}{\lambda_1} - \log\frac{\lambda_2}{\lambda_1} - 1$$

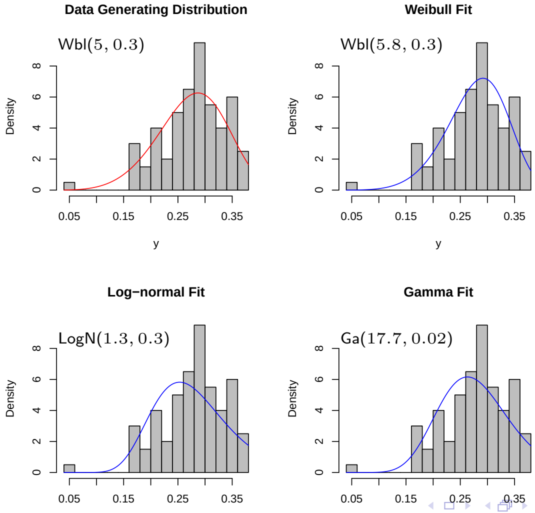
Introduction (3)

- Ideally, want to choose model $\gamma \in \Gamma$ is

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \left\{ \Delta_n(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma}) \right\}$$

- Example: $\mathbf{y} \sim \text{Weibull}(k = 5, \lambda = 0.3)$, $(n = 100)$
  - $\Gamma = \{$Weibull, Log-normal, Gamma$\}$

Introduction (4)



Introduction (5)

- Ranking based on KL divergence $\Delta_n(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma})$
  - Requires knowledge of p^*.
  - Not possible!

Takeuchi Information Criterion (TIC) (1)

- Takeuchi noted $-\log p(\mathbf{y}|\hat{\boldsymbol{\theta}};\gamma)$ is a downwardly biased estimator of

$$\mathrm{E}_{\mathbf{x}\sim p^*}\left[-\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right]$$

- Exact bias adjustment generally not computable
  - Asymptotic adjustment possible

Takeuchi Information Criterion (TIC) (2)

- Model selection criterion

$$\mathrm{TIC}(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2\,\mathrm{tr}\left(\boldsymbol{\Omega}^{-1}(\gamma;\mathbf{y})\boldsymbol{\Sigma}(\gamma;\mathbf{y})\right)$$

where

$$\begin{aligned}
\boldsymbol{\Omega}(\mathbf{y};\gamma) &= -\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma} \\
\boldsymbol{\Sigma}(\mathbf{y};\gamma) &= \sum_{i=1}^{n}\left(\frac{\partial \log p(y_i|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma}\right)\left(\frac{\partial \log p(y_i|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma}\right)'
\end{aligned}$$

Takeuchi Information Criterion (TIC) (3)

- To use TIC for model selection, choose

$$\hat{\gamma}_{\text{TIC}}(\mathbf{y}) = \arg\min_{\gamma \in \Gamma} \left\{ \text{TIC}(\gamma; \mathbf{y}) \right\}.$$

- Model with smallest TIC score
  - Closest to p^* in KL divergence
- TIC is asymptotically unbiased estimate of KL divergence!

Akaike Information Criterion (AIC) (3)

- TIC can be simplified
  - Assume p^* is contained in model γ
- Akaike Information Criterion (AIC)

$$\mathrm{AIC}(\gamma; \mathbf{y}) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2k$$

where k is dimensionality of $\boldsymbol{\theta} \in \Theta_\gamma$

Akaike Information Criterion (AIC) (4)

- AIC is an asymptotically unbiased estimator of the KL divergence
- Excellent estimate when
  - Sample size n is large
  - Number of parameters k is small

Small Sample Correction to AIC (1)

- AIC *should not* be used if
  - Sample size n is small, or
  - Number of parameter k is large relative to n
- A small-sample correction for AIC

$$\text{AIC}_c(\gamma; \mathbf{y}) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + \frac{2kn}{n-k-1}$$

Small Sample Correction to AIC (1)

- AIC *should not* be used if
  - Sample size n is small, or
  - Number of parameter k is large relative to n
- A small-sample correction for AIC

$$\mathrm{AIC}_c(\gamma; \mathbf{y}) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + \frac{2kn}{n-k-1}$$

Small Sample Correction to AIC (2)

- AIC_c derived for linear regression setting
- As $n \to \infty$, AIC_c is equivalent to regular AIC
- Empirical evidence
  - AIC_c performs better at model selection than AIC
  - Largely true, irrespective of the problem

The Kullback Information Criterion (KIC) (1)

- KL divergence is an asymmetric measure
- Symmetric KL can be used to derive new criteria

$$J_n(p^*, p_{\boldsymbol{\theta}_\gamma}) = \mathrm{E}_{\mathbf{x}\sim p^*}\left[\log \frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)}\right] + \mathrm{E}_{\mathbf{x}\sim \boldsymbol{\theta}_\gamma}\left[\log \frac{p(\mathbf{x}|\boldsymbol{\theta};\gamma)}{p^*(\mathbf{x})}\right]$$

The Kullback Information Criterion (KIC) (1)

- The symmetric Kullback information criterion (KIC)

$$\mathrm{KIC}(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 3k$$

- Small sample correction (KIC_c)

$$\mathrm{KIC}_c(\gamma;\mathbf{y}) = \mathrm{AIC}_c(\gamma;\mathbf{y}) - n\psi\left(\frac{n-k+1}{2}\right) + n\log\frac{n}{2}$$

Example: Linear Regression

- Total number of parameters is $k = |\gamma| + 1$
- Model selection criteria

$$\begin{aligned}
\mathrm{AIC}(\gamma;\mathbf{y}) &= n\log\left(2\pi\hat{\tau}(\mathbf{y};\gamma)\right) + n + 2k \\
\mathrm{AIC}_c(\gamma;\mathbf{y}) &= n\log\left(2\pi\hat{\tau}(\mathbf{y};\gamma)\right) + n + \frac{2kn}{n-k-1} \\
\mathrm{KIC}(\gamma;\mathbf{y}) &= n\log\left(2\pi\hat{\tau}(\mathbf{y};\gamma)\right) + n + 3k \\
\mathrm{KIC}_c(\gamma;\mathbf{y}) &= \mathrm{AIC}_c(\gamma;\mathbf{y}) - n\psi\left(\frac{n-k+1}{2}\right)
\end{aligned}$$

Summary

- All examined distance criteria derived for nested Γ
- Consistency
  - None of the examined distance based criteria are consistent!
  - Avoid using if the aim is to do model selection
- Efficiency
  - AIC and KIC, and corrected variants, are asymptotically efficient
  - Good prediction performance

Summary

- All examined distance criteria derived for nested Γ
- Consistency
  - None of the examined distance based criteria are consistent!
  - Avoid using if the aim is to do model selection
- Efficiency
  - AIC and KIC, and corrected variants, are asymptotically efficient
  - Good prediction performance

Summary

- All examined distance criteria derived for nested Γ
- Consistency
  - None of the examined distance based criteria are consistent!
  - Avoid using if the aim is to do model selection
- Efficiency
  - AIC and KIC, and corrected variants, are asymptotically efficient
  - Good prediction performance

Outline

Basic Idea (1)

- Uncertainty about models and parameters is defined in terms of probability
- Need a *prior* distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)$ over parameters $\boldsymbol{\theta} \in \Theta_\gamma$
    - Quantifies uncertainty about $\boldsymbol{\theta} \in \Theta_\gamma$
    - Subjective priors, objective priors

Basic Idea (1)

- Uncertainty about models and parameters is defined in terms of probability
- Need a *prior* distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)$ over parameters $\boldsymbol{\theta} \in \Theta_\gamma$
  - Quantifies uncertainty about $\boldsymbol{\theta} \in \Theta_\gamma$
  - Subjective priors, objective priors

Basic Idea (1)

- Uncertainty about models and parameters is defined in terms of probability
- Need a *prior* distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \gamma)$ over parameters $\boldsymbol{\theta} \in \Theta_\gamma$
  - Quantifies uncertainty about $\boldsymbol{\theta} \in \Theta_\gamma$
  - Subjective priors, objective priors

Basic Idea (2)

- Parameter estimation
  - Posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{y};\gamma) = \frac{p(\mathbf{y}|\boldsymbol{\theta};\gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)}{m(\mathbf{y};\gamma)}$$

$$m(\mathbf{y};\gamma) = \int_{\Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta};\gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)d\boldsymbol{\theta}$$

  - Marginal distribution $m(\mathbf{y};\gamma)$
  - Posterior mean, posterior mode, posterior maximum?

Basic Idea (3)

- Example: Parameter estimation
  - Likelihood, $y_i \sim \mathrm{Exp}(\lambda), (i = 1, 2, \ldots, n)$
  - Prior density, $\lambda \sim \mathrm{Ga}(\alpha, \beta)$
  - Posterior density, $\lambda|\mathbf{y} \sim \mathrm{Ga}(\alpha + n, \beta + \sum_{i=1}^{n} y_i)$

Basic Idea (4)

- How do we choose γ from $p(\boldsymbol{\theta}|\mathbf{y};\gamma)$?
- Posterior distribution over models
  - Define a prior density $\pi_\gamma(\gamma)$ over models $\gamma \in \Gamma$

$$p(\gamma|\mathbf{y}) = \frac{m(\mathbf{y};\gamma)\pi_\gamma(\gamma)}{\sum_{\gamma\in\Gamma} m(\mathbf{y};\gamma)\pi_\gamma(\gamma)}$$

- Model selection
  - Choose the model that maximises

$$\hat{\gamma}(\mathbf{y}) = \arg\max_{\gamma\in\Gamma} \left\{ \frac{m(\mathbf{y};\gamma)\pi_\gamma(\gamma)}{\sum_{\gamma\in\Gamma} m(\mathbf{y};\gamma)\pi_\gamma(\gamma)} \right\}$$

Basic Idea (5)

- Posterior odds in favour of model γ_1 over γ_0

$$\mathrm{BF}(\gamma_1, \gamma_0) = \frac{m(\mathbf{y}; \gamma_1)\pi_\gamma(\gamma_1)}{m(\mathbf{y}; \gamma_0)\pi_\gamma(\gamma_0)}$$

- Computational complexity regarding $m(\mathbf{y}; \gamma)$
    - Difficult to compute
    - No closed-form solution in general!

Basic Idea (5)

- Posterior odds in favour of model γ_1 over γ_0

$$\mathrm{BF}(\gamma_1, \gamma_0) = \frac{m(\mathbf{y}; \gamma_1)\pi_\gamma(\gamma_1)}{m(\mathbf{y}; \gamma_0)\pi_\gamma(\gamma_0)}$$

- Computational complexity regarding $m(\mathbf{y}; \gamma)$
  - Difficult to compute
  - No closed-form solution in general!

Bayesian Information Criterion (1)

- BIC approach to model selection
  - Assume certain regularity conditions, e.g., $n \to \infty$ and

$$\mathbf{J}_1(\boldsymbol{\theta}; \gamma) = \lim_{n \to \infty} \frac{\mathbf{J}_n(\boldsymbol{\theta}; \gamma)}{n}$$

  - Use Laplace approximation to the integral in $m(\mathbf{y}; \gamma)$

$$-\log \int_{\Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta}; \gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \gamma) d\boldsymbol{\theta} = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}; \gamma) + \frac{k}{2} \log n + O(1)$$

Bayesian Information Criterion (2)

- Model selection with BIC
  - Compute criterion for all $\gamma \in \Gamma$

$$\mathrm{BIC}(\gamma;\mathbf{y}) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}};\gamma) + \frac{k}{2}\log n - \log \pi_\gamma(\gamma)$$

  - Choose model with smallest BIC score

$$\hat{\gamma}(\mathbf{y}) = \arg\min_{\gamma\in\Gamma}\{\mathrm{BIC}(\gamma;\mathbf{y})\}$$

Properties of BIC

- Derived for both nested and non-nested model selection
- Independent of prior density $\pi_{\theta}(\theta; \gamma)$
- BIC is asymptotically consistent!
  - Under certain assumptions
  - Important for model selection
- Strong empirical performance
  - If generating process has a small number of strong effects

Properties of BIC

- Derived for both nested and non-nested model selection
- Independent of prior density $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)$
- BIC is asymptotically consistent!
  - Under certain assumptions
  - Important for model selection
- Strong empirical performance
  - If generating process has a small number of strong effects

Properties of BIC

- Derived for both nested and non-nested model selection
- Independent of prior density $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \gamma)$
- BIC is asymptotically consistent!
  - Under certain assumptions
  - Important for model selection
- Strong empirical performance
  - If generating process has a small number of strong effects

Properties of BIC

- Derived for both nested and non-nested model selection
- Independent of prior density $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta};\gamma)$
- BIC is asymptotically consistent!
  - Under certain assumptions
  - Important for model selection
- Strong empirical performance
  - If generating process has a small number of strong effects

Example: Linear Regression

- Total number of parameters is $k = |\gamma| + 1$
- Bayesian Information Criterion (BIC)

$$\mathrm{BIC}(\gamma; \mathbf{y}) = n \log\left(2\pi\hat{\tau}(\mathbf{y};\gamma)\right) + n + k \log n$$

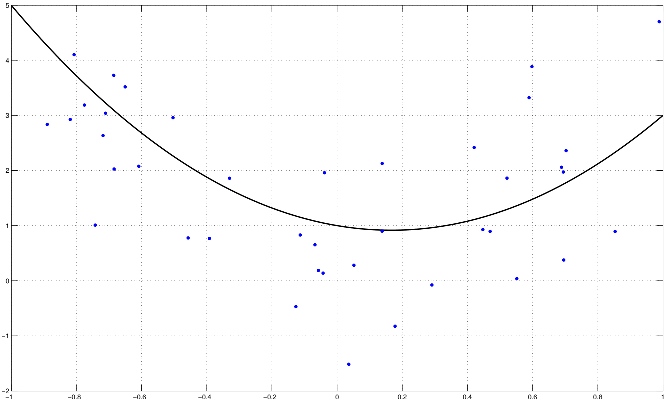
Outline

Simulation (1)

- Simulation procedure
  - Generate $x_i \in (-1, 1)$, $(i = 1, 2, \ldots, n)$
  - Create matrix of covariates $\mathbf{X} = (\mathbf{x}^0, \mathbf{x}^2, \ldots, \mathbf{x}^{15})$
  - Generate targets $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathrm{N}_n(0, \tau\mathbf{I}_n)$
  - Ask each criterion to nominate the best model given $(\mathbf{X}, \mathbf{y})$
  - Repeat each test 10^4 times
  - Performance metrics
    - Squared prediction error (SPE)
    - Polynomial order

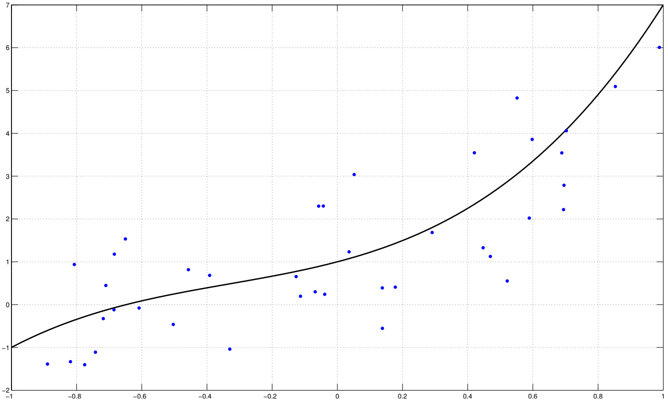
Simulation (2)

Test Function (1): $y = 1 - x + 3x^2$



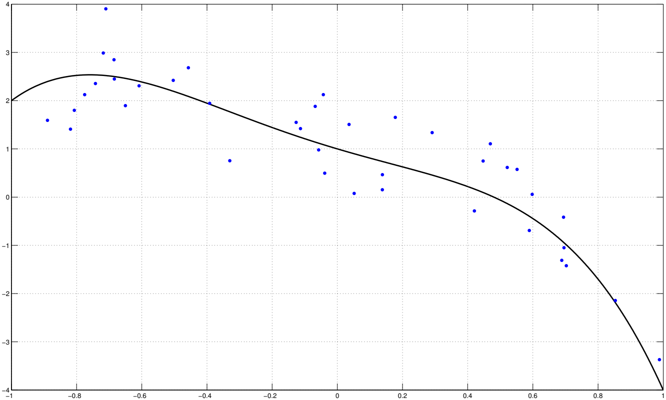
Simulation (3)

Test Function (2): $y = 1 + 2x + 2x^2 + 2x^3$



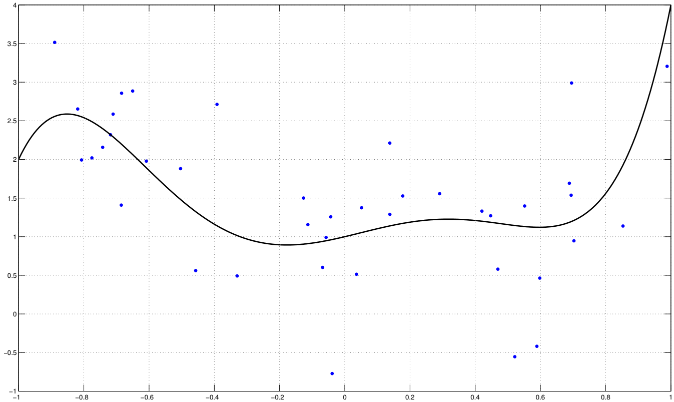
Simulation (4)

Test Function (3): $y = 1 - 2x + x^2 - x^3 - 3x^4$



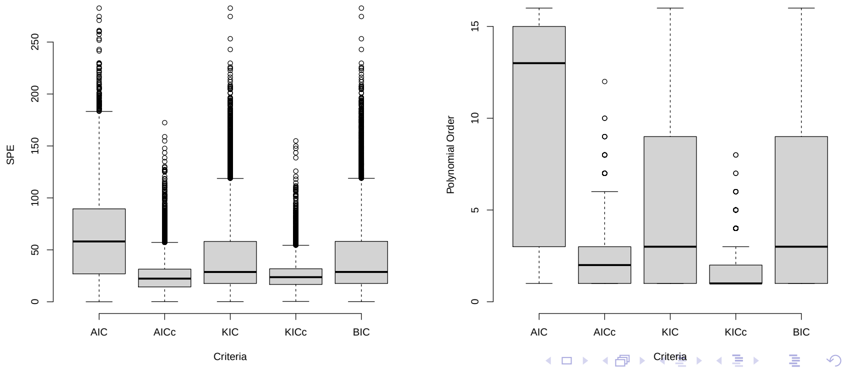
Simulation (5)

Test Function (4): $y = 1 + x + x^2 - 7x^3 + x^4 + 7x^5$



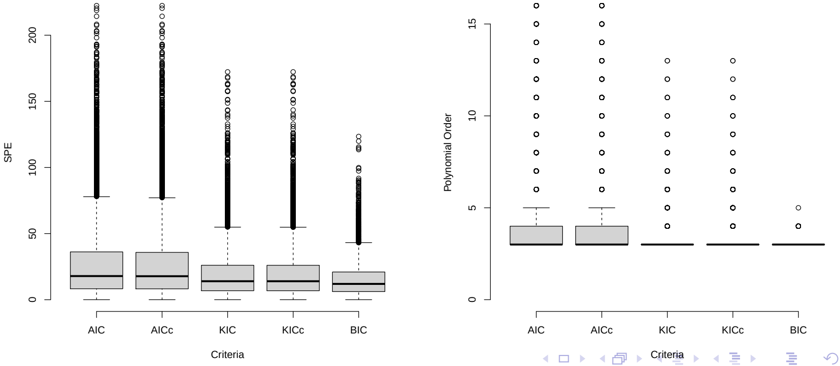
Simulation (6)

Test Function (1), $n = 20$, SNR $= 1$



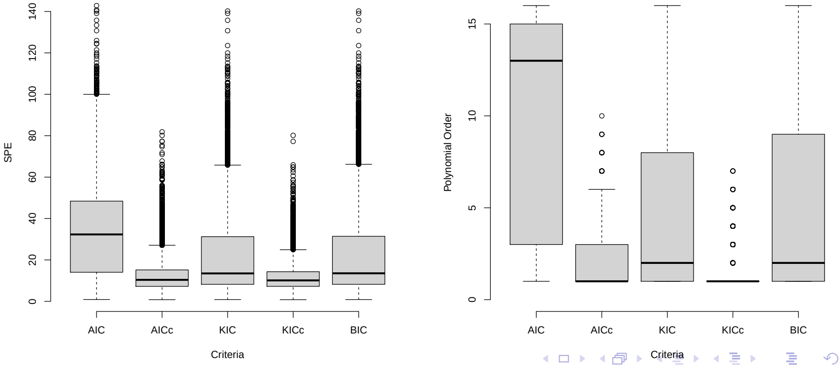
Simulation (7)

Test Function (1), $n = 2000$, SNR $= 1$



Simulation (8)

Test Function (4), $n = 20$, SNR $= 1$



Simulation (9)

Test Function (4), $n = 2000$, SNR $= 1$