Almost the best of three worlds
The switch model selection criterion for single-parameter exponential families

S.L. van der Pas

Joint work with prof. dr. P.D. Grünwald

Axioma symposium, June 4th 2014
Outline

The model selection problem

Example

Desirable properties

The AIC-BIC dilemma

The switch distribution

Optional stopping

Conclusions
The model selection problem
A single-parameter exponential family \( \{ p_\theta \mid \theta \in \Theta \} \) on \( \mathcal{X} \), where \( \Theta \subset \mathbb{R} \), is a model such that any member \( p_\theta \) can be written as

\[
p_\theta(x) = \frac{1}{z(\theta)} e^{\theta \phi(x)} r(x).
\]

We only consider regular exponential families.
A single-parameter exponential family $\{p_\theta | \theta \in \Theta\}$ on $\mathcal{X}$, where $\Theta \subset \mathbb{R}$, is a model such that any member $p_\theta$ can be written as

$$p_\theta(x) = \frac{1}{z(\theta)} e^{\theta \phi(x)} r(x).$$

We only consider regular exponential families.

Example:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = \frac{1}{e^{-\theta^2/2}} e^{\theta x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$
The model selection problem

Suppose \( \{p_\mu | \mu \in M\} \) is a regular single-parameter exponential family in its mean-value parameterization and we have data \( X_i \sim p_{\mu^*}, \mu^* \in M, i = 1, \ldots, n \). Let \( \mu_0 \in M \) be a constant. We wish to select one of the following two models:

\[
M_0 = \{p_{\mu_0}\} \quad \text{and} \quad M_1 = \{p_\mu, \mu \in M \setminus \mu_0\}.
\]
Example
Example

- Soal & Bateman, Modern Experiments in Telepathy (1954)
- 130 experiments with Mrs Stewart between 1945 and 1950
Example
Example

Results:

- 37,100 trials
- 9,410 'direct hits'
Results:

- 37,100 trials
- 9,410 'direct hits'

\( H_0: \) Mrs Stewart does not possess telepathic powers.

- \( p\)-value: \( 5.8 \cdot 10^{-139} \).

Estimated probability of guessing the correct card:
\[ \hat{\theta} = \frac{9,410}{37,100} \approx 0.25. \]
Desirable properties
Desirable properties

- Consistency
- Minimax-rate optimality
- Insensitivity to optional stopping
Consistency

Collection of parametric models \( \mathcal{M}_k = \{ p_\mu | \mu \in M_k \} \).

- A criterion \( \delta \) is consistent if for any \( \mu \) in any \( M_k \):

  \[
  \lim_{n \to \infty} \mathbb{P}_\mu(\delta(X^n) = k) = 1.
  \]

- ‘Essentially, all models are wrong, but some are useful.’ (Box and Draper, 1987)
Desirable properties

- Consistency
- Minimax-rate optimality
- Insensitivity to optional stopping
Minimax-rate optimality

- \( \hat{\mu}(x^n) \): estimator of \( \mu \).

- Quadratic loss: \( L(\mu, \hat{\mu}(x^n)) = (\mu - \hat{\mu}(x^n))^2 \).

- Quadratic risk:

\[
R(\mu, \hat{\mu}, n) = \mathbb{E}_\mu \left[ (\mu - \hat{\mu}(X^n))^2 \right].
\]
Minimax-rate optimality

- $\hat{\mu}(x^n)$: estimator of $\mu$.

- Quadratic loss: $L(\mu, \hat{\mu}(x^n)) = (\mu - \hat{\mu}(x^n))^2$.

- Quadratic risk:

  $$R(\mu, \hat{\mu}, n) = \mathbb{E}_\mu \left[ (\mu - \hat{\mu}(X^n))^2 \right].$$

- Quadratic risk of $\delta$:

  $$R(\mu, \delta, n) = E_\mu \left[ (\mu - \hat{\mu}_k(x^n))^2 \right].$$
The AIC-BIC dilemma
Desirable properties - the AIC-BIC dilemma

- Consistency
- Minimax-rate optimality
- Insensitivity to optional stopping
AIC and BIC

Collection of parametric models $\mathcal{M}_k = \{ p_\mu | \mu \in M_k \}$.

Selected model:

$$\arg \max_k 2L(x^n | \hat{\mu}_k) - 2 \cdot \text{(number of parameters)} \cdot c_n,$$

where

$$c_n = \begin{cases} 1 & \text{AIC} \\ \frac{1}{2} \log n & \text{BIC} \end{cases}.$$
$\mathcal{M}_0 = \{N(0, 1)\}, \mathcal{M}_1 = \{N(\mu, 1), \mu \in \mathbb{R}\}$. $\mathcal{M}_1$ is selected if

$$\left| \sum_{i=1}^{n} x_i \right| > \sqrt{2n} \quad \text{AIC}$$

$$\left| \sum_{i=1}^{n} x_i \right| > \sqrt{n \log n} \quad \text{BIC}$$
Complexity penalty
Complexity penalty
### Desirable properties - AIC and BIC

<table>
<thead>
<tr>
<th>Property</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Worst-case risk</td>
<td>( \frac{1}{n} )</td>
<td>( \log n )</td>
</tr>
<tr>
<td>Insensitive to optional stopping</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Theorem (analogue of Yang (2005))

Suppose $\{p_\mu | \mu \in M\}$ is a regular single-parameter exponential family in its mean-value parameterization and we have data $X_i \sim p_{\mu^*}, \mu^* \in M, i = 1, \ldots, n$. Let $\mu_0 \in M$ be a constant. Consider the following models:

$$M_0 = \{p_{\mu_0}\} \quad \text{and} \quad M_1 = \{p_{\mu}, \mu \in M \setminus \mu_0\}$$

Let $\delta$ be a consistent model selection criterion. Then:

$$n \sup_{\mu} R(\mu, \delta, n) \to \infty,$$

where $R$ is the standardized squared error risk.
The most important observation

Let $\hat{\mu}_k^n$ be the estimator of $\mu$ after selection of model $\mathcal{M}_k$. Then

$$R(\mu, \delta, n) = \mathbb{E}_\mu \left[ I(\mu)(\mu - \hat{\mu}_k^n)^2 \right]$$

is equal to

$$\mathbb{E}_\mu \left[ I(\mu)(\mu - \hat{\mu}_1^n)^2 1_{\{\delta(X^n) = 1\}} \right] + I(\mu)(\mu - \mu_0)^2 \mathbb{P}_\mu(\delta(X^n) = 0).$$
The switch distribution
Example
Data from a $\mathcal{N}(\mu^*, 1)$-distribution, $(\mu^*)^2 = \frac{1}{100}$. Candidate models:

$\mathcal{M}_0 = \{\mathcal{N}(0, 1)\}$ vs. $\mathcal{M}_1 = \{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$

Simple model:

$$E_{\mu^*} \left[ (\mu^* - 0)^2 \right] = \frac{1}{100}$$

Complex model:

$$E_{\mu^*} \left[ (\mu^* - \hat{\mu}_n)^2 \right] = \frac{1}{n}.$$
Frequentist Bayes

Example: coin toss, $\theta = \text{probability of heads}$. 

- Prior: $\theta \sim \text{Unif}[0,1]$.
- Data: $n$ coin tosses, $h$ heads, $t$ tails.
- Posterior: $\pi(\theta | \text{data}) \sim \text{Beta}(h+1, t+1)$.
- Estimator: $E[\theta | \text{data}] = \frac{h+1}{n+2}$. 

Frequentist Bayes: assume $\theta_0$ generates the data.
Frequentist Bayes

Example: coin toss, $\theta =$ probability of heads.

- Prior: $\theta \sim \text{Unif}[0,1]$.
- Data: $n$ coin tosses, $h$ heads, $t$ tails.
- Posterior: $\theta|\text{data} \sim \text{Beta}(h + 1, t + 1)$. 
Example: coin toss, $\theta =$ probability of heads.

- Prior: $\theta \sim \text{Unif}[0,1]$.
- Data: $n$ coin tosses, $h$ heads, $t$ tails.
- Posterior: $\theta|\text{data} \sim \text{Beta}(h + 1, t + 1)$.
- Estimator: $\mathbb{E}[\theta|\text{data}] = \frac{h+1}{n+2}$.
Frequentist Bayes

Example: coin toss, $\theta =$ probability of heads.

- Prior: $\theta \sim \text{Unif}[0,1]$.
- Data: $n$ coin tosses, $h$ heads, $t$ tails.
- Posterior: $\theta | \text{data} \sim \text{Beta}(h + 1, t + 1)$.
- Estimator: $\mathbb{E}[\theta | \text{data}] = \frac{h+1}{n+2}$.

*Frequentist* Bayes: assume $\theta_0$ generates the data.
The switch distribution

Candidate models:

\[ \mathcal{M}_0 = \{p_{\mu_0}\} \quad \text{vs.} \quad \mathcal{M}_1 = \{p_{\mu}\mid \mu \in \mathcal{M}\backslash\mu_0\} \]

Associated prediction strategies:

\[ p_0(x_n|x^{n-1}) = p_{\mu_0}(x_n|x^{n-1}), \quad p_1(x_n|x^{n-1}) = \frac{p_B(x^n)}{p_B(x^{n-1})}, \]

where

\[ p_B(x^n) = \int_{\mathcal{M}} p_{\mu}(x^n)\omega(\mu)d\mu. \]

Prediction strategy that switches after \( j \) observations:

\[ \bar{p}_j(x_n|x^{n-1}) = \begin{cases} p_{\mu_0}(x_n|x^{n-1}) & n \leq j \\ p_1(x_n|x^{n-1}) & n > j \end{cases}. \]
The switch distribution

Switching allowed after $2^0, 2^1, \ldots, 2^{\lceil \log_2(n-1) \rceil}$ observations.

Prior on strategies that switch after $2^i$ observations, for example:

$$\pi(i) = \frac{1}{(i + 2)(i + 3)}.$$

Prior on strategies predicting exclusively with simple or complex model: $\frac{1}{4}$ each.

Yields posterior on the strategies. Let $K_{n+1}(s)$ be the model used by strategy $s$ after $n$ observations. Then:

$$\delta_{sw}(x^n) = \arg\max_k p_{sw}(K_{n+1} = k|x^n).$$
The criterion $\delta_{\text{sw}}$

$$\delta_{\text{sw}}(x^n) = \begin{cases} 
1 & \text{if } \frac{\rho_{\text{sw}, 1}(x^n)}{\rho_{\text{sw}, 0}(x^n)} > 1 \\
0 & \text{if } \frac{\rho_{\text{sw}, 1}(x^n)}{\rho_{\text{sw}, 0}(x^n)} \leq 1 
\end{cases}$$

where

$$\rho_{\text{sw}, 1}(x^n) = \sum_{i=0}^{\lfloor \log_2 n \rfloor} \pi(i) \bar{p}_{2i}(x^n) + \frac{1}{4} \rho_B(x^n)$$

$$\rho_{\text{sw}, 0}(x^n) = \left( \frac{3}{4} - \sum_{i=0}^{\lfloor \log_2 n \rfloor} \pi(i) \right) \rho_{\mu_0}(x^n)$$
The criterion $\delta_{SW}$

$$
\delta_{SW}(x^n) = \begin{cases} 
1 & \text{if } \frac{p_{SW, 1}(x^n)}{p_{SW, 0}(x^n)} > 1 \\
0 & \text{if } \frac{p_{SW, 1}(x^n)}{p_{SW, 0}(x^n)} \leq 1
\end{cases},
$$

where

$$
p_{SW, 1}(x^n) = \sum_{i=0}^{\lfloor \log_2 n \rfloor} \pi(i) \bar{p}_{2i}(x^n) + \frac{1}{4} p_B(x^n)
$$

and

$$
p_{SW, 0}(x^n) = \left( \frac{3}{4} - \sum_{i=0}^{\lfloor \log_2 n \rfloor} \pi(i) \right) p_{\mu_0}(x^n)
$$

$\delta_{SW}$ is consistent.
Main theorem

Let \( \{ p_\mu | \mu \in M \} \) be a regular single-parameter exponential family in its mean-value parameterization. Let \( \mu_0 \in M \) be a constant. Consider the following models:

\[
M_0 = \{ p_{\mu_0} \} \quad \text{and} \quad M_1 = \{ p_\mu, \mu \in M \backslash \mu_0 \}.
\]

When \( \delta_{sw} \) is used with

(i) a prior \( \pi(i) \) on the strategies that switch after \( 2^i \) observations such that \( \pi(i) \propto i^{-k} \) for some \( k \geq 2 \);

(ii) a predictive distribution based on the Bayesian marginal likelihood with a prior \( \omega \) that admits a strictly positive, continuous density;

and the parameter \( \mu \) is estimated within \( M_1 \) by an efficient estimator \( \hat{\mu}_1 \), then the worst-case standardized quadratic risk is of order \( \log \log n / n \).
Simulations - quadratic loss

\( X_i \sim \mathcal{N}(1/10, 1), i = 1, \ldots, n. \)
Simulations - consistency

Simple model $M_0$ true: $X_i \sim \mathcal{N}(0, 1), i = 1, \ldots, n$. 
Simulations - consistency

Complex model $\mathcal{M}_1$ true: $X_i \sim \mathcal{N}(\mu, 1), i = 1, \ldots, n.$
Optional stopping
Desirable properties

- Consistency
- Minimax-rate optimality
- Insensitivity to optional stopping
Optional stopping

“A psychologist who received a special grant for the purpose might perform a limited number of experiments with impunity so long as the investigation appeared to prove that telepathy did not happen. But if an academic man showed any enthusiasm and a tendency to go on in the face of discouragement, he would soon be frowned upon and accused of wasting his time. His sanity might even be doubted.” (Soal and Bateman, p.23)
Optional stopping

“A psychologist who received a special grant for the purpose might perform a limited number of experiments with impunity so long as the investigation appeared to prove that telepathy did not happen. But if an academic man showed any enthusiasm and a tendency to go on in the face of discouragement, he would soon be frowned upon and accused of wasting his time. His sanity might even be doubted.” (Soal and Bateman, p.23)
Optional stopping

More conservative version of $\delta_{sw}$ is insensitive to optional stopping:

$$
\delta_{sw}(X^n) = \begin{cases} 
1 & \text{if } \frac{p_{sw, 1}(x^n)}{p_{sw, 0}(x^n)} > 4k - 1 \\
0 & \text{if } \frac{p_{sw, 1}(x^n)}{p_{sw, 0}(x^n)} \leq 4k - 1
\end{cases},
$$

where $k > 1$.

For any $n$ and any $k > 1$, by a version of Doob’s martingale inequality (see Shafer et al. (2011)):

$$
\mathbb{P}_{\mu_0} \left( \exists n : \delta_{sw}(X^n) = 1 \right) \leq \frac{1}{k}.
$$
### Overview

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>Switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Worst-case risk</td>
<td>$\frac{1}{n}$</td>
<td>$\log n$</td>
<td>$\log \log n$</td>
</tr>
<tr>
<td>Insensitive to optional stopping</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Conclusions

- Three desirability criteria: consistency, minimax-rate optimality, insensitivity to the stopping rule.

- A choice is inevitable.

- $\delta_{sw}$ combines consistency and insensitivity to optional stopping with near optimal worst-case instantaneous risk.