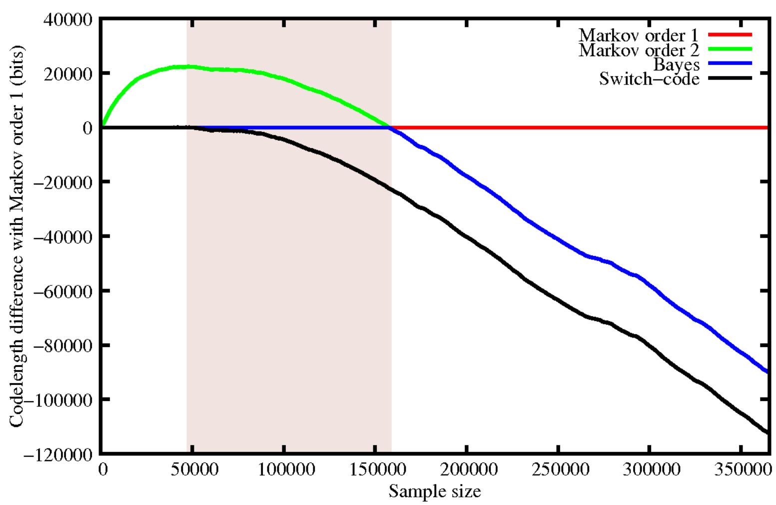
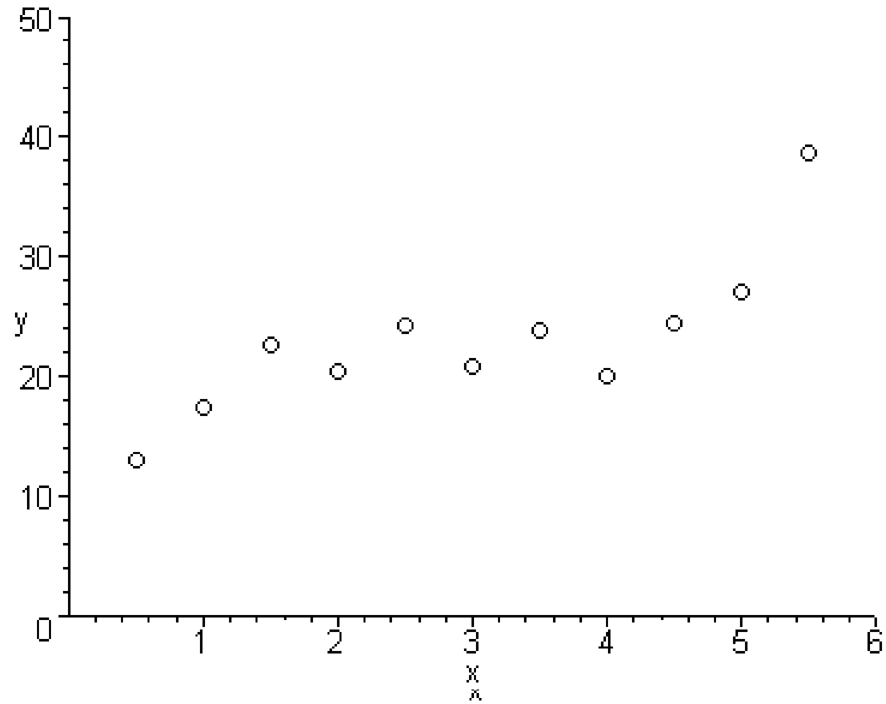
The Catch-Up Phenomenon

Peter Grünwald
www.grunwald.nl

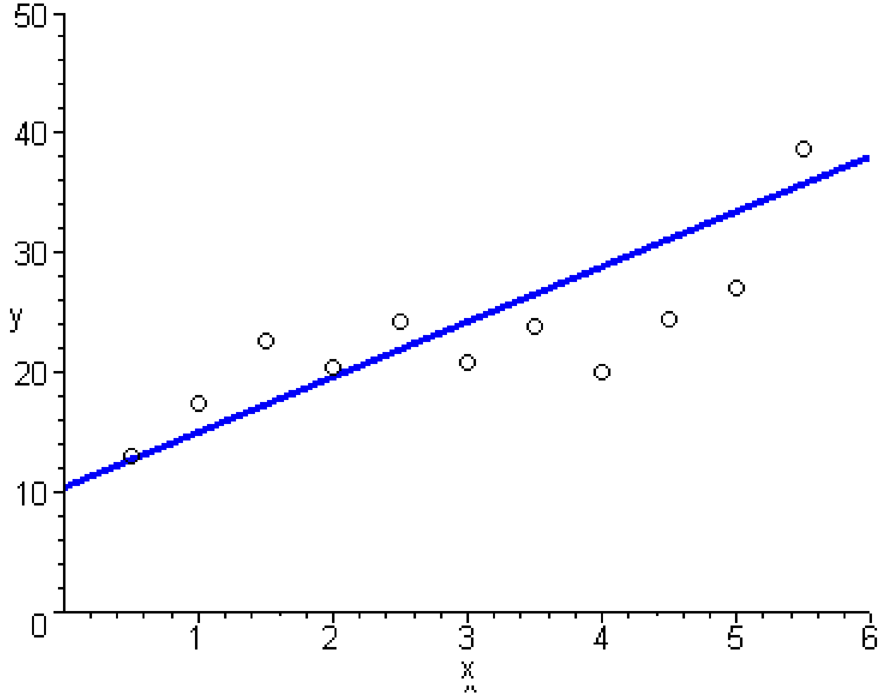
Joint work with Tim van Erven, Steven de Rooij, Wouter Koolen



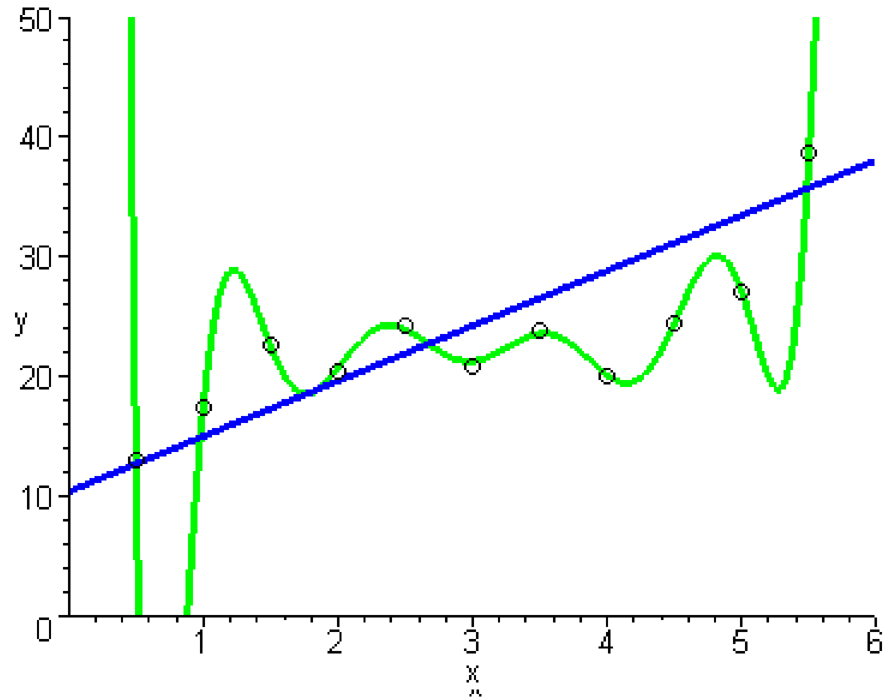
Model Selection



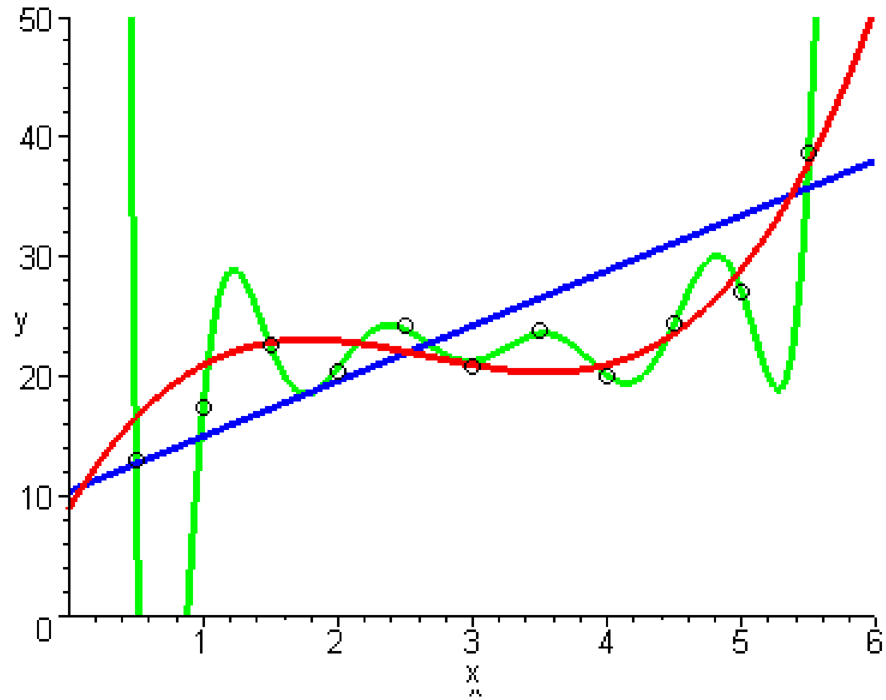
Model Selection



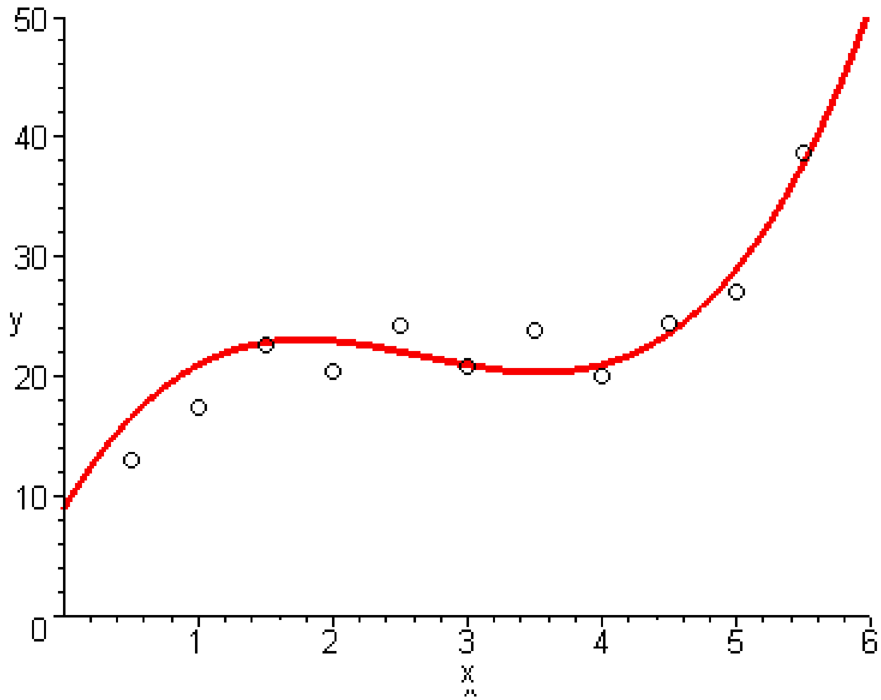
Model Selection



Model Selection



Model Selection



Model Selection Methods

- Suppose we observe data $y^n = y_1, \ldots, y_n \in \mathcal{Y}^n$
- We want to know which model in our list of candidate models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ best explains the data

- In this talk, $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k \subseteq \mathbb{R}^k\}$ is k-parameter set of probability distributions
  - polynomials with Gaussian noise (regression)
  - histograms with varying number of bins
  - Markov chains of increasing order

Model Selection Methods

- Suppose we observe data $y^n = y_1, \ldots, y_n \in \mathcal{Y}^n$
- We want to know which model in our list of candidate models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ best explains the data

- A model selection method

$$\widehat{k} : \bigcup_{n \geq 1} \mathcal{Y}^n \rightarrow \mathbb{N}$$

is a function mapping data sequences of arbitrary length to model indices

  - $\widehat{k}(y^n)$ is model chosen for data y^n

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion (AIC, 1973)

$$\widehat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\widehat{\theta}_k}(x^n) + k$$

2. **BIC-type**
   - Bayesian Information Criterion (BIC, 1978)

$$\widehat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\widehat{\theta}_k}(x^n) + \frac{k}{2}\log n$$

Googling "AIC *and* BIC": 355000 hits

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   – Akaike Information Criterion (AIC, 1973)

$$\hat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\hat{\theta}_k}(x^n) + k$$

2. **BIC-type**
   – Bayesian Information Criterion (BIC, 1978)

$$\hat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\hat{\theta}_k}(x^n) + \frac{k}{2}\log n$$

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion (AIC, 1973)
   - **leave-one-out cross-validation**
   - DIC, C_p
2. **BIC-type**
   - Bayesian Information Criterion (BIC, 1978)
   - prequential validation
   - **Bayes factor model selection**
   - standard Minimum Description Length (MDL)

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion
   - **leave-one-out cross-validation**
   - DIC, C_p
2. **BIC-type**
   - Bayesian Information Criterion
   - prequential validation
   - **Bayes factor model selection**
   - standard MDL

asymptotic overfitting

inconsistent

consistent

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion
   - **leave-one-out cross-validation**
   - DIC, C_p
2. **BIC-type**
   - Bayesian Information Criterion
   - prequential validation
   - **Bayes factor model selection**
   - standard MDL

inconsistent 

optimal rate 

consistent 

slower rate 

asymptotic underfitting

The Best of Both Worlds

We present the first model selection criterion that is provably **both consistent** and **optimal** in terms of **prediction and estimation**

Example: Histograms

- Assume $Y_1, Y_2, \ldots$ are identically and independently distributed according to some p^* on $\mathcal{Y} = [0, 1]$

- We model data using k-bin equal-width histograms, and try to determine k based on data y^n

Example: Histograms

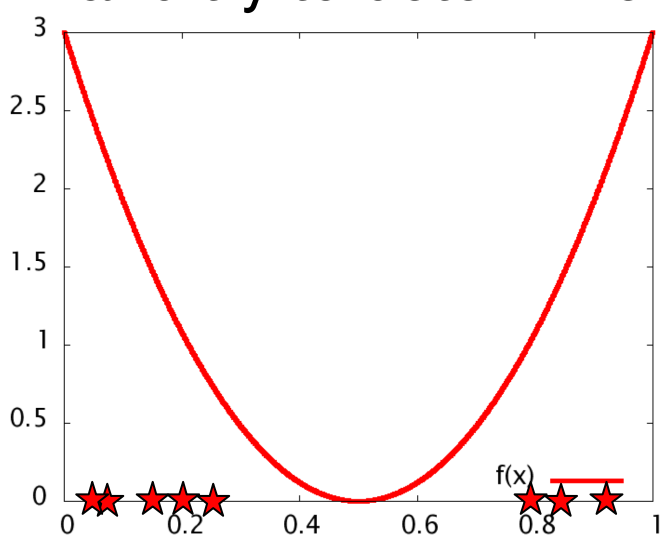
- Assume $Y_1, Y_2, \ldots$ are identically and independently distributed according to some p^* on $\mathcal{Y} = [0, 1]$

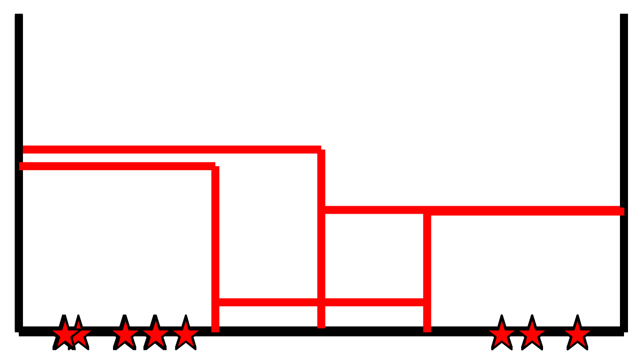
- We model data using k-bin equal-width histograms, and try to determine k based on data y^n

Example: Histograms

- Assume $Y_1, Y_2, \ldots$ are identically and independently distributed according to some p^* on $\mathcal{Y} = [0, 1]$

- We model data using k-bin equal-width histograms, and try to determine k based on data y^n



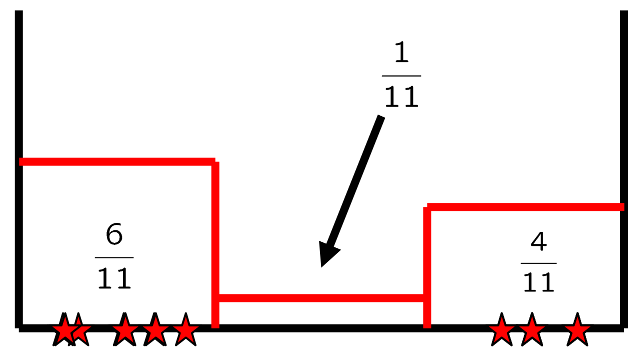


Example: Histograms

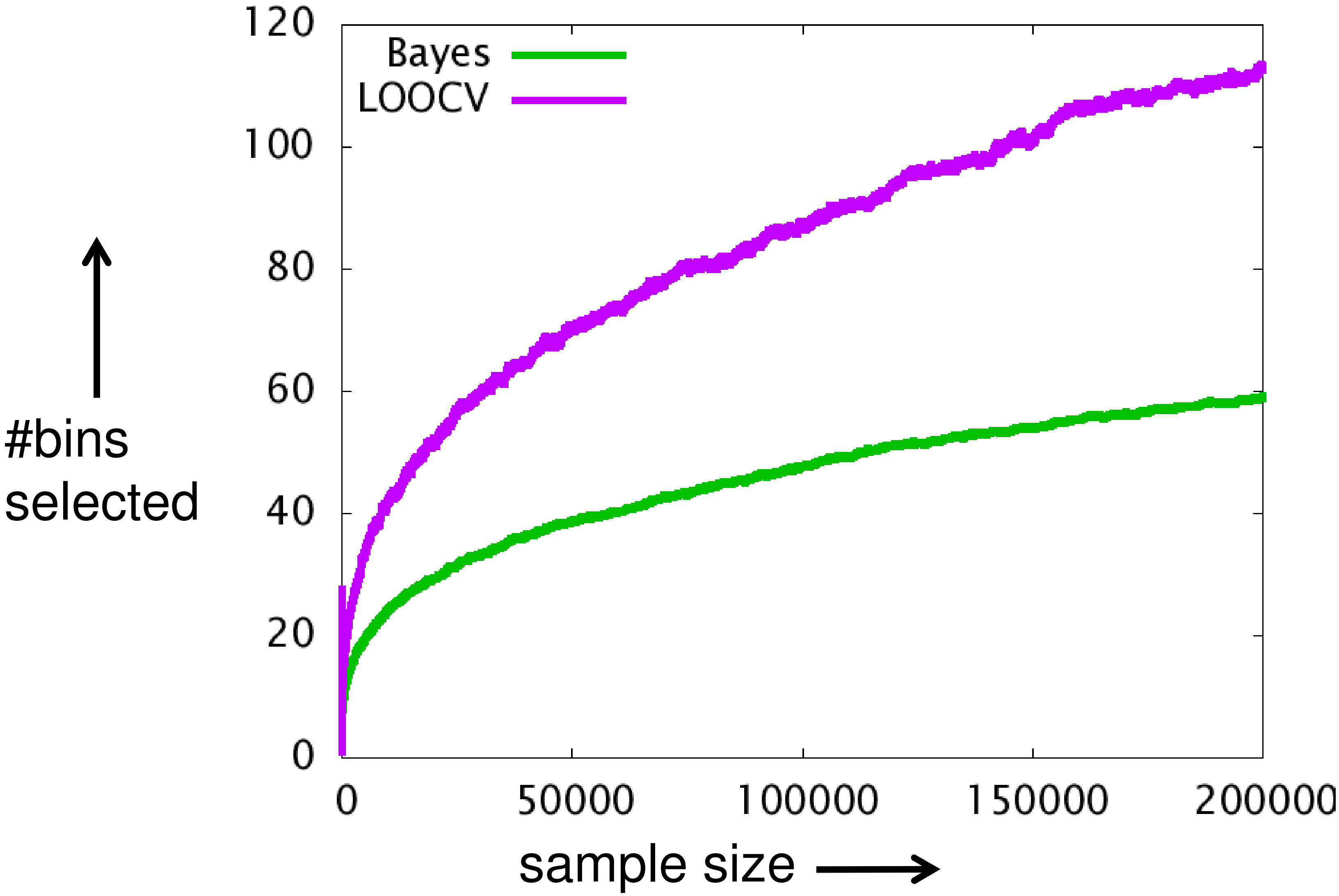
- $\mathcal{M}_k$ is family of k-bin histograms with equal widths
- Given $\mathcal{M}_k$ predict/estimate using Laplace estimator, for $j = 1..k$,

$$\bar{p}_k \left(Y_{n+1} \text{ falls in bin } j \mid y^n \right) = \frac{(\#\text{ points in } y^n \text{ in bin } j\ ) + 1}{n + k}$$

- As in Rissanen, Speed, Yu (1993)
- Equivalent to Bayes predictive distribution with uniform (Dirichlet(1, .., 1)) prior



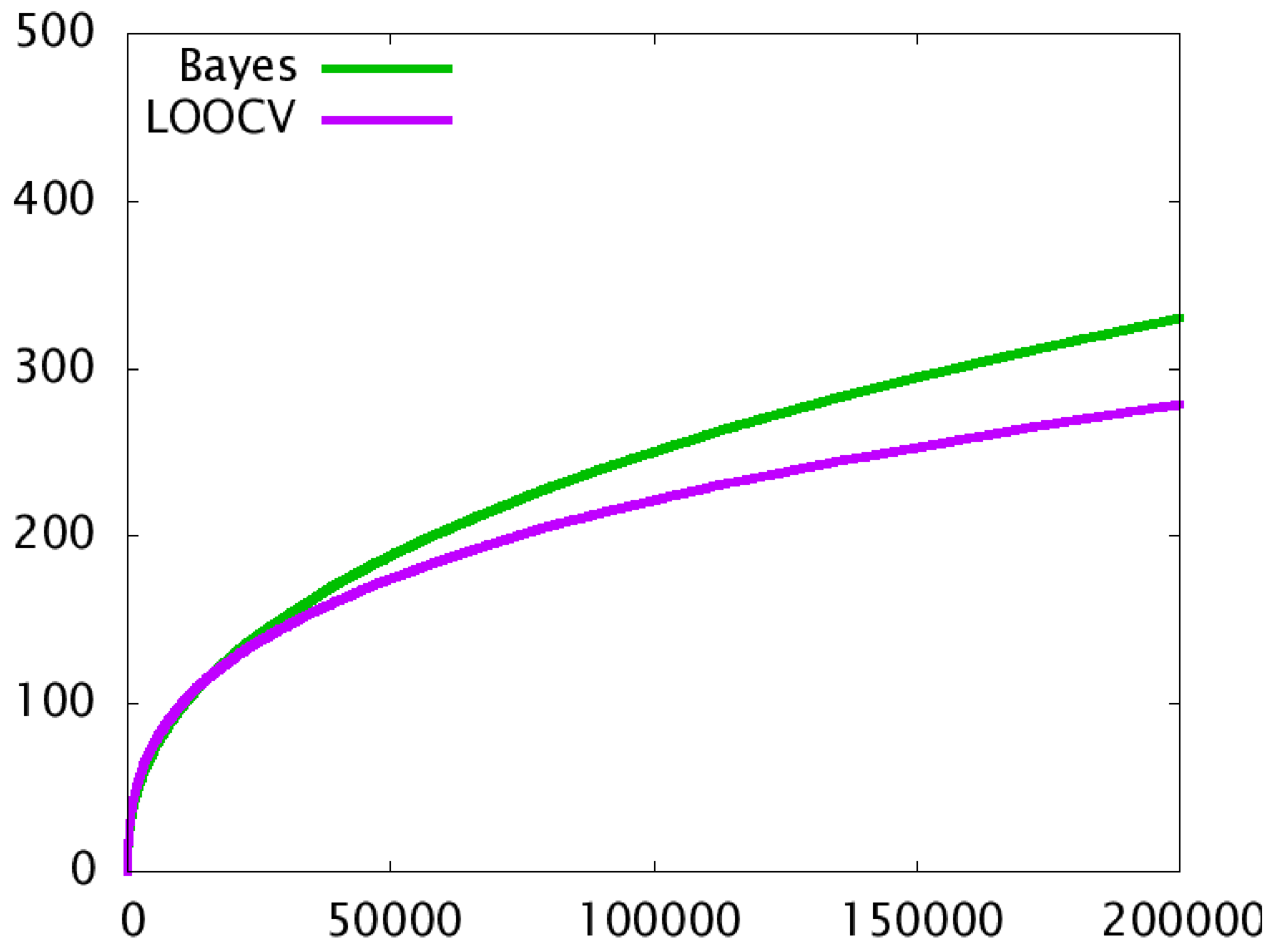
CV selects more bins than Bayes



CV predicts better than Bayes

accumulated
prediction error
measured in
log-loss

$$\sum_{i=1}^{n} -\log \bar{p}_{\hat{k}(y^{i-1})}(y_i \mid y^{i-1})$$



sample size →

CV predicts better than Bayes

- Data sampled from P^* that is not in set of models $\bigcup_{k \geq 1} \mathcal{M}_k$, but in their closure

- LOO-CV, AIC converge at optimal rate,
- Bayesian model selection/averaging is too slow (underfits!)

↑

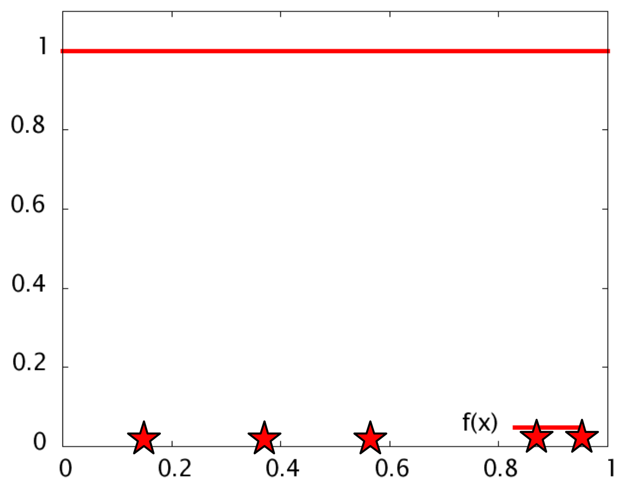
accumulated prediction error measured in log-loss

$$\sum_{i=1}^{n} - \log \bar{p}_{\hat{\theta}_{\hat{k}}(y^{i-1})}(y_i)$$

sample size ⟶

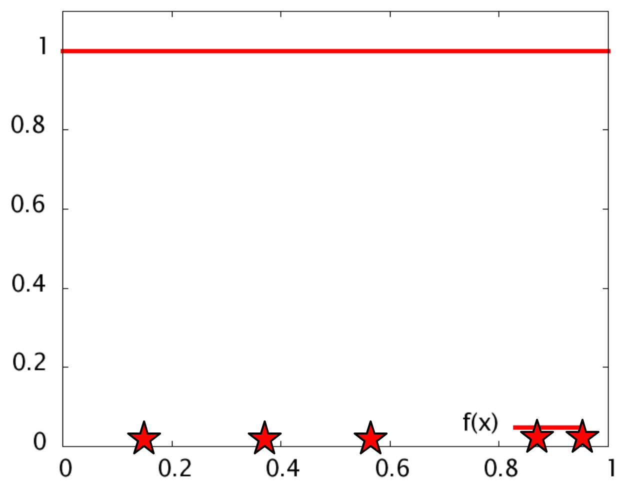
...but CV is inconsistent!

- Now suppose data are sampled from the uniform distribution...

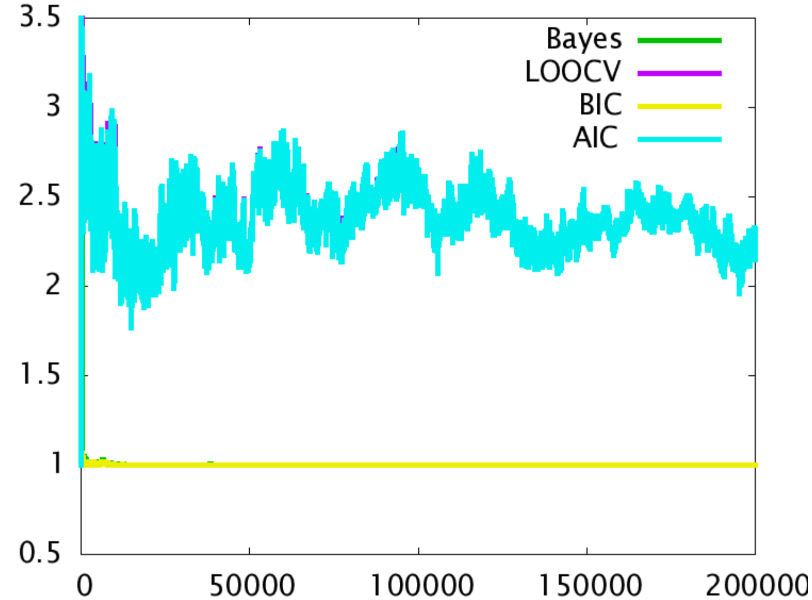


...but CV is inconsistent!

- Now suppose data are sampled from the uniform distribution...



#bins selected



sample size

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*
- This allows us to define a model selection/averaging method that, in a wide variety of circumstances,
  1. is provably **consistent**
  2. provably achieves **optimal convergence rates**

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*
- This allows us to define a model selection/averaging method that, in a wide variety of circumstances,
  1. is provably **consistent**
  2. provably achieves **optimal convergence rates**
- …even though it had been suggested that this is impossible! Yang 2005, Forster 2001, Sober 2004
- For many model classes, method is computationally feasible

Menu

1. Bayes Factor Model Selection
   - Predictive interpretation

2. The Catch-Up Phenomenon

   …. as exhibited by the Bayes factor method

3. Solving the AIC-BIC Dilemma
   - Theorems

Bayes Factor Model Selection

$\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\} \qquad \Theta_k \subseteq \mathbb{R}^k \qquad k \in \mathcal{K} \subset \mathbb{N}$

$\hat{k}(y^n)$ is k maximizing a posteriori probability

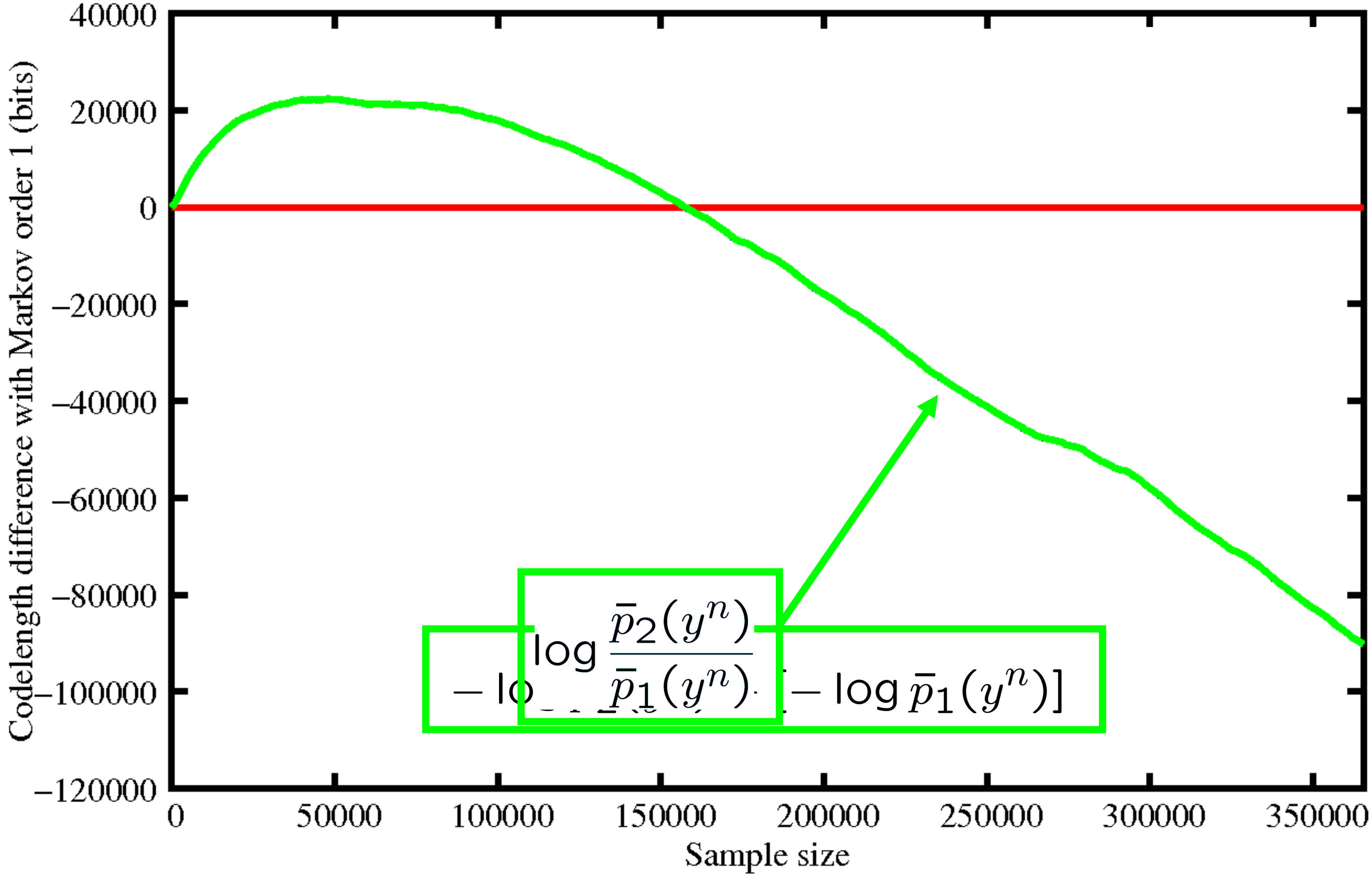
$$p(\mathcal{M}_k \mid y^n) = \frac{p(y^n \mid \mathcal{M}_k)\pi(k)}{\sum_{k \in \mathcal{K}} p(y_n \mid \mathcal{M}_k)\pi(k)}$$

$$\bar{p}_k := p(y^n \mid \mathcal{M}_k) = \int_{\theta \in \Theta_k} p_\theta(y^n) w_k(\theta) d\theta$$

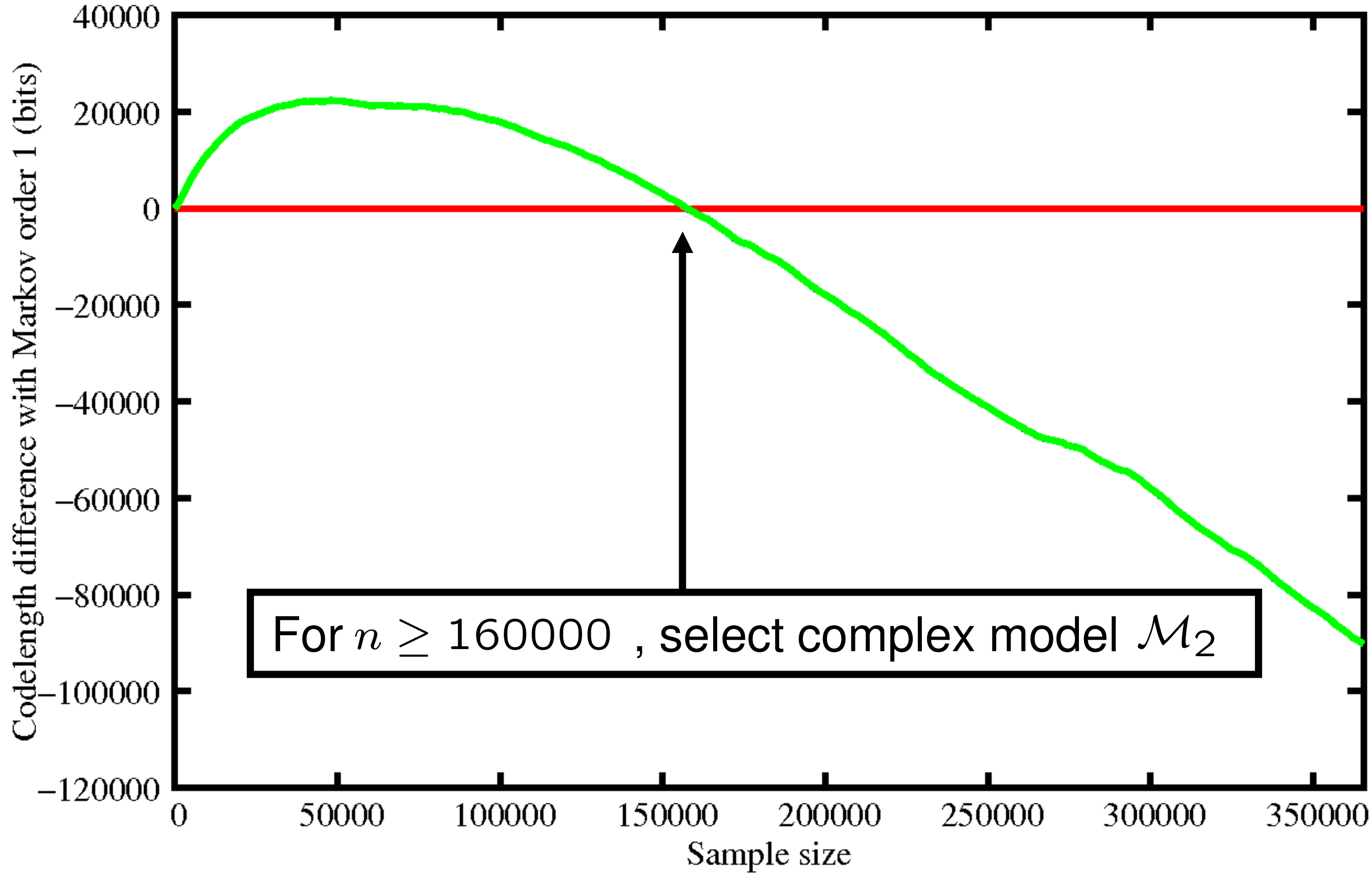
$\pi(k)$ is prior

$w_1, w_2, \ldots$ are priors

$\hat{k}(y^n)$ is & minimizing $-\log \bar{p}_k(y^n) - \log \pi(k) \approx -\log \bar{p}_k(y^n)$



Bayes factor model selection between 1st-order and 2nd-order Markov model for "The Picture of Dorian Gray"



Bayes factor model selection between 1st-order and 2nd-order Markov model for “The Picture of Dorian Gray”

The Catch-Up Phenomenon

- Suppose we select between "simple" model $\mathcal{M}_1$ and "complex" model $\mathcal{M}_2$
- Common Phenomenon: for some n_{switch}

  simple model predicts better if $n < n_{\text{switch}}$

  complex model predicts better if $n \geq n_{\text{switch}}$

  – this seems to be the very reason why it makes sense to prefer a simple model even if the complex one is true
- We would expect Bayes factor method to switch at about $n \approx n_{\text{switch}} \ldots$

  but is this really where Bayes switches!?

Menu

1. Bayes Factor Model Selection
   - Predictive interpretation

2. The Catch-Up Phenomenon

   …. as exhibited by the Bayes factor method

3. Solving the AIC-BIC Dilemma
   - Theorems
   - Discussion

Bayesian prediction

- Given model $\mathcal{M}_k$, Bayesian marginal likelihood is

$$\bar{p}_k(y^n) = p(y^n \mid \mathcal{M}_k) := \int_{\Theta_k} p_\theta(y^n) w(\theta) d\theta$$

- Given model $\mathcal{M}_k$, predict by predictive distribution

$$\bar{p}_k(y_{n+1} \mid y^n) = \frac{\bar{p}_k(y^{n+1})}{\bar{p}_k(y^n)} = \int_{\Theta_k} p_\theta(y_{n+1} \mid y^n) w(\theta \mid y^n) d\theta$$

Logarithmic Loss

- If we measure prediction quality by 'log loss',

$$\text{loss}(y, p) := -\log p(y)$$

then accumulated loss satisfies

$$\sum_{i=1}^{n} \text{loss}(y_i, p(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} \left[-\log p(y_i \mid y^{i-1}) \right]$$

$$= -\log \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1}) = -\log \prod_{i=1}^{n} \frac{p(y^i)}{p(y^{i-1})}$$

$$= -\log p(y_1, \ldots, y_n)$$

so that **accumulated log loss = minus log likelihood**

The Most Important Slide

- Bayes picks the k minimizing

$$-\log \bar{p}_k(y_1, \ldots, y_n) = \sum_{i=1}^{n} \text{loss}(y_i, \bar{p}_k(\cdot \mid y^{i-1}))$$

- **Prequential interpretation** of Bayes model selection: select the model $\mathcal{M}_k$ such that, when used as a sequential prediction strategy, $\bar{p}_k = p(\cdot \mid \mathcal{M}_k)$ minimizes accumulated sequential prediction error

Dawid '84, Rissanen '84

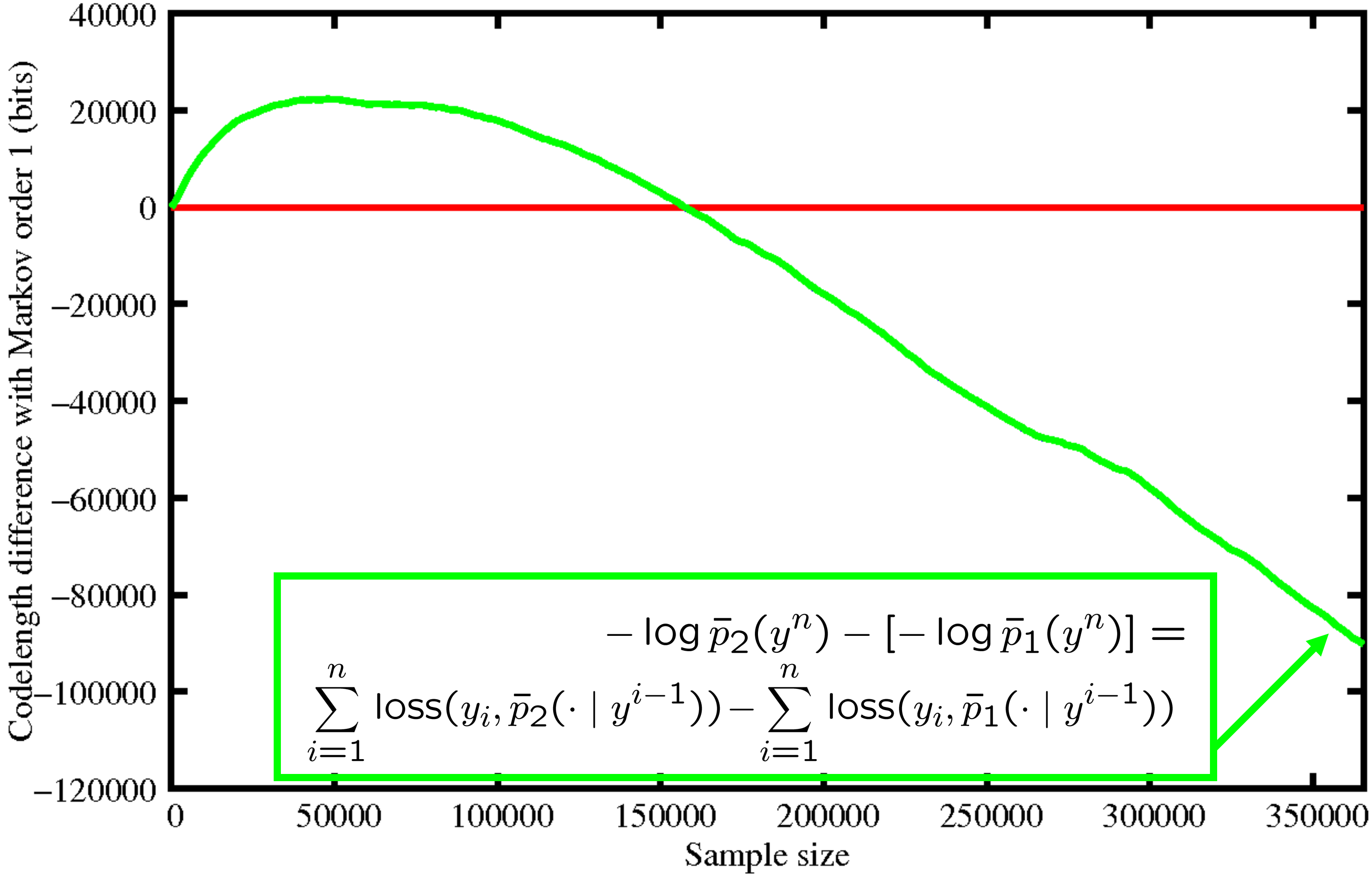
Menu

1. Bayes Factor Model Selection
   - Predictive interpretation

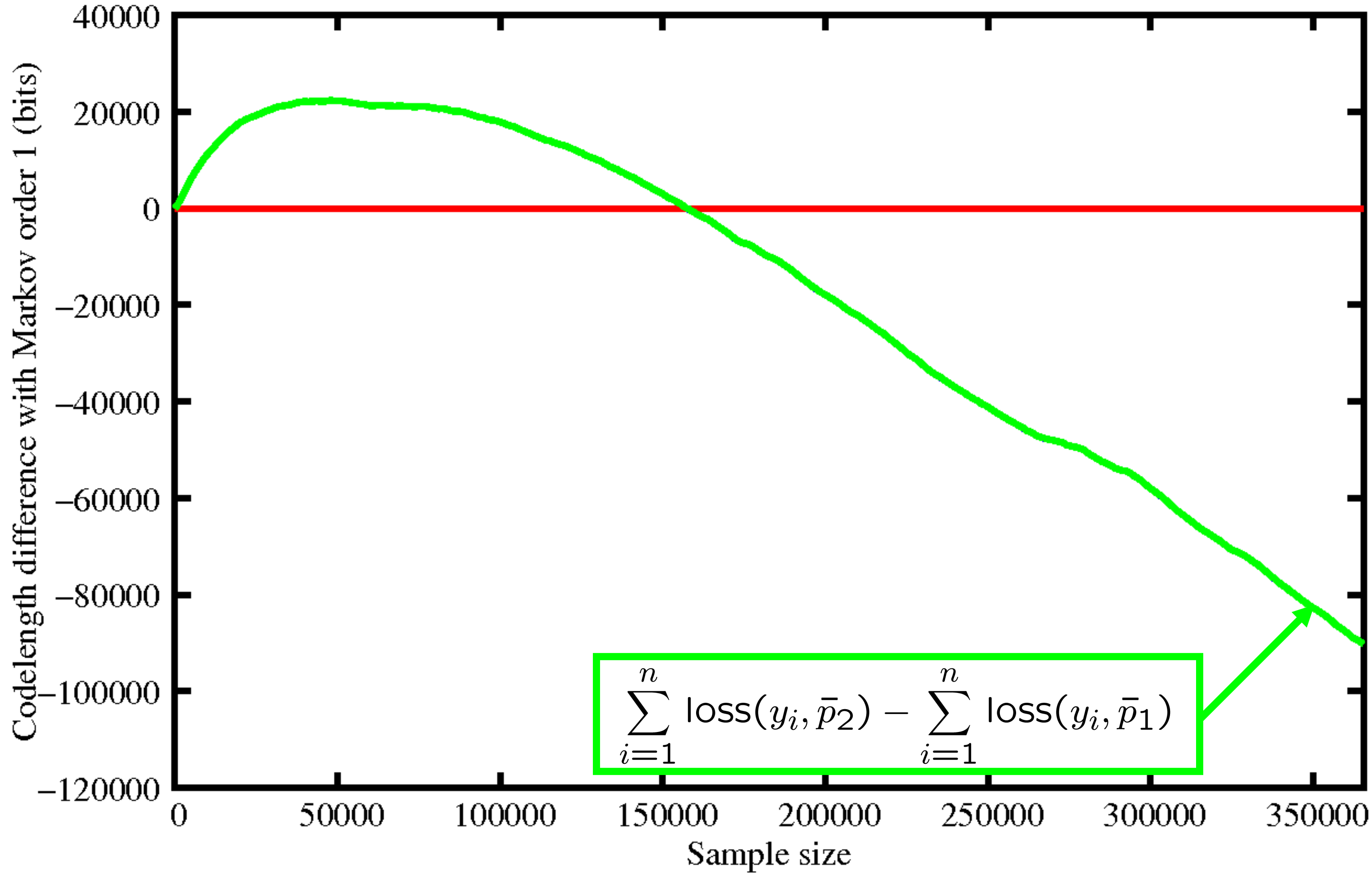
2. The Catch-Up Phenomenon

   …. as exhibited by the Bayes factor method

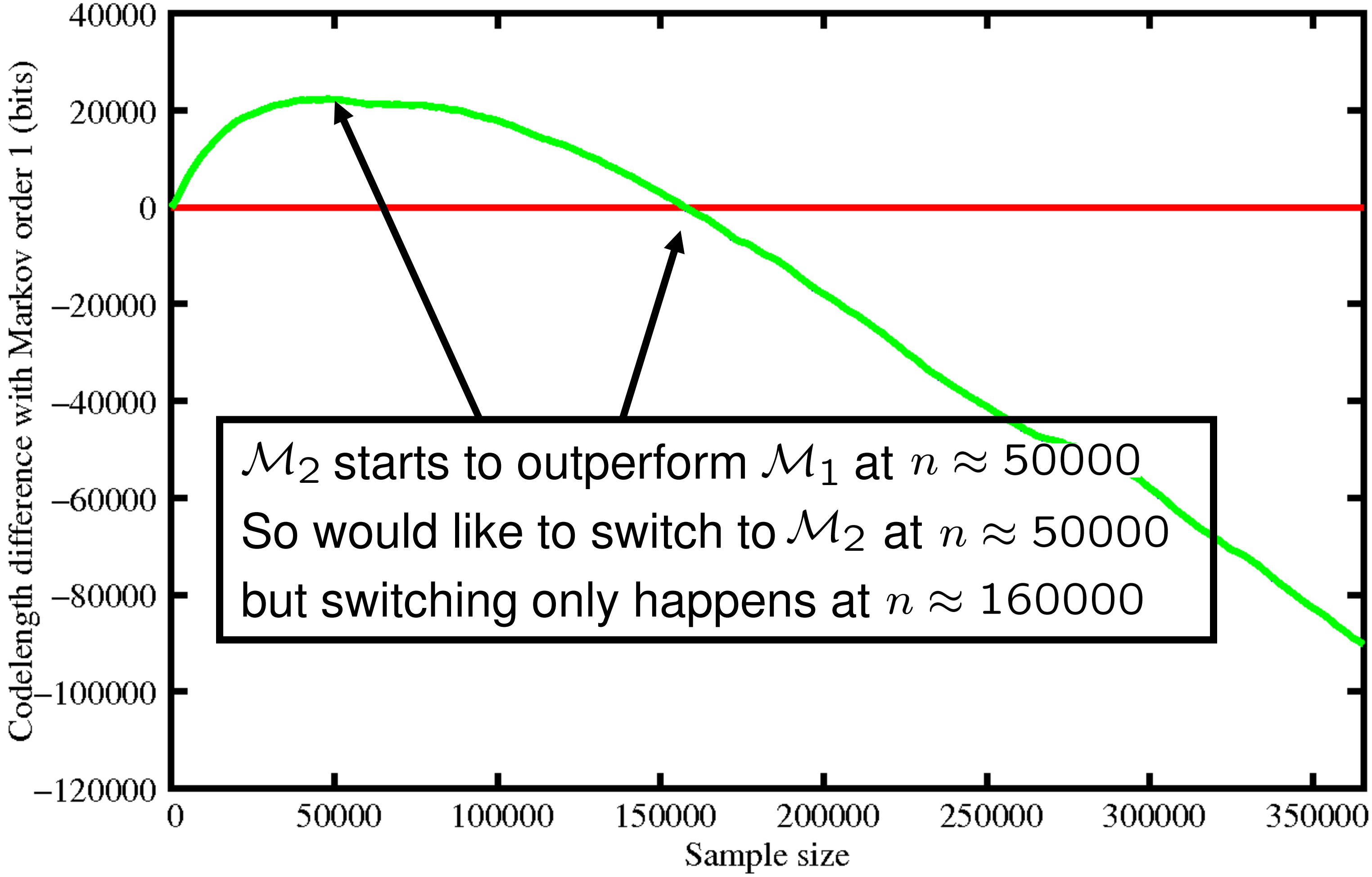
3. Solving the AIC-BIC Dilemma
   - Theorems
   - Discussion
   - Initial Experiments



Green curve depicts difference in accumulated prediction error between predicting with $\mathcal{M}_2$ and predicting with $\mathcal{M}_1$



Green curve depicts difference in accumulated prediction error between predicting with $\mathcal{M}_2$ and predicting with $\mathcal{M}_1$



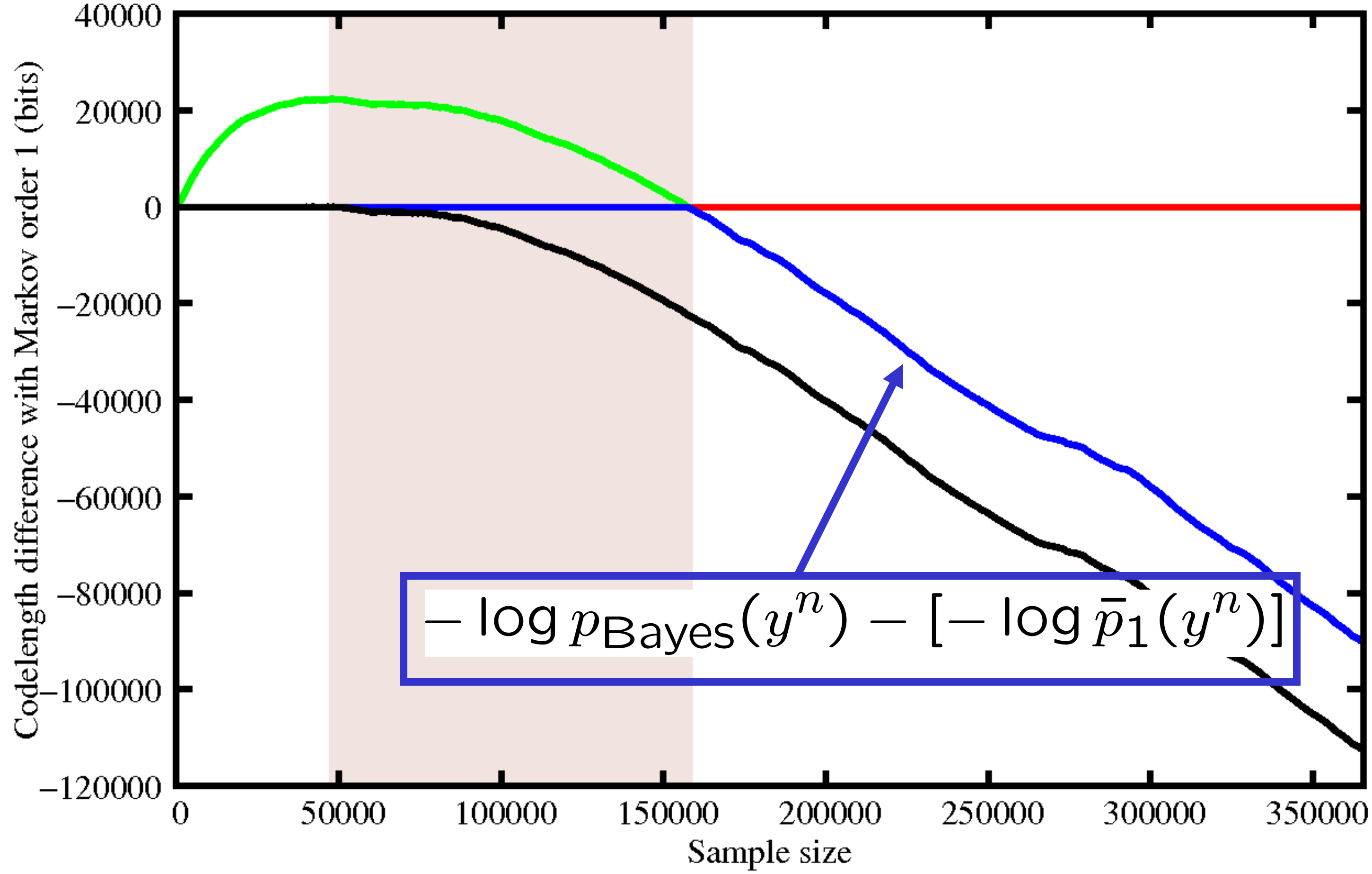
$$\sum_{i=1}^{50000} \mathsf{loss}(y_i, \bar{p}_2) - \sum_{i=1}^{50000} \mathsf{loss}(y_i, \bar{p}_1) = \mathbf{20000}$$

$$\sum_{i=1}^{160000} \mathsf{loss}(y_i, \bar{p}_2) - \sum_{i=1}^{160000} \mathsf{loss}(y_i, \bar{p}_1) = \mathbf{0}$$

The Catch-Up Phenomenon

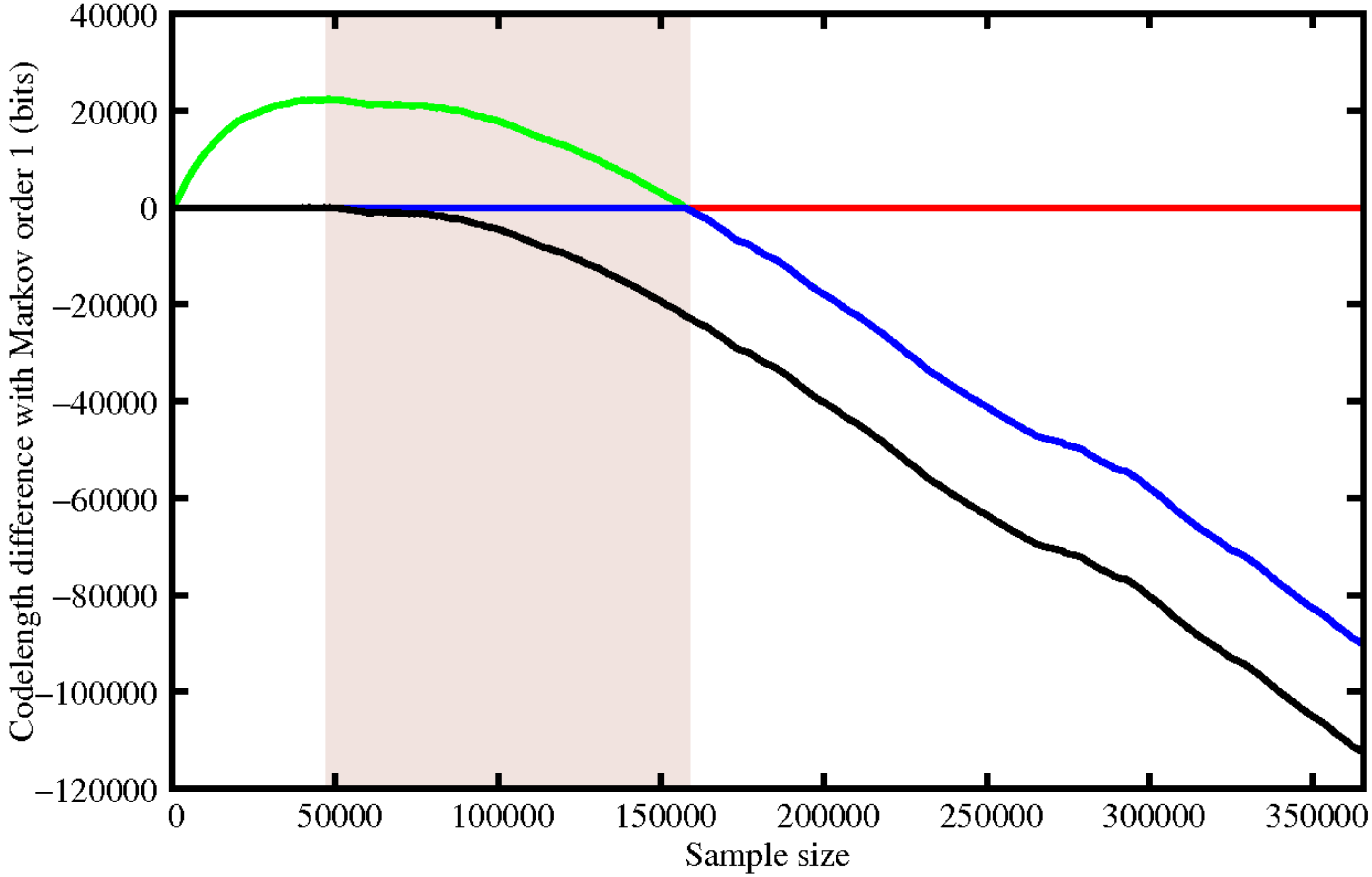
- Suppose we select between "simple" model $\mathcal{M}_1$ and "complex" model $\mathcal{M}_2$
- Common Phenomenon: for some n_{switch}
  simple model predicts better if $n < n_{\text{switch}}$
  complex model predicts better if $n \geq n_{\text{switch}}$

- Bayes exhibits **inertia**: complex model has to "**catch up**", so we prefer simpler model for a while even after $n \geq n_{\text{switch}}$



Model averaging does not help!

$$p_{\text{Bayes}}(y^n) = \frac{1}{2}\bar{p}_1(y^n) + \frac{1}{2}\bar{p}_2(y^n)$$



Can we modify Bayes so as to do as well as the black curve? Almost!

The Switch Distribution

- Suppose we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at sample size ***s***
- Our total prediction error is then

$$\sum_{i=1}^{s} \text{loss}(y_i, \bar{p}_1) + \sum_{s+1}^{n} \text{loss}(y_i, \bar{p}_2)) =$$

$$- \log \bar{p}_1(y^s) - \log \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

The Switch Distribution

- Suppose we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at sample size ***s***
- Our total prediction error is then

$$\sum_{i=1}^{s} \text{loss}(y_i, \bar{p}_1) + \sum_{s+1}^{n} \text{loss}(y_i, \bar{p}_2)) =$$

$$- \log \bar{p}_1(y^s) - \log \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

- If we define

$$\bar{p}_{\text{switch}}(y^n \mid s) = \bar{p}_1(y^s) \cdot \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

then total prediction error is $- \log \bar{p}_{\text{switch}}(y^n \mid s)$

- $\bar{p}_{\text{switch}}$ may be viewed both as a prediction strategy and as a distribution over infinite sequences

The Switch Distribution

- We want to predict $y_1, y_2, \ldots$ using some distribution $\bar{p}$ such that *no matter what data are observed*, i.e. for *all* $y^n \in \mathcal{Y}^n$,

$$-\log \bar{p}(y^n) \approx -\log \bar{p}_{\text{switch}}(y^n \mid \hat{s}(y^n))$$

where $\hat{s}(y^n)$ maximizes $\bar{p}_{\text{switch}}(y^n \mid s)$

- We achieve this by treating s as a parameter, putting a prior on it, and then integrating s out

(adopt a Bayesian solution to a Bayesian problem…)

The Switch Distribution

- Put "flat" prior on switch-point:

$$\pi(s) = \frac{1}{s(s+1)} \qquad\qquad -\log \pi(s) \leq 2 \log s + 1$$

- Define

$$\bar{p}_{\mathsf{switch}}(y^n) = \sum_{s \in \mathbb{N}} \pi(s) \bar{p}_{\mathsf{switch}}(y^n \mid s)$$

- Then

$$\begin{aligned} -\log \bar{p}_{\mathsf{switch}}(y^n) = -\log \sum_{s \in \mathbb{N}} \pi(s) \bar{p}_{\mathsf{switch}}(y^n \mid s) \leq \\ -\log \bar{p}_{\mathsf{switch}}(y^n \mid \widehat{s}(y^n)) - \log \pi(\widehat{s}(y^n)) \leq \\ -\log \bar{p}_{\mathsf{switch}}(y^n \mid \widehat{s}(y^n)) + 2 \log \widehat{s}(y^n) + 1 \end{aligned}$$

The Switch Distribution

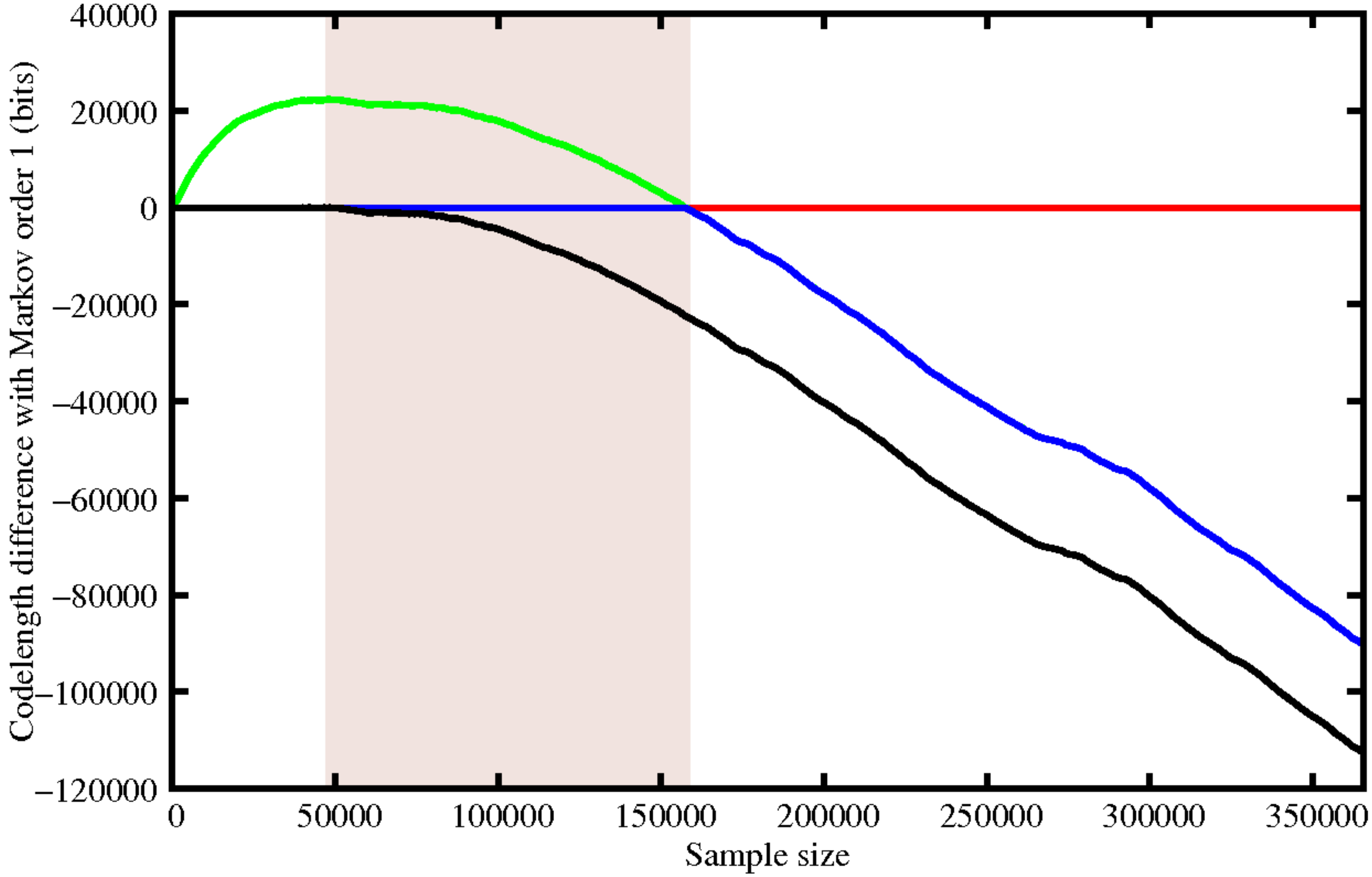
The switch distribution gains substantially over Bayes factor at a negligible price!

$$-\log \bar{p}_{\text{switch}}(y^n) \leq$$

$$-\log \bar{p}_{\text{switch}}(y^n \mid \hat{s}(y^n)) + 2\log(\hat{s}(y^n) + 1)$$

Markov: gain 20000 bits over p_{Bayes}

lose $2 \log 50001 < 32$



40000
20000
0
−20000
−40000
−60000
−80000
−100000
−120000
Codelength difference with Markov order 1 (bits)
0
50000
100000
150000
200000
250000
300000
350000
Sample size

Menu

1. Bayes Factor Model Selection

2. The Catch-Up Phenomenon

3. Solving the AIC-BIC Dilemma
   - Multi-Switch Distribution
   - Switching is consistent (Theorem 1)
   - Switching converges fast (Theorem 2)
   - Discussion

More than 2 Models

- Switch-distribution for 2 models:
  - Even in worst-case, we never lose more than 1 bit compared to standard Bayesian model averaging
  - In favourable case, we win substantially, but gain remains bounded as n increases

More than 2 Models

- Switch-distribution for 2 models:
  - Even in worst-case, we never lose more than 1 bit compared to standard Bayesian model averaging
  - In favourable case, we win substantially, but gain remains bounded as n increases
- Switch-distribution for infinite number of models:
  - Gain over Bayes increases every time we switch
  - If we keep selecting more complex models as n increases, we win infinitely many bits compared to Bayes!
  - i.e. in the case where AIC outperforms Bayes, we also outperform Bayes when doing prediction; **and also when doing estimation**

Multi-Switch Distribution

- m : number of times you switch
- $\mathbf{t} = (1, t_1, \ldots, t_m)$: "switch points"
  (sample sizes at which you switch)
- $\mathbf{k} = (k_0, k_1, \ldots, k_m)$: models you switch to
- Define $\bar{p}_{\mathbf{t},\mathbf{k}}$ as:

Multi-Switch Distribution

- m : number of times you switch
- $\mathbf{t} = (1, t_1, \ldots, t_m)$: "switch points"
  (sample sizes at which you switch)
- $\mathbf{k} = (k_0, k_1, \ldots, k_m)$: models you switch to
- Define $\bar{p}_{\mathbf{t},\mathbf{k}}$ as:
  for $1 \leq n < t_1$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_0}(y_n \mid y^{n-1})$

Multi-Switch Distribution

- m : number of times you switch
- $\mathbf{t} = (1, t_1, \ldots, t_m)$: "switch points"

  (sample sizes at which you switch)
- $\mathbf{k} = (k_0, k_1, \ldots, k_m)$: models you switch to
- Define $\bar{p}_{\mathbf{t},\mathbf{k}}$ as:

  for $1 \leq n < t_1$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_0}(y_n \mid y^{n-1})$

  for $t_1 \leq n < t_2$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_1}(y_n \mid y^{n-1})$

Multi-Switch Distribution

- m : number of times you switch
- $\mathbf{t} = (1, t_1, \ldots, t_m)$: "switch points"
  (sample sizes at which you switch)
- $\mathbf{k} = (k_0, k_1, \ldots, k_m)$: models you switch to
- Define $\bar{p}_{\mathbf{t},\mathbf{k}}$ as:

for $1 \leq n < t_1$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_0}(y_n \mid y^{n-1})$

for $t_1 \leq n < t_2$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_1}(y_n \mid y^{n-1})$

for $t_2 \leq n < t_3$: $\bar{p}_{\mathbf{t},\mathbf{k}}(y_n \mid y^{n-1}) = \bar{p}_{k_2}(y_n \mid y^{n-1})$

…and so on

$$\bar{p}_{\mathbf{t},\mathbf{k}}(y^n) := \prod_{i=1}^{n} \bar{p}_{\mathbf{t},\mathbf{k}}(y_i \mid y^{i-1})$$

Multi-Switch Distribution

$$\bar{p}_{\mathbf{t},\mathbf{k}}(y^n) := \prod_{i=1}^{n} \bar{p}_{\mathbf{t},\mathbf{k}}(y_i \mid y^{i-1})$$

may be thought of both as a sequential prediction strategy and as defining a likelihood under "meta-model" with "parameters" $(\mathbf{t}, \mathbf{k})$

$$-\log \bar{p}_{\mathbf{t},\mathbf{k}}(y^n)$$

is the accumulated prediction error you make when you switch to k_1 at $n = t_1$, to k_2 at $n = t_2$, etc.

Multi-Switch Distribution

- Put prior v on all $(\mathbf{t}, \mathbf{k})$ of each dimension as follows:
- For $\mathbf{t}, \mathbf{k} \in \mathbb{N}^{m+1}$, set

$$v(\mathbf{t}, \mathbf{k} \mid m) = w(k_0) \cdot \prod_{j=1}^{m} w(k_j) w(t_j \mid t_j > t_{j-1})$$

where $w(n) = \dfrac{1}{n(n+1)}$

- Set $v(m) = 2^{-m-1}$, $v(\mathbf{t}, \mathbf{k}) = v(\mathbf{t}, \mathbf{k} \mid m) v(m)$

- Define $\bar{p}_{\text{switch}}(y^n) = \sum_{\mathbf{t}, \mathbf{k}} v(\mathbf{t}, \mathbf{k}) p_{\mathbf{t}, \mathbf{k}}(y^n)$

Model Selection by Switching

- Use Bayes' theorem to define "posterior"

$$\bar{p}_{\text{switch}}(\mathbf{t}, \mathbf{k} \mid y^n) := \frac{v(\mathbf{t}, \mathbf{k}) p_{\mathbf{t},\mathbf{k}}(y^n)}{\sum_{\mathbf{t}',\mathbf{k}'} v(\mathbf{t}', \mathbf{k}') p_{\mathbf{t}',\mathbf{k}'}(y^n)}$$

- Define

$$\bar{p}_{\text{switch}}(k^* \mid y^n) = \sum_{m \geq 0, \mathbf{t}, \mathbf{k} \in \mathbb{N}^{n+1}, k_m = k^*} \bar{p}_{\text{switch}}(\mathbf{t}, \mathbf{k} \mid y^n)$$

- Define the switch method for model selection as: $\widehat{k}_{\text{switch}}(y^n)$ is the k^* maximizing $\bar{p}_{\text{switch}}(k^* \mid y^n)$

Switching is Consistent

- "Theorem": Bayes consistent → Switching consistent

  Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be a sequence of models as before.

  Let $\hat{k}_{\text{Bayes}}$ be Bayesian model selection, defined for priors $\pi, w_1, w_2, \ldots$ with, for all k, $\pi(k) > 0$ and for all $\theta \in \Theta_k$, $w_k(\theta) > 0$ and $w_k(\theta)$ continuous.

  Let $p^* \in \mathcal{M}_{k^*}$ for some $k^* \in \mathbb{N}$ . .

If, with p^*-probability 1, $\lim_{n \to \infty} \hat{k}_{\text{Bayes}}(Y^n) = k^*$

then, with p^*-probability 1, $\lim_{n \to \infty} \hat{k}_{\text{switch}}(Y^n) = k^*$

Rate-of-Convergence

- A model selection/averaging method together with an estimation method within each model induces a combined estimator/predictor $\bar{p}_{|y^n}$
  1. e.g. first use AIC to choose model k, then use maximum likelihood estimator $\widehat{\theta}_k^{\text{ml}}$ within model:

$$\bar{p}_{|y^n} := p_{\widehat{\theta}^{\text{ml}}_{\widehat{k}_{\text{AIC}}(y^n)}}(y^n)$$

Rate-of-Convergence

- A model selection/averaging method together with an estimation method within each model induces a combined estimator/predictor $\bar{p}_{|y^n}$

1. e.g. first use AIC to choose model k, then use maximum likelihood estimator $\widehat{\theta}_k^{\text{ml}}$ within model:

$$\bar{p}_{|y^n} := p_{\widehat{\theta}^{\text{ml}}_{\widehat{k}_{\text{AIC}}(y^n)}}(y^n)$$

2. …or use Bayesian model averaging:

$$\bar{p}_{|y^n} := \sum_k p(\cdot \mid y^n, \mathcal{M}_k) p(\mathcal{M}_k | y^n)$$

Rate-of-Convergence

- A model selection/averaging method together with an estimation method within each model induces a combined estimator/predictor $\bar{p}_{|y^n}$
  1. e.g. first use AIC to choose model k, then use maximum likelihood estimator $\hat{\theta}_k^{\text{ml}}$ within model:

  $$\bar{p}_{|y^n} := p_{\hat{\theta}^{\text{ml}}_{\hat{k}_{\text{AIC}}(y^n)}}(y^n)$$

  2. …or use Bayesian model averaging:

  $$\bar{p}_{|y^n} := \sum_k p(\cdot \mid y^n, \mathcal{M}_k) p(\mathcal{M}_k | y^n)$$

  3. …or use our Switch Distribution as defined before:

  $$\bar{p}_{|y^n} := p_{\text{switch}}(Y_{n+1} = \cdot \mid y^n)$$

Rate-of-Convergence

- The **risk** is the expected distance between 'true' p^* and estimate $\bar{p}_{|y^n}$:

$$\text{risk}_n(p^*, \bar{p}) = E_{Y^{n-1} \sim p^*}[D(p^*, \bar{p}_{|Y^{n-1}})]$$

- Here D is some fixed distance/divergence measure
  - Here: KL divergence (Hellinger^2 distance also works)

Switching achieves Minimax Rate

- Let $\mathcal{M}^* \subset \left\{ p^* : \inf_{q \in \bigcup_{k \geq 1} \mathcal{M}_k} D(p^*, q) = 0 \right\}$
- **"Theorem 2":** Under variety of conditions:

$$\frac{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{\text{switch}})}{\inf_{\bar{p}} \sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p})} \rightarrow \text{something finite}$$

- Examples:
  - histogram/spline density estimation, $\mathcal{M}^*$ is class of smooth densities with *r* bounded derivatives
  - nonparametric linear regression
- Typically convergence rate is $\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{\text{switch}}) \asymp n^{-\gamma}$ for some $0 < \gamma < 1$

Switching achieves Minimax Rate

- Let $\mathcal{M}^* \subset \left\{ p^* : \inf_{q \in \bigcup_{k \geq 1} \mathcal{M}_k} D(p^*, q) = 0 \right\}$
- **"Theorem 2":** Under variety of conditions:

$$\frac{\sup_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(p^*, \bar{p}_{\text{switch}})}{\inf_{\bar{p}} \sup_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(p^*, \bar{p})} \rightarrow \text{something finite}$$

- Examples:
  - histogram/spline density estimation, $\mathcal{M}^*$ is class of smooth densities with *r* bounded derivatives
  - nonparametric linear regression
- Typically convergence rate is $\sup_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} R_i(p^*, \bar{p}_{\text{switch}}) \asymp n^{1-\gamma}$ for some $0 < \gamma < 1$

Switch-distribution converges fast

- The Upshot:

The Switch-distribution essentially converges at least as fast **as any other method at all**, in particular, as fast as AIC/leave-one-out CV

The AIC-BIC Dilemma

- AIC-group converges faster when $p^* \notin \mathcal{M}$ but can be arbitrarily well-approximated by $p_1, p_2, \ldots \in \mathcal{M}$
- BIC-group performs better (is consistent) when $p^* \in \mathcal{M}$
- **In "typical" situations switch-distribution achieves both!**

...both in **theory** and in **practice**

Computational Complexity

- Is switching computationally efficient?
- Answer is YES … Time complexity $O(n \cdot k_{\max})$
  - (usually) comparable to AIC and BIC
  - Algorithm similar to "fixed share" (Herbster & Warmuth 98), , developed in ***tracking the best expert*** literature
  - optimal model for prediction at sample size *n* may be viewed as **hidden state in a Hidden Markov Model**
  - use forward algorithm

De Rooij and Koolen, COLT 2008, tomorrow 5.15 PM

(Potential) Applications

- Nonparametric density estimation (work in progress)
  - variable-width histograms, splines, kernel density estimation
- Time Series Prediction
- Regression (challenge: subset selection)
- ....

"Bayesian"?

- Formally, our procedure is Bayesian
- But a real subjective Bayesian would probably not use the switch-distribution
  - It corresponds (…) to a belief that data "follow" $\mathcal{M}_1$ until some critical s, and afterwards, they follow $\mathcal{M}_2$
  - But we certainly do not believe this! If anything, we believe that **all** $y_1, y_2, \ldots$ follow the **same** $\mathcal{M}_k$ …
  - Nevertheless, because of the catch-up phenomenon, we get better predictions and estimations if we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at some point, under some conditions

Subjective Bayesian Objections

- GIGO (Garbage In, Garbage Out)
  - If model and priors are "correct", predicting according to standard Bayesian predictive distribution must be optimal
  - "…so instead of the switch distribution on a bad model, should use standard Bayes on good model"

Subjective Bayesian Objections

- GIGO (Garbage In, Garbage Out)
  - –If model and priors are “correct”, predicting according to standard Bayesian predictive distribution must be optimal
  - –“…so instead of the switch distribution on a bad model, should use standard Bayes on good model”

Wrong!

Subjective Bayesian Objections

- GIGO (Garbage In, Garbage Out)
  - –If model and priors are “correct”, predicting according to standard Bayesian predictive distribution must be optimal
  - –“…so instead of the switch distribution on a bad model, should use standard Bayes on good model”

Wrong!

- A Better Bayesian Objection:
  - –if you think that data come from distribution that is not in any of the $\mathcal{M}_k$, but rather in their closure, you have a “nonparametric belief” and should use a nonparametric prior rather than the hierarchical parametric prior used here!
  - –True; but in fact we can think of our approach as using Bayes with a very unusual type of nonparametric prior!

It's MDL, Jim, but not as we know it!

- Bayesian interpretation of $\bar{p}_{\mathsf{switch}}$ is tenuous
- Yet $\bar{p}_{\mathsf{switch}}$ makes eminent sense from
  1. Dawid's **prequential**…
  2. Rissanen's **MDL**…
  3. **Universal prediction**… point of views
  – We are trying to predict/estimate as well as the best sequence of models, rather than the best single model
- Nevertheless, apparently nobody in MDL field has ever thought of explicitly coding switch points before

Thank you for your attention!

Paper is on my webpage, www.grunwald.nl

Shameless plug:

For more on MDL and “prequential” ideas, see my book

The Minimum Description Length Principle

MIT Press 2007