CWI

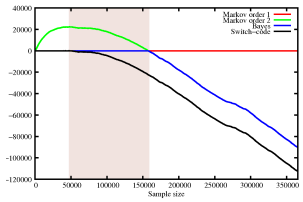
Catching Up Faster by Switching Sooner

Tim van Erven, Peter Grünwald, Steven de Rooij
CWI, Amsterdam

CWI is the national research center for mathematics and computer science in the Netherlands

In a Nutshell

- Bayesian Model Selection and Model Averaging often converge at a suboptimal rate in sequential prediction



In a Nutshell

- Bayesian Model Selection and Model Averaging often converge at a suboptimal rate in sequential prediction
- We identify the **catch-up phenomenon** as a novel, prequential explanation of the slow convergence

In a Nutshell

- Bayesian Model Selection and Model Averaging often converge at a suboptimal rate in sequential prediction
- We identify the **catch-up phenomenon** as a novel, prequential explanation of the slow convergence
- Based on the analysis, we propose the switch distribution, a modification of Bayesian Model Averaging/Selection that "catches up faster"

Menu

1. Bayesian Model Selection/Averaging
2. Catch-up phenomenon
3. Switch Distribution
4. **Main Theorem:**
   switch distribution converges at the optimal rate in parametric and nonparametric settings
5. Optimal prediction implies optimal estimation (to some extent)

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, $\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k \subseteq \mathbb{R}^{d_k}\}$ is parametric 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
  $$\hat{k} : \bigcup_{n \geq 1} \mathcal{Y}^n \to \mathbb{N}$$
  is a function mapping data sequences of arbitrary length to model indices
  - $\hat{k}(y^n)$ is model chosen for data y^n

Examples of Model Selection Methods

- Akaike's Information Criterion (AIC, 1973)

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

- Bayesian Information Criterion (BIC, 1978)

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

- Bayes factor model selection, DIC, Cross-Validation, L_1-methods, …

Bayes Factor Model Selection

$\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}$ $\quad k \in \mathcal{K} \subset \mathbb{N}$

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

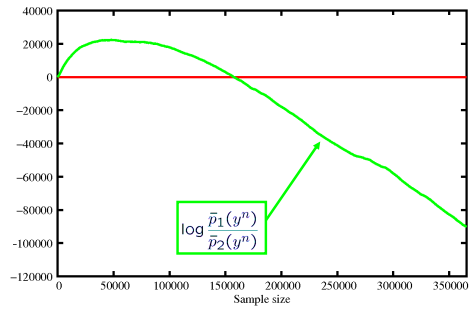
$$p(\mathcal{M}_k \mid y^n) = \frac{p(y^n \mid \mathcal{M}_k)\pi(k)}{\sum_{k \in \mathcal{K}} p(y_n \mid \mathcal{M}_k)\pi(k)}$$

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

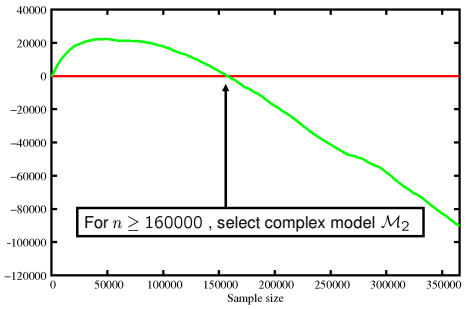
$\pi(k)$ is prior

$w_1, w_2, \ldots$ are priors

$\hat{k}(y^n)$ is k 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 $n \approx n_{\text{switch}}$….
  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. The Switch Distribution

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 cumulative loss satisfies

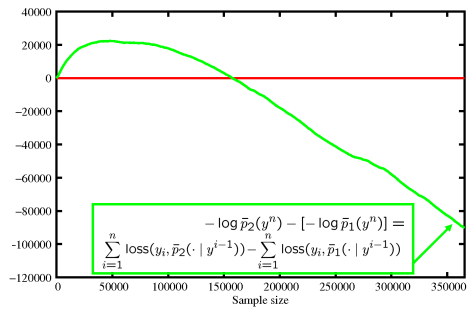
$$\sum_{i=1}^n \text{loss}(y_i, p(\cdot \mid y^{i-1})) = \sum_{i=1}^n \left[-\log p(y_i \mid y^{i-1})\right]$$
$$= -\log \prod_{i=1}^n p(y_i \mid y_1, \ldots, y_{i-1}) = -\log \prod_{i=1}^n \frac{p(y^i)}{p(y^{i-1})}$$
$$= -\log p(y_1, \ldots, y_n)$$

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

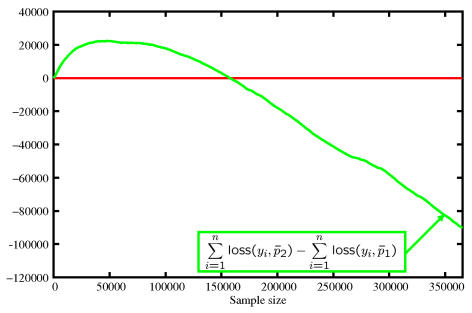
The Most Important Slide

- 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 cumulative sequential prediction error

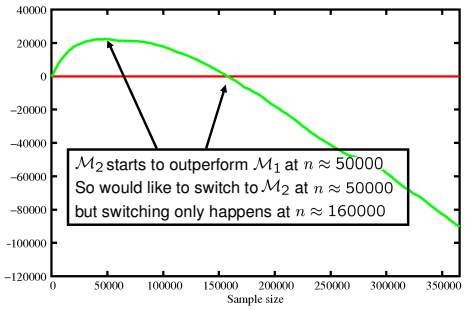
  Dawid '84, Rissanen '84



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



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

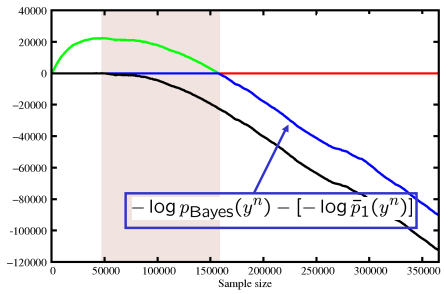


$$\sum_{i=1}^{50000} \text{loss}(y_i, \bar{p}_2) - \sum_{i=1}^{50000} \text{loss}(y_i, \bar{p}_1) \approx 20000$$

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

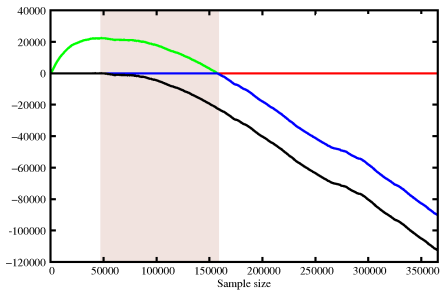
The Catch-Up Phenomenon

- 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!

Why try to modify Bayes?

- Frequentist Objection: in your example, other methods work fine. Why not use those?
  - e.g. model selection by leave-one-out cross-validation, prediction by ML within chosen model works fine here

Why try to modify Bayes?

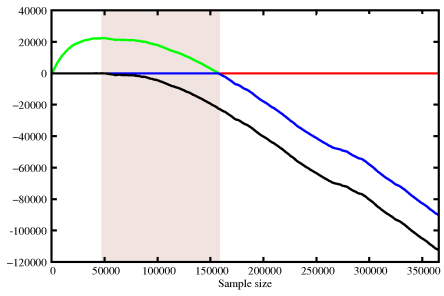
- Frequentist Objection: in your example, other methods work fine. Why not use those?
  - e.g. model selection by leave-one-out cross-validation, prediction by ML within chosen model works fine here
- Answer: our method
  - Retains nice properties of Bayes (consistency, prequential useability, ease of incorporationg prior information)
  - can be proven to be "optimal" under very mild conditions

Why try to modify Bayes - II?

- Bayesian Objection: GIGO (Garbage In, Garbage Out)
  - A. The only coherent statistical approach is the Bayesian. Thus, if large catch-up phenomenon occurs, this can only mean that our **models and/or priors are wrong**
  - B. So come up with better model, rather than modify method!

Why try to modify Bayes - II?

- Bayesian Objection: GIGO (Garbage In, Garbage Out)
  - A. The only coherent statistical approach is the Bayesian. Thus, if large catch-up phenomenon occurs, this can only mean that our **models and/or priors are wrong**
  - B. So come up with better model, rather than modify method!
- Answer:
  - As to A: it is true that a large catch-up phenomenon indicates model/prior misspecificatoin
  - As to B: of course we should try....but in practice we often fail...and then the switch distribution will really help!



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:

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

- Define

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

- Then

$$-\log \bar{p}_{\text{switch}}(y^n) = -\log \sum_{s \in \mathbb{N}} \pi(s) \bar{p}_{\text{switch}}(y^n \mid s) \leq$$
$$-\log \bar{p}_{\text{switch}}(y^n \mid \hat{s}(y^n)) - \log \pi(\hat{s}(y^n)) \leq$$
$$-\log \bar{p}_{\text{switch}}(y^n \mid \hat{s}(y^n)) + 2 \log \hat{s}(y^n) + 1$$

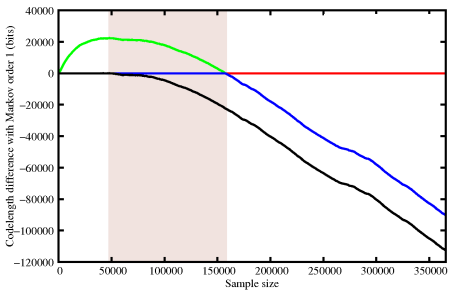
The Switch Distribution

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 over p_{Bayes}

lose $2 \log 50001 < 32$



"Bayesian"?

- Formally, our procedure is still 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 if we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at some point

More than 2 Models

- Switch-distribution for 2 models:
  - Even in worst-case, we never lose more than 1 loss-unit compared to standard Bayesian model averaging in sequential prediction
  - 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 loss-unit compared to standard Bayesian model averaging in sequential prediction
  - In favourable case, we win substantially, but gain remains bounded as n increases
- Switch-distribution for countably infinite nr of models:
  - Gain over Bayes increases every time we switch
  - If we keep selecting more complex models as n increases, we win unboundedly compared to Bayes!

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
- Meta-prediction strategy $\bar{p}_{\mathbf{t},\mathbf{k}}$:
  - Predict with k_0 until sample size t_1
  - Predict with k_1 from sample size t_1 to t_2
  - ....
  - From sample size t_m onwards, predict with k_m

Multi-Switch Distribution

- Put special prior v on all $(\mathbf{t}, \mathbf{k})$ of each dimension
- Define

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

- …and use this for sequential prediction.
- This is Bayesian model averaging with prior on sequences of models, rather than on single models

Does it work?

Cumulative Risk (i.i.d. case)

$$R_n(p^*, \bar{p}) := E_{Y^n \sim p^*}\left[\sum_{i=1}^n \mathsf{loss}(Y_i, \bar{p}_{|Y^{i-1}}) - \sum_{i=1}^n \mathsf{loss}(Y_i, p^*)\right]$$

conditional distribution of Y_i given Y^{i-1} according to prediction strategy/distribution $\bar{p}$

Cumulative Risk (i.i.d. case)

$$R_n(p^*, \bar{p}) := E_{Y^n \sim p^*}\left[\sum_{i=1}^n \mathsf{loss}(Y_i, \bar{p}_{|Y^{i-1}}) - \sum_{i=1}^n \mathsf{loss}(Y_i, p^*)\right]$$

conditional distribution of Y_i given Y^{i-1} according to prediction strategy/distribution $\bar{p}$

$R_n(p^*, \bar{p}) \geq 0$; $R_n(p^*, p^*) = 0$

Optimal Parametric Rate

- Let $\mathcal{M}$ be a parametric model with d parameters. Under mild conditions

$$\min_{\bar{p}} \max_{p^* \in \mathcal{M}} R_n(p^*, \bar{p}) = \frac{d}{2} \log n + O(1)$$

Optimal Nonparametric Rate

- Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be nested parametric models, and let $\mathcal{M}^*$ be a "smooth" subset of

$$\left\{ p^* : \inf_{q \in \bigcup_{k \geq 1} \mathcal{M}_k} D(p^* \| q) = 0 \right\}$$

then "typically" for some $0 < \gamma < 1/2$

$$\min_{\bar{p}} \max_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}) \asymp n^{\gamma}$$

Examples:
histogram density estimation, $\mathcal{M}^*$class of α-smooth densities for unknown α
nonparametric linear regression with random design, Gaussian noise, $E[Y \mid X] = f(X)$, f in Besov-type space

Main Result

- Suppose $\hat{k}$ is a model selection criterion that, for all n, each sample of size n, selects a model in a set $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{k(n)}\}$
that grows at most polynomially in n, e.g. $k(n) = n^{10}$

Main Result

- Suppose $\hat{k}$ is a model selection criterion that, for all n, each sample of size n, selects a model in a set $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{k(n)}\}$
that grows at most polynomially in n
- Suppose $\mathcal{M}^* = \mathcal{M}_{k^*}$ for some k^* (parametric case)
or $\frac{(\log n)^{2.01}}{\inf_{\bar{p}} \sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p})} \to 0$ (nonparametric case)

Main Result

- Suppose $\hat{k}$ is a model selection criterion that, for all n, each sample of size n, selects a model in a set $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{k(n)}\}$
that grows at most polynomially in n
- Suppose $\mathcal{M}^* = \mathcal{M}_{k^*}$ for some k^* (parametric case)
or $\frac{(\log n)^{2.01}}{\inf_{\bar{p}} \sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p})} \to 0$ (nonparametric case)

Then: $\limsup_{n \to \infty} \frac{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{\text{switch}})}{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{\hat{k}(y^{n-1})})} \leq 1$

Main Result, Boldly Stated

- Suppose data are i.i.d. $\sim p^*$
- We want to predict outcomes sequentially based on parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$. Then:

Switch Distribution Achieves Optimal Cumulative Log Loss Rate under **hardly any** conditions on p^* and $\mathcal{M}_1, \mathcal{M}_2, \ldots$

Caveats:
- only works for slower (by factor n) version of switch distr.
- We are comparing model averaging to model selection

Proof Idea, Nonparametric Case

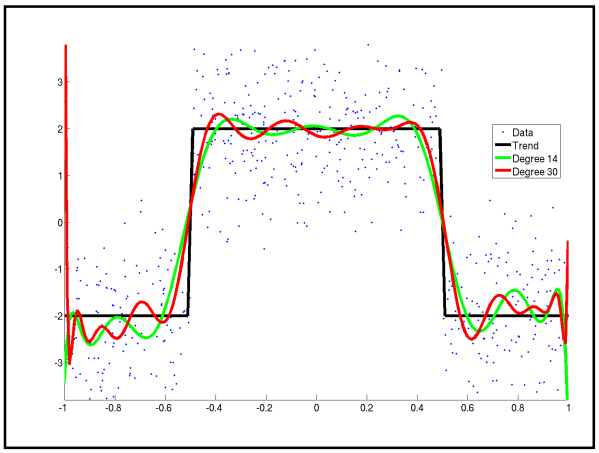
- For every (arbitrary) model selection criterion $\hat{k}$ and every p^* there exists "lazy" version $\hat{k}_{\text{lazy}}$ that is only allowed to switch at $n = 2, 4, 8, 16, \ldots$and with:

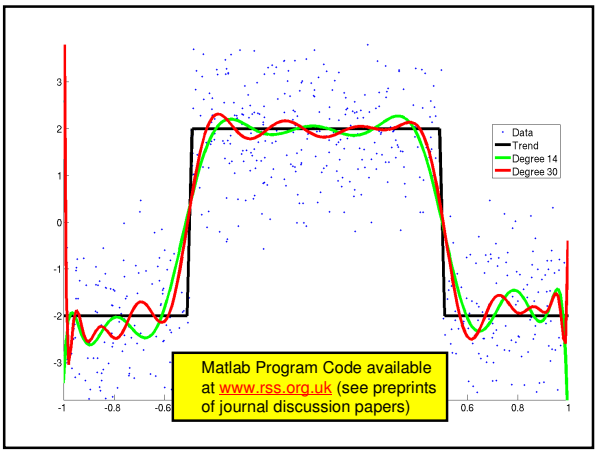
$$R_n(p^*, \bar{p}_{\hat{k}_{\text{lazy}}}) \leq 2 \cdot R_n(p^*, \bar{p}_{\hat{k}})$$

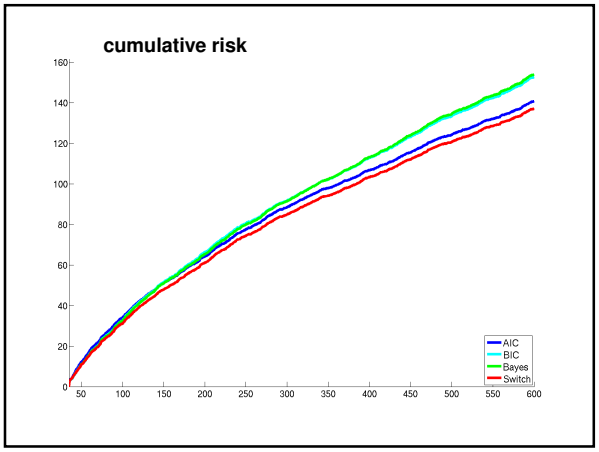
- Note that $\hat{k}_{\text{lazy}}$ depends on true p^* but this is o.k.
- Prior mass of sequence (t, k) chosen by $\hat{k}_{\text{lazy}}$ satisfies

$-\log \pi(t, k) = O((\log n)^2)$ [$\log n$ switches to $\log n^{\tau}$ models]
so

$$R_n(p^*, \bar{p}_{\text{switch}}) = R_n(p^*, \bar{p}_{\hat{k}_{\text{lazy}}}) + O((\log n)^2)$$







Switching, AIC and BIC

	cumulative risk rate		instantaneous risk rate		consistent
	parametric	non-parametric	parametric	non-parametric	
AIC LOOCV	optimal	optimal			
BIC BayesMS	optimal	suboptimal			
Switch (slow v.)	optimal	optimal			

under very weak conditions and "prequentially": our main advantages

In many cases we expect switch to do better than Bayes but we have no means to formally state/prove this

Switching, AIC and BIC

	cumulative risk rate		instantaneous risk rate		consistent
	parametric	non-parametric	parametric	non-parametric	
AIC LOOCV	optimal	optimal			
BIC BayesMS	optimal (well...)	suboptimal			
Switch (slow v.)	optimal	optimal			

under very weak conditions and "prequentially": our main advantages

Switching, AIC and BIC

	cumulative risk rate		instantaneous risk rate		consistent
	parametric	non-parametric	parametric	non-parametric	
AIC LOOCV	optimal	optimal			No
BIC BayesMS	optimal (well...)	suboptimal			Yes
Switch (slow v.)	optimal	optimal			**Yes**

Estimation and "Standard" Risk

- The **instantaneous risk** is 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.
- Remarkably, if we take KL divergence we have

$$R_n(p^*, \bar{p}) = \sum_{i=1}^{n} \text{risk}_i(p^*, \bar{p}).$$

- Small Cumulative Risk implies Small Individual Risk at "most" sample sizes

Switching, AIC and BIC

	cumulative risk rate		instantaneous risk rate		consistent
	parametric	non-parametric	parametric	non-parametric	
AIC LOOCV	optimal	optimal	optimal	optimal	No
BIC BayesMS	optimal (well...)	suboptimal	suboptimal	suboptimal	Yes
Switch (slow v.)	Optimal	optimal	suboptimal **YANG**	"optimal" CESARO	Yes

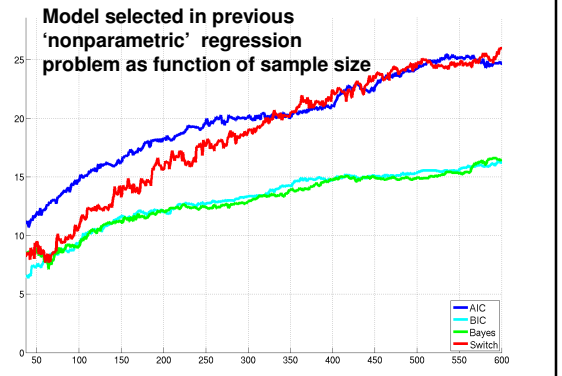
Thank you for your attention!

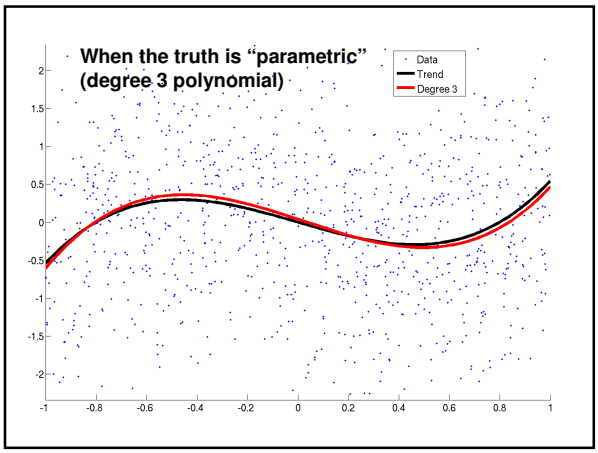
Extra Slides

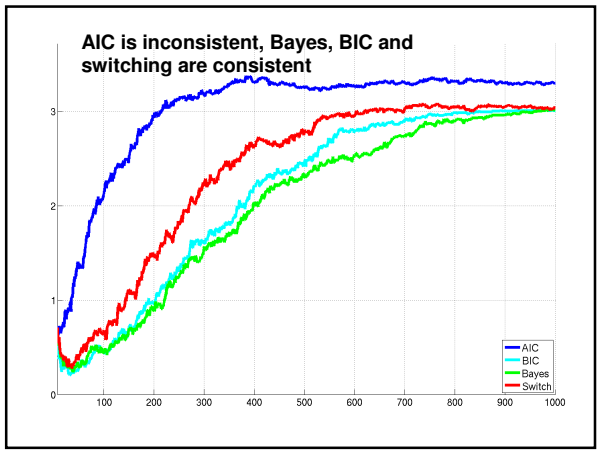
Model Selection by Switching

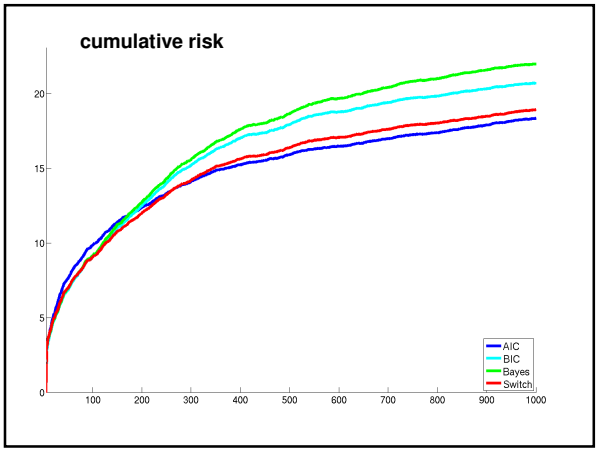
- Define $\bar{p}_{\text{switch}}(k^* \mid y^n) = \sum \bar{p}_{\text{switch}}(\mathrm{t}, \mathrm{k} \mid y^n)$

  where sum is over all structures (t, k) that say "use k^* to predict Y_{n+1} and never switch again"
- Define the switch method for model selection as: $\hat{k}_{\text{switch}}(y^n)$ is the k^* maximizing $\bar{p}_{\text{switch}}(k^* \mid y^n)$









"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** $y_1, y_2, \cdots$ 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

"Bayesian"? - II

- Indeed, let $p_{\text{Bayes}}(y^n) = \frac{1}{2}p(y^n \mid \mathcal{M}_1) + \frac{1}{2}p(y^n \mid \mathcal{M}_2)$
- We have
  $E_{P_{\text{Bayes}}}[-\log \bar{p}_{\text{switch}}(Y^n)] > E_{P_{\text{Bayes}}}[-\log p_{\text{Bayes}}(Y^n)]$

  so if $\bar{p}_{\text{Bayes}}$ really describes your subjective beliefs, you should predict by $\bar{p}_{\text{Bayes}}$, not $\bar{p}_{\text{switch}}$

No Hyperprediction

- Moreover ("no-hyperprediction inequality", Grü07), for all n, all K :

$$P_{\text{Bayes}}\left(-\log \bar{p}_{\text{switch}}(Y^n) \leq -\log p_{\text{Bayes}}(Y^n) - K\right) \leq 2^{-K}$$

- If we are serious about our prior, we strongly believe that no substantial catch-up phenomenon will occur
- Still, in practice, it does
  - Pragmatic, rather than subjective, prior is used
  - Models are wrong ($K = 20000(!)$)

Details on Defining Cumulative Convergence Rate in Main Theorem

- 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:

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

- A model selection/averaging method together with an estimation/averaging 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:

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

- 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:

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

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
  - try out with our software at RSS site

De Rooij and Koolen, COLT 2008

"No Smoothness Assumptions Needed"

- Pleasant Property of Switch-Distribution as compared to many other model selection methods:
- No weighting factors λ_n whose optimal values may depend on smoothness assumptions about the underlying density/distributions
  - In fact, adaptation happens entirely automatically: proof of cumulative risk theorem based on **individual-sequence result** for **every** sequence $y^n = y_1, \ldots, y_n$ we have: if y^n is predicted well by a sequence of distributions in $\bigcup_k \mathcal{M}_k$ then it is also predicted well by the switch distribution!