Bayesian Inference
successes and problems

CWI Peter Grünwald

Universiteit Leiden

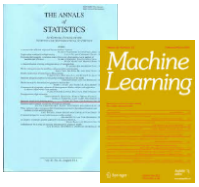
Centrum Wiskunde & Informatica – Amsterdam
Mathematical Institute – Leiden University

Joint work with Thijs van Ommen, Rianne de Heide, Wouter Koolen, Tim van Erven, Nishant Mehta - Preprint on arXiv

CWI

Context

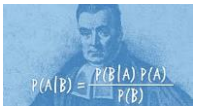
- Around 30% of all statistics papers are "Bayesian"
- Before "deep learning revolution" (2011), Bayes was also the trend in machine learning



CWI

Context

- Bayesian statistics is actually much older than 'standard' statistics
  - Bayes/**Laplace** (+- 1800)



CWI

Context

- Bayesian statistics is actually much older than 'standard' statistics
  - Bayes/**Laplace** (+- 1800)
  - Heavily criticized from 1850s onwards
  - Basics of 'standard' statistics developed between 1900-1940 (Pearson Sr,Jr, Student, Fisher, Neyman)
  - Slow comeback since 1950s, accelerated in 1990s

  (fast computers ⟶ it could finally be applied!)

CWI

Context

- According to many, **ideally** all statistics/learning from data should be done in a Bayesian manner
  - "All rational methods for learning are equivalent to a Bayesian method"

  (Ramsey ('23), De Finetti ('37), Cox ('46), Savage ('54), Jefffreys ('61))

  - **Your only excuse not to be 'Bayesian' is that it might be computationally too intensive**

CWI

Context

- According to many, **ideally** all statistics/learning from data should be done in a Bayesian manner
  - "All rational methods for learning are equivalent to a Bayesian method"

  (Ramsey ('23), De Finetti ('37), Cox ('46), Savage ('54), Jefffreys ('61))

  - **Your only excuse not to be 'Bayesian' is that it might be computationally too intensive**
  - **....or is it?**

CWI

Menu

1. A Problem for Bayes when **Model is Wrong**
2. The Learning Rate & The Safe Bayesian

CWI

Menu

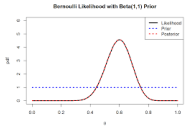
0. Bayesian Inference
   – Bayes theorem
   – Two ways of doing Regression with Bayes:
   Ridge + Bayes Factor Model Sel/Averaging
1. A Problem for Bayes when **Model is Wrong**
2. The Learning Rate & The Safe Bayesian

Bayesian Inference

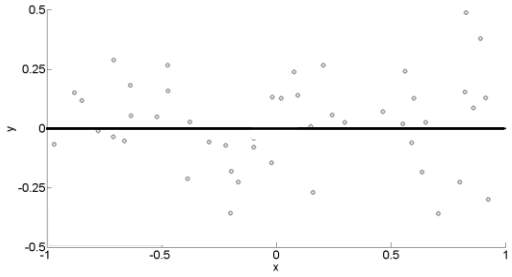
User supplies:

- Statistical model $\mathcal{M} = \{p_\theta \mid \theta \in \Theta\}$
- Prior probability density $w(\theta)$
- Bayes theorem gives the posterior

$$p(\theta \mid y^n) = \frac{p_\theta(y^n) \cdot w(\theta)}{\int_{\theta \in \Theta} p_\theta(y_n) w(\theta) d\theta}$$



CWI



Bayesian Linear Regression Model

Raftery et al. '96

- Model $\mathcal{M}_k = \{p_{\vec{\beta},\sigma^2} \mid \sigma^2 \in \mathbb{R}^+, \vec{\beta} \in \mathbb{R}^{k+1}\}$

  expresses $Y = \sum_{j=0}^{k} \beta_j g_j(X) + \epsilon$

  where ϵ is 0-mean, σ^2 –variance Gaussian random variable

- Model assumes that data are independently sampled
- $g_j(x)$ is j-th basis function (e.g. x^j or x_j or $\sin 2\pi j x$)

Bayesian Linear Regression Model

Raftery et al. '96

- Model $\mathcal{M}_k = \{p_{\vec{\beta},\sigma^2} \mid \sigma^2 \in \mathbb{R}^+, \vec{\beta} \in \mathbb{R}^{k+1}\}$

  expresses $Y = \sum_{j=0}^{k} \beta_j g_j(X) + \epsilon$

  where ϵ is 0-mean, σ^2 –variance Gaussian random variable, extended to n outcomes by independence:

  $p_{\vec{\beta},\sigma^2}(y^n \mid x^n) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=0}^{k}\beta_j g_j(x_i))^2}$

  still need to equip with **priors** on β, σ^2

Bayesian Ridge Regression

- Model $\mathcal{M}_k = \{p_{\vec{\beta},\sigma^2} \mid \sigma^2 \in \mathbb{R}^+, \vec{\beta} \in \mathbb{R}^{k+1}\}$

$$p_{\vec{\beta},\sigma^2}(y^n \mid x^n) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \sum_{j=0}^k \beta_j g_j(x_i))^2}$$

- Take prior $\beta_0, \ldots, \beta_k$ i.i.d. $N(0, \sigma^2)$ on β
  spherical Gaussian – large absolute values severely penalized

$$p(\vec{\beta} \mid y^n, x^n, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \sum_{j=0}^k \beta_j g_j(x_i))^2 - \frac{1}{2\sigma^2}\sum_{j=0}^k \beta_j^2}$$

- Take σ^2 fixed or Inverse Gamma prior

Experiment: Polynomial Ridge Regression

- Model instantiated to $Y = \sum_{j=0}^k \beta_j X^j + \epsilon, k = 50$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
  $X_i \sim \text{Unif.}[-1,1]$, i.i.d.
  $Y_i = X_i + \epsilon_i, \epsilon_i \sim \text{Normal}(0,1)$, i.i.d.
- Note: model is (for now!) **correct**

Experiment

- Model instantiated to $Y = \sum_{j=0}^k \beta_j X^j + \epsilon, k = 50$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
  $X_i \sim \text{Unif.}[-1,1]$, i.i.d.
  $Y_i = X_i + \epsilon_i, \epsilon_i \sim \text{Normal}(0,1)$, i.i.d.
- Note: model is (for now!) **correct**
- ...and Bayes works perfectly well: posterior concentrates around $\beta = (0,1,0,\ldots,0)$ after just a few outcomes and keeps on doing so for ever

Experiment

- Model instantiated to $Y = \sum_{j=0}^k \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
  $X_i \sim \text{Unif.}[-1,1]$, i.i.d.
  $Y_i = X_i + \epsilon_i, \epsilon_i \sim \text{Normal}(0,1)$, i.i.d.
- Note: model is (for now!) **correct**
- ...and Bayes works perfectly well: posterior concentrates around $\beta = (0,1,0,\ldots,0)$ after just a few outcomes and keeps on doing so for ever
- **Least Squares/ML would perform terribly here!**

Bayesian High Dimensional Regression

Two standard approaches:

- Bayesian Ridge/Lasso/Horseshoe Regression
- Bayesian Model Selection/Model Averaging

Bayes Factor Model Selection (Hypothesis Testing)

$\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\}, k = 1, 2, \ldots$

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

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

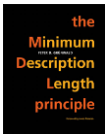
$$p(y^n \mid \mathcal{M}_k) = \int_{\theta \in \Theta_k} p_\theta(y^n)\pi(\theta \mid k)d\theta$$

$\pi(k)$ is prior **on** models

$\pi(\theta \mid 1), \pi(\theta \mid 2)$, are priors **within models**

Bayes Factor Model Selection

- Standard Bayesian method to select a model based on the data
- Can be used to select degree of polynomial
  - Bayes has a built-in **Occam's Razor:** automatic regularization
  - Complex models 'penalized' automatically (even if flat priors are used within these models)
- Close Relation to information-theoretic Minimum Description Length (**MDL**) Model Selection



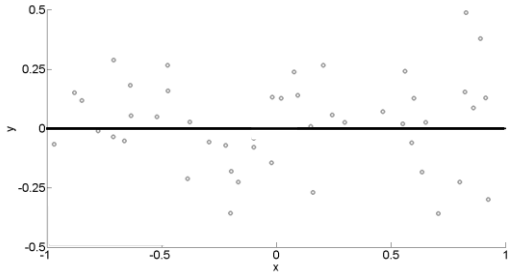
Bayes Factor Model Selection

- Standard Bayesian method to select a model based on the data
- Can be used to select degree of polynomial
  - Bayes has a built-in **Occam's Razor:** automatic regularization
- If the model is correct i.e. well-specified, then this guarantees that the 'right' degree will be selected given enough time, with probability 1 (consistency)
  - In contrast standard least squares would always select a polynomial with 0 error on the data and degree equal to number of data points -1

Experiment

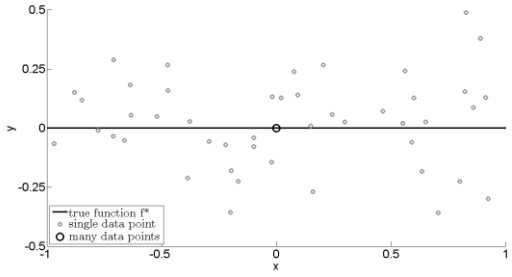
- Model instantiated to $Y = \sum_{j=0}^{k} \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
  $X_i \sim \text{Unif.}[-1, 1], \text{i.i.d.}$
  $Y_i = 0 + \epsilon_i, \epsilon_i \sim \text{Normal}(0, 1), \text{i.i.d.}$
- Note: model is (for now!) **correct**
- ...and Bayes factor model selection works perfectly well, selects 0-degree model after just a few outcomes and keeps on doing so for ever

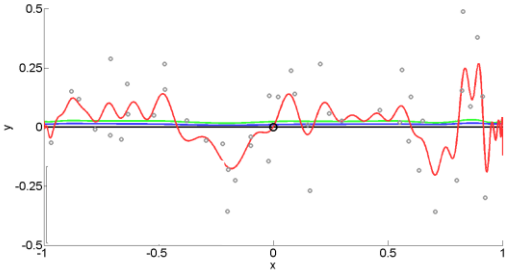
CWI

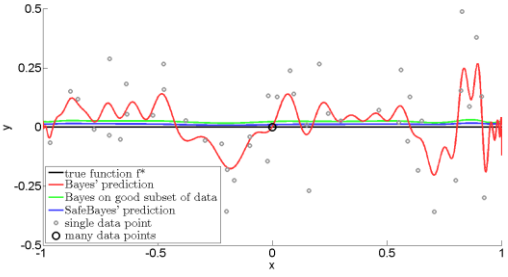


Experiment

- Model instantiated to $Y = \sum_{j=0}^{k} \beta_j X^j + \epsilon$
- Let's experiment to see what happens if data are sampled from following "true" distribution:
- At each i, we independently toss a fair coin
  - if coin lands heads, as before:
    $X_i \sim \text{Unif.}[-1, 1], \text{i.i.d.}$
    $Y_i = 0 + \epsilon_i, \epsilon_i \sim \text{Normal}(0, 1), \text{i.i.d.}$
  - if tails, we generate an **easy** example ("**in-lier**")
    $(X_i, Y_i) = (0, 0)$

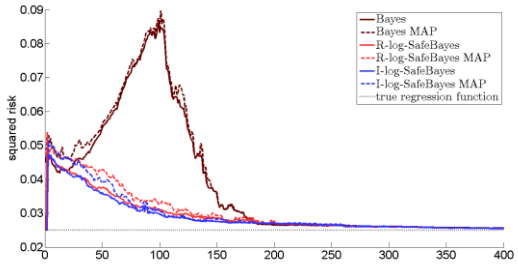






$\sigma^2 = 1/20 \to 1/40 = 0.025$

Risk Graph



Risk measured in Expected Squared Loss on a new outcome

Important Remark

- If nr of basis functions is finite, then problem does go away *at some point*
- **Real issue** is; if we take an infinite nr of basis functions (e.g. polynomials of all degree)
  - Bayes converges straight away if model correct
  - Bayes *never* converges if model contains 50% easy points

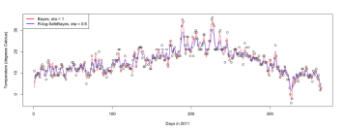
Important Remarks - II

- Problem has **nothing to do with** choice of **polynomials** as basis functions; same behaviour observed e.g. if $X_i = (X_{i1}, \ldots, X_{ik}), X_{ij}$ i.i.d. $\sim N(0,1)$
- Problem persists if we choose regression function in, say, $\mathcal{M}_4$ rather than $\mathcal{M}_0$ and if easy points not at $X = \mathbf{0}$

Problem is Real

Problem persists in **Bayesian Ridge/LASSO Setting** (with trigonometric basis functions)

- In Lasso setting we found dramatic improvements of 'Safe Bayes' over standard Bayes on some **real-world data sets** (Seattle weather, London Air Pollution)



CWI

Context

- According to many, **ideally** all statistics/learning from data should be done in a Bayesian manner
  - "All rational methods for learning are equivalent to a Bayesian method"

  (Ramsey ('23), De Finetti ('37), Cox ('46), Savage ('54), Jefffreys ('61))

  - **Your only excuse not to be 'Bayesian' is that it might be computationally too intensive**
  - **...or that you'll end up with a wrong model (unavoidable!!!)**

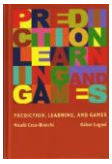
A Solution...

- Remaining 'almost' Bayesian while avoiding this type of problems..
- G. (2012) *The Safe Bayesian*
- G.& Van Ommen (2015), G. (2016)
- R. de Heide (2016),
  Safe Bayesian Regression R Package,
  CRAN *R* Repository

1. **Bayesian inference** with misspecification

$$\pi(\theta \mid \text{data}) \propto (p(\text{data} \mid \theta)) \cdot \pi(\theta)$$

2. **On-Line Sequential Prediction Literature**



$$\pi(\theta \mid \text{data}) \propto e^{-\eta \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

Freund, Schapire, Cesa-Bianchi, Lugosi, Hazan, Kale, ..., Koolen, Van Erven

1. **Bayesian inference** with misspecification

$$\pi(\theta \mid \text{data}) \propto (p(\text{data} \mid \theta)) \cdot \pi(\theta)$$

2. **On-Line Sequential Prediction**

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

Linear regression can be interpreted both ways!

1. **Bayesian inference** with misspecification

$$\pi(\theta \mid \text{data}) \propto (p(\text{data} \mid \theta)) \cdot \pi(\theta)$$

can fail **dramatically**

2. **On-Line Sequential Prediction**

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

can fail **dramatically** if η set to 1/(2variance)

1. **Bayesian inference** with misspecification

$$\pi(\theta \mid \text{data}) \propto (p(\text{data} \mid \theta)) \cdot \pi(\theta)$$

can fail **dramatically**

2. **On-Line Sequential Prediction**

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

η often set to $1/\sqrt{n}$ by theorists

1. **Bayesian inference** with misspecification

$$\pi(\theta \mid \text{data}) \propto (\, p(\text{data} \mid \theta)\,)^{\eta} \cdot \pi(\theta)$$

can fail **dramatically** with $\eta = 1$
Still o.k. for some much smaller η
But which one?
(setting η too small leads to overly slow learning)

2. **On-Line Prediction ; PAC-Bayes ; Lasso/Ridge**

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

Often set to $1/\sqrt{n}$
O.K for adversarial data, but in practice larger η usually much better! **But which ones?**

Theory

- **SafeBayes** (G. 2011, 2014) is a **way to provably learn the optimal learning rate from the data**
  - Something which *does not* work: putting a prior on the learning rate and integrating it out...
  - SafeBayes is 'a little bit' like **cross-validation** but with a **non-standard loss function**

Explaining SafeBayes via Ridge

- Frequentist (non-Bayesian) Ridge Regression:
- For each λ select $\widehat{\beta_\lambda}$ minimizing

$$\sum_{i=1}^{n} (y_i - \sum_{j=0}^{k} \beta_j g_j(x_i))^2 + \lambda \sum_{j=0}^{k} \beta_j^2$$

- Select final $\hat{\beta} = \hat{\beta}_{\lambda_{cv}}$ by cross-validating over λ
- $\hat{\beta}_\lambda$ is also the posterior mean/MAP of Bayesian ridge regression with fixed variance $\sigma^2 = 2\lambda$

Explaining SafeBayes via Ridge

- For each λ select $\widehat{\beta_\lambda}$ minimizing

$$\sum_{i=1}^{n} (y_i - \sum_{j=0}^{k} \beta_j g_j(x_i))^2 + \lambda \sum_{j=0}^{k} \beta_j^2$$

- Select final $\hat{\beta} = \hat{\beta}_{\lambda_{cv}}$ by cross-validating over λ
- $\hat{\beta}_\lambda$ is also the posterior mean/MAP of Bayesian ridge regression with fixed variance $\sigma^2 = 2\lambda$
- under 'bad' misspecification, putting prior on λ leads to posterior concentrating **on way smaller** λ than λ_{cv}

SafeBayes (Simple Version)

- Pick $\eta \left(\approx \frac{1}{\lambda}\right)$ minimizing

$$\sum_{i=1}^{n} \left(-\log P_{\bar{\beta}_{\eta,Z^{i-1}}}(Y_i \mid X_i)\right)$$

...where $\bar{\beta}_{\eta,Z^{i-1}}$ is the mean of the η-generalized posterior based on data $Z^{i-1} = (X^{i-1}, Y^{i-1})$

SafeBayes (Simple Version)

(only) for regression with fixed variance

- Pick $\eta \left(\approx \frac{1}{\lambda}\right)$ minimizing

$$\sum_{i=1}^{n} \left(-\log P_{\bar{\beta}_{\eta,Z^{i-1}}}(Y_i \mid X_i)\right) = \sum_{i=1}^{n} \left(Y_i - \bar{\beta}_{\eta,Z^{i-1}}^T \cdot g(X_i)\right)^2$$

...where $\bar{\beta}_{\eta,Z^{i-1}}$ is the mean of the η-generalized posterior based on data $Z^{i-1} = (X^{i-1}, Y^{i-1})$

forward (rather than cross-) validation

loss function in general non-standard

Main Result as a Slogan

- "Generalized Bayes is great...
  ...once you know the right learning rate"
- "Safe Bayes is great...
  ...even if you don't know the learning rate"

Main Result as a Slogan

- "Generalized Bayes is great...
  ...once you know the right learning rate"
- "Safe Bayes is great...
  ...even if you don't know the learning rate"

Most Extensive Explanation So-Far: *Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It* http://arxiv.org/abs/1412.3730

Final Remarks - I

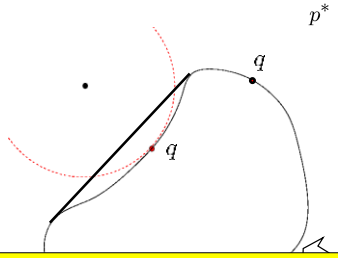
- Bayes can be in trouble when model is wrong but useful; adding learning rate helps
- There are other issues with Bayes as well, e.g. in nonparametrics. These are ***not*** the classical objections 'it is subjective' or 'where do you get your prior'
  - see 'Larry and Jamie take on a Nobel Prize Winner' on Larry Wasserman's blog
- ...but also amazing successes

Final Remarks

- For sequential prediction, the learning-rate approach is common among theorists and **extremely robust**
- Just a few practical applications, but these are **very successful**
- **...e.g. online prediction of electricity demand in Paris region**
  - M Devaine, P Gaillard, Y Goude, G Stoltz, Machine Learning 90 (2), 231-260
- **They actually used SafeBayes (unconsciously)**

Extra Material

Bad and Good Misspecification



Standard Bayes can exploit bad misspecification to get small KL risk; Safe Bayes cannot

What are these "right" learning rates?

- Model $\mathcal{P}$ of distributions correct or convex: $\eta_{\text{crit}} = 1$
  –
- Model $\mathcal{H}$ of predictors
  – **mixable** loss function, eg.squared, logistic loss, support of Y under P^* bounded: $\eta_{\text{crit}} = c > 0$ (Vovk '90)
  – 0/1-loss, if $(P^*, \mathcal{H})$ satisfies Tsybakov-Mammen margin condition then (Audibert '04)
  $\eta_{\text{crit}} \asymp n^{-\alpha}$, for some $\alpha \in [0, 1/2]$